The Janus Faced Scholar

a Festschrift in honour of Peter Ingwersen

Larsen, Birger; Schneider, Jesper Wiborg; Åström, Fredrik

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Larsen, B., Schneider, J. W., & Åström, F. (Eds.) (2010). *The Janus Faced Scholar: a Festschrift in honour of Peter Ingwersen*. Det Informationsvidenskabelige Akademi. http://www.issi-society.info/peteringwersen/

the janus faced scholar

a festschrift in honour of peter ingwersen

The Janus Faced Scholar

A Festschrift in Honour of Peter Ingwersen

The Janus Faced Scholar

A Festschrift in Honour of Peter Ingwersen

special volume of the e-zine of the
international society for scientometrics and informetrics

vol. 06-S June 2010

Published by
Det Informationsvidenskabelige Akademi
(Royal School of Library and Information Science, Copenhagen)

Under the auspices of
ISSI

The Janus Faced Scholar. A Festschrift in Honour of Peter Ingwersen.
Special volume of the e-zine of the ISSI, vol. 06-S June 2010

ISBN 978-87-741-5318-4 (print)
ISBN 978-87-741-5319-1 (electronic)

Cover & technical editing: Balázs Schlemmer
Printed by Jespersen Tryk + Digital
June 2010

Contents

Foreword

With this Festschrift we wish to honour Professor Peter Ingwersen on his retirement from the Royal School of Library and Information (RSLIS) and concomitant appointment as the first Professor Emeritus at the Royal School.

Why we wish to honour Peter

As the list of contributors and congratulators demonstrate, Peter Ingwersen's influences are manifold and widespread. He has been an active teacher and researcher at RSLIS since 1973, and for nearly four decades Peter's teaching abilities have been appreciated by numerous students on all levels – among them the editors of this volume. In fact Peter Ingwersen was one of the driving forces behind the establishment of a master's degree in Library and Information Science in Denmark, as well as a later PhD program. Peter Ingwersen has been a supervisor for several PhD students in Denmark as well abroad. He has been an appreciated opponent on numerous international PhD defences in information science, information retrieval and informetrics. And Professor Peter Ingwersen has also been a driving force in establishing the South African information science community.

Peter is well known for his mentoring and especially social skills. Master students, PhD students and colleagues, literally all over the world, has benefited from Peter's intellectual depth, always constructive comments, and not least wit. He has an ability to fascinate and above all *inspire* especially young researchers, always asking about their interests, giving comments and suggestions – thus learning about the newest and brightest ideas. Many friendships have been initiated through Peter's insistent networking abilities; he brings people together. Indeed collaboration has been trademark for Peter Ingwersen. He has been a visiting professor at several international research institutions. He has organized, or participated in, numerous international conferences and PhD courses, as well as being an active host for guest scholars and students at the RSLIS. As a testimony to his collaboration, almost 60% of the 183 publications in his bibliography are co-authored.

His contribution

Professor Peter Ingwersen is an interesting case when it comes to his research profile; a recurring topics in several contributions in this Festschrift. He has been active in the three main research areas of information science: “information behaviour”, “information retrieval” and “informetrics”. More specifically, Professor Peter Ingwersen has contributed to the integration of information retrieval and information seeking research by advocating “Interactive information retrieval”. Peter Ingwersen’s theoretical stance is the cognitive view point, which he has a primer promoter of since the early 1980s. Notably the later focus on “information interaction” and the principle of “polyrepresentation” culminating in the co-authored book “The Turn, contextualizes and gives a holistic framework for interactive information retrieval. This unifying research is recognized in both the IR and IS communities.

Interestingly, Professor Peter Ingwersen’s research profile goes beyond “interactive information retrieval”. In the spirit of his holistic thinking, Peter Ingwersen also has a research profile within “informetrics”, again trying to bring bond this field with for example information retrieval. Peter’s research main areas have been webometrics and scientometrics. He actually coined the term webometrics and invented its first indicator, the Web Impact Factor. Together with colleagues, Peter Ingwersen has worked persistently on developing science and technology indicators, perhaps most notably the “diachronic impact factor”.

Peter Ingwersen is one of the most cited researchers in library and information science. In the current bibliometric maps of information science, Peter Ingwersen’s position is often at the centre of the map or network, where the three major subfields “Information behaviour”, “IR” and “informetrics” are placed around him. His position in the maps indicates that he is active and cited in all three subfields – testimony to his versatility, influence and integrative approach.

The Festschrift

Despite an impossibly short deadline the Festschrift contains more than 30 papers by 50 authors. This bears witness to the dedications that Peter invokes in the people that know him. The contributions fall into three main themes: Information Retrieval, Informetrics and Information Science. And there are broadly speaking three types of contributions: regular scientific papers that report on the current interests and future visions of the contributors, celebratory papers with congratulatory anecdotes about Peter, and finally those that have a bit of both. The topics span very widely, from *reflections on the nature of commuting between Malmö and Copenhagen*, *the historic dimension in museum contexts* over *search procedures* and *chemoinformatics*, *blogometrics* and *web impact factors* to a highly conceptual model of *polyrepresentation based on formalisms from quantum mechanics*. The Festschrift concludes with a bibliography of Peter's impressing academic production – more 180 publications on all levels and of all types. One may note that 2010 looks like a strong year with several published papers and many accepted for publication already.

Dear Peter!

With this Festschrift, we wish to honour you on the retirement as Full Professor, and to show our appreciation for you as a colleague, friend, mentor and teacher. We find it apt that you will become the first Professor Emeritus at the Royal School, and hope to draw on your wisdom and experience for many years to come. We are many academics all over the world that owe you a lot. We hope that you will enjoy this volume – there is plenty of *Nagagga* in it!

All the best wishes for your retirement, and your new role as Professor Emeritus!

Copenhagen, Aalborg and Lund, June 25, 2010

Birger Larsen
Royal School of Library and Information Science,
Birketinget 6, DK-2300 Copenhagen S (Denmark)
Email: blar[at]iva.dk

Jesper Wiborg Schneider
Royal School of Library and Information Science,
Fredrik Bajers Vej 7K, DK- 9220 Aalborg Ø (Denmark)
Email: jws[at]iva.dk

Fredrik Åström
Lund University Libraries,
Head Office, P.O. Box 134, SE-22100 Lund (Sweden)
Email: fredrik.astrom[at]lub.lu.se

Information Retrieval

On the Evaluation of Interactive Information Retrieval Systems

Nicholas J. Belkin

Rutgers University, New Brunswick, USA

Abstract. This paper briefly discusses the history of the standard information retrieval evaluation criteria, measures and methods, and why they are unsuitable for the evaluation of interactive information retrieval. A new framework for evaluation of interactive information retrieval is proposed, based on the criterion of usefulness.

Keywords: Interactive information retrieval, information retrieval evaluation.

1 Introduction

It is both a great honor, and a great pleasure for me to contribute to this celebration of the career of my long-time friend and colleague, Peter Ingwersen. Furthermore, it turns out to be, at least in one respect, a relatively easy task, in that Peter has made significant contributions in so many areas of information science, that finding a topic both relevant to his interests, and to my current research concerns, is not a great problem. Of more moment, of course, is to achieve his level of insight.

Among Peter's continuing concerns has been the evaluation of interactive information retrieval systems (e.g. [1] [2]), and it is this particular issue that I wish to address in this paper. For well on 20 years now (see, e.g. [3]), it has been quite clear that the standard Cranfield/TREC model of information retrieval (IR) system evaluation is very badly suited to the evaluation of interactive IR systems. Since IR is an inherently interactive activity, from a theoretical point of view (e.g, [4]), and has been from a practical point of view since the 1970s, it is a severe problem that almost all criteria, measures and methods used in formal IR system evaluation continue to be those which have been designed to test non-interactive IR.

In this paper, I discuss just why the standard IR evaluation criteria, measures and methods are not suited, in the general case, to the evaluation of interactive IR (IIR), suggest that the criterion of *relevance*, long held to be the central concept of IR, if not of information science itself (cf. [5]), is inappropriate (again, in the general case), and propose that considering the *usefulness* of an

IIR episode, and of its components, with respect to its contribution to the accomplishment of the task that led to the episode, can lead to both realistic and informative evaluation of IIR systems.

2 Why have IR systems been evaluated as they have been?

There is a history to the evaluation of IR systems, and I believe that it is rooted in the practices of documentation, and especially of science librarianship. Bradford's discovery of bibliographic regularities arose through his analysis of the work that he did as a science librarian [6]. That work was the compilation of subject bibliographies, primarily on request of a scientist or a group of scientists. The goal of such bibliographies was to identify all of the documents pertaining to the subject, and to not include in the bibliography any documents which did not pertain to the subject. It is not difficult to see how Cyril Cleverdon, himself a science librarian (and others, of course), could accept these as goals for an IR system, understanding the phrase "pertaining to the subject" as meaning (eventually) "relevant to the inquirer's query", making relevance of a document the basic criterion of evaluation, and therefore leading to the measures of recall and precision, emulating the "all and only" of the subject bibliography.

The very first evaluations of IR systems, as at Cranfield [7] and Western Reserve [8], and their critics (e.g. Swanson, [9]), clearly recognized that there were some inherent problems with this general analogy, and with the concept of relevance, mostly having to do with the inherent subjectivity of relevance judgments. The response to these problems by the IR research community was to attempt to remove the person from the equation, thereby eliminating subjectivity. Both Cleverdon and his regular adversary, Jason Farradane [10] accepted that this was the only manner in which "scientific" evaluation of IR systems could be conducted.

Salton's SMART project recognized another difficulty with the standard model; that is, that a person's initial expression of an "information need" in some query was quite unlikely to be the best possible such expression. In Rocchio's [11] interpretation of this fact, the problem was seen as finding the "ideal" query, and the answer was for the IR system to interpret the searcher's evaluations of document relevance (or not) as evidence for query modification. Thus, there was implied in this formulation some idea of the searcher *interacting* with the IR system, but in a strangely passive mode. More substantive interaction, involving the searcher as an active participant, and also one whose information need, as represented by a query, might change through the course of an interaction, was explicitly not considered. Thus, the evaluation model, even in this partially interactive mode, remained the evaluation of the results of one specific query, with the same "all and only" measures.

3 Why shouldn't IR systems be evaluated as they have been?

The reasons which lead people to engage in information seeking, and therefore in interaction with information retrieval systems, seem only rarely to be equivalent to the goal of the subject bibliography (cf. [12] [13] [14] [15]). Indeed, a more apt example from the same era as Bradford's, might rather be the exploration of a library in order to discover relationships among ideas which one had not thought of before, such as interacting in the library of the Warburg Institute [16]; another might be to learn about a new domain of interest, through exploration of its canonical texts; yet another might be the desire to find one document which answers a specific question; a fourth could well be to obtain advice about possible courses of action in a given situation. It would be simple to continue this list for quite some time, if not quite endlessly. An alternative is to consider the possible circumstances underlying the *problematic situation*, as initially described in Schutz & Luckmann [17]), and applied in various ways to the contexts of information science and IR by, e.g., Belkin, Seeger & Wersig [18] Wersig [19]. Schutz & Luckmann quite plainly outline at least the knowledge-oriented reasons that might lead people to engage in information seeking; none of them, however, seems to lead to that which underlies the standard IR evaluation methods and measures. Even their quite extended and explicit discussion of relevance is of a concept quite different from that normally used in IR. Indeed, when considering the range of reasons that might lead people to engage with IR systems, we find that the situations in which finding all of the documents relevant to a query (or its underlying information "need") constitute a rather small minority, which suggests that a more general evaluation model, encompassing the range of reasons or goals of information seeking might be more appropriate.

It is also the case that many, if not most information seeking interactions take place not as isolated, single queries, but rather as information seeking episodes, during which various activities, including, but definitely not limited to the posing of different queries, take place (cf. Belkin, 1996 [20]: Fuhr, 2009 [21]). It thus makes sense to consider an evaluation paradigm which undertakes the evaluation of the search episode as a whole. But the relevance criterion and the "all and only" measures are suited (indeed designed) to evaluate the success of a single query, and it seems at the very least exceedingly difficult to adapt them to the evaluation of an entire search episode. The struggles, and eventual failure of the TREC Interactive Track Dumais and Belkin 2005 [22] in its attempt to evaluate IIR within the strictures of the standard evaluation paradigm give testimony to aspects of this problem. Järvelin, et al., 2008 [23] is an example, perhaps the only extant example, of an attempt at directly using relevance as the criterion for evaluation of an entire search episode, albeit with a quite different measure than recall or precision. The difficulties that they faced, and the problems that arose in the test of their measure and methods, illustrate the extreme difficulty of using relevance for this purpose. More often, when considering the evaluation of IIR, relevance and its companion measures have just been discarded, or, as in the TREC Interactive Track, supplemented by a variety of alternative measures. Su [24] suggested a measure which could, in principle, be applied to the entire search episode, "value of search results as a whole', which in fact does away completely with ideas of recall and precision, and perhaps even relevance, at least as commonly understood. Similarly, "satisfaction", measured according to multiple criteria, including satisfaction with the search episode (often operationalized as the interaction with a library and a librarian) has long been suggested (and used) as a more holistic criterion than just relevance for evaluation of IIR (e.g. Tagliacozzo [25]).

Furthermore, the nature of IIR is such that the information seeker's state of knowledge is quite likely to change during the course of the information seeking episode [14], leading to new ideas of what might be useful, as could even the person's understanding of the problem or task that led to information seeking [18]. As Bates [12] and Oddy [26] have proposed, just seeing some new text during the course of information seeking could lead to quite new ideas about what other texts it would be nice to encounter. But the only kind of interaction that the normal IR evaluation paradigm readily allows, relevance feedback leading to an ideal query, takes no account of these sorts of changes.

Thus, the standard IR evaluation paradigm fails to respond to the fundamental nature of IIR, in terms of the kinds of goals for information seeking that it presupposes, in terms of its inability to evaluate entire information seeking episodes, and in terms of its inability to account for the changes in the searcher that are inherent in interactive information seeking.

4 Usefulness as the criterion for evaluation of interactive information retrieval

Assume that the ultimate goal of IR is to support people in the resolution of their problematic situations [18] [20]. An operationalization of this goal that has been accepted by the IR community is the provision of texts relevant to a query. But quite different operationalizations can be, and have been imagined. Cooper [27], for instance, suggested that the *utility* of a search result is a more realistic criterion. My colleagues and I at Rutgers have questioned relevance as an appropriate criterion for evaluation of IIR, and suggested elsewhere that *usefulness* could be a much more realistic criterion [28] [29] [30]. Here, I draw on that work, sketching an outline of the argument in favor of usefulness, with some discussion of how it could be applied.

We begin by considering the issue of how to evaluate an IIR system in terms of the goal that we have assumed. The question that immediately arises is: how to relate what the system does (or doesn't do) to the resolution of the problematic situation. The issue here is how to know to what extent the problematic situation has been resolved; already in 1974, John Martyn [31] pointed out that our concern should be with the *use* of the information gained through interaction with the information system, yet we still lack methods, or a sound framework for directly understanding this relationship. One possibility for addressing this problem is to specify, quite concretely, the *task* which the searcher intends to accomplish, and then to measure to what extent, or how well that task has actually been accomplished, after the information retrieval interaction. To some extent, the method proposed by Borlund and Ingwersen [1] attempts to address this issue. The major difficulty remains the ability to establish a direct connection between what the system did, and what effect that had on the task outcome. Jean Tague's [32] proposal of a measure of *informativeness* was an early step in this direction, which has unfortunately not been followed up in subsequent research.

Our proposal for addressing this problem is to consider the *usefulness* of the IR interaction with respect to the motivating task at three distinct levels:

1. The usefulness of the entire interaction with respect to the motivating task;
2. The usefulness of each step in the information seeking episode with respect to accomplishing the goal of the interaction, and with respect to its contribution to accomplishment of the motivating task;
3. The usefulness of system support with respect to the goal of each individual step in the interaction.

Our contention is that, by decomposing the tasks/goals of an information seeking episode in this way, it will be possible to relate system support behaviors associated with each individual step during the course of the information seeking episode with the extent to which the motivating task has been resolved, combining both summative (motivating task) and analytic (individual step goals) evaluation methods.

The method, in the abstract, is as follows. First, the motivating task is elicited (in the case of participants searching for their own purposes) or controlled (as proposed in [1]), as are criteria and measures for evaluating the extent to which the task will be or has been accomplished, respectively. The goal of the information seeking episode itself is treated in the same manner. Then, the searcher engages in the IIR system, and the task (in the case of controlled searching) completed. All activities during the information seeking episode are logged/recorded.[1] At this point, task accomplishment is evaluated, and searcher evaluation of the usefulness of the information seeking interaction with respect to task accomplishment is elic-

[1] In the case of uncontrolled searching, at the end of the search, both motivating task and information seeking goal are again elicited, in order to confirm that they did not change; if they did change, we engage in the elicitation and measurement activity with respect to these, and consider when and why the changed in subsequent elicitation.

ited, as is the goal of the information seeking episode itself. Then, each step in the information seeking episode is examined, sequentially, eliciting from the searcher the goal of each step, in and of itself, and with respect to the accomplishment of the episode's information seeking goal, and the extent to which the goal of the specific step was achieved, and the usefulness of that step toward the accomplishment of the information seeking goal.

This procedure allows not only the establishment of the relationship of each support technique (associated with the individual steps) with the outcome of the searching process, and task accomplishment, but also can evaluate the sequencing of the steps, as a process leading to information seeking goal and task accomplishment. We have not considered in this description a number of factors that would need to be controlled or taken account of, in order to interpret the data appropriately. These would include, *inter alia*, characteristics of the searcher such as searching, topic and domain knowledge, cognitive abilities, and other individual differences. But we already have examples of how this could be done in a variety of IIR experiments.

Clearly, the method as outlined above is likely to be too cumbersome to be enacted in whole in a realistic (i.e. relatively large) evaluation exercise. But, one can imagine how various aspects of the evaluation could be accomplished without the great involvement of the searcher that is described. For instance, using the method of [1], suitably enhanced, can eliminate searcher involvement in the first step. Examining the search log to see what uses have been made of each step in subsequent steps could substantially reduce searcher involvement in evaluation of usefulness of each step toward the information seeking goal. Inferring individual step goals from the specific behaviors within each step, and applying appropriate evaluation measures, could again reduce searcher involvement. And, examining the sequence of steps for "aberrant" sequences (e.g. repetitions, backtracking) could inform the identification of an "ideal" sequence, and an evaluation of the system's support for helping the searcher to engage in that sequence. Of course, being able to do these sorts of abstractions will require substantial preliminary research using the full, searcher intensive method, but this should not deter us from moving toward the goal of truly good evaluation of IIR.

In summary, the criterion of usefulness, properly construed, can not only incorporate previous criteria, such as relevance, as special cases appropriate for evaluating specific steps within an information seeking episode, but also offers the opportunity to evaluate the effectiveness of an IIR system in such a way as to relate the support characteristics of that system to the success of the information seeking episode as a whole, in supporting the resolution of the searcher's problematic situation, and the accomplishment of the task that led the searcher to engage in information seeking behavior.

Acknowledgements: This research was supported by IMLS Grant LG-06-07-0105-07. Much of the content of this paper is due to my colleagues in the PoODLE Project2 at Rutgers University, although I take sole responsibility for opinions expressed herein.

5 References

1. Borlund, P., Ingwersen, P.: The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. Journal of Documentation. 53, 225-250 (1997)
2. Borlund, P., Ingwersen, P. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 324-331. ACM Press, New York (1998)
3. Roberson, S.E., Hancock-Beaulieu, M.M.: On the Evaluation of IR Systems. Information Processing and Management. 28, 457-466 (1992)
4. Belkin, N.J.: Information Retrieval as Interaction with Information. In: Information Retrieval '93: von der Modellierung zur Anwendung, pp. 55-66. Konstanz, DE, Universitäts Verlag Konstanz (1993)
5. Saracevic, T.: Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Journal of the American Society for Information Science, 26, 321-343 (1975)
6. Bradford, S.C.: Documentation. C. Lockwood, London (1948)
7. Cleverdon, C.: The Cranfield Tests on Index Language Devices. Aslib Proceedings. 19, 173-194 (1967)
8. Saracevic, T.: An Inquiry into Testing of Information Retrieval Systems, Part i: Objectives, Methodology, Design and Control. Final technical report, Grant Phs Fr-00118 (1968).
9. Swanson, D. R.: Some Unexplained Aspects of the Cranfield Tests of Indexing Performance Factors. Library Quarterly. 41, 221-228 (1971)
10. Farradane, J.: The Nature of Information. Journal of Information Science, 1, 13-17 (1979).
11. Rocchio, J.J.: Relevance Feedback in Information Retrieval. In: Salton, G. (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313-323. Prentice-Hall, Englewood Cliffs, NJ (1971)
12. Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review, 13, 407-423 (1989)

[2] http://comminfo.rutger.edu/imls/poodle

13. Belkin, N.J.: Anomalous States of Knowledge as a Basis for Information Retrieval. Canadian Journal of Information Science. 5, 133-143 (1980)

14. Belkin, N.J., Oddy, R.N., Brooks, H.M.: ASK for Information Retrieval. Part I:. Background and Theory. Part II: Results of a Design Study. Journal of Documentation, 38, 61-71, 145-164 (1982)

15. Marchionini, G.: Exploratory Search: From Finding to Understanding. Communications of the ACM, 49 4, 41-46 (2006)

16. Gombrich, E.H.: Abby Warburg: An Intellectual Biography, 2nd edition. University of Chicago Press, Chicago (1986)

17. Schutz, A., Luckmann, T.: The Structures of the Life World. Northwestern University Press, Evanston, IL (1973)

18. Belkin, N.J., Seeger, T., Wersig, G.: Distributed Expert Problem Treatment as a Model for Information System Analysis and Design. Journal of Information Science, 5, 153-167 (1983)

19. Wersig, G.: Information – Kommunikation – Dokumentation. Pullach bei München, Verlag Dokumentation (1971)

20. Belkin, N.J.: Intelligent Information Retrieval: Whose Intelligence? In: ISI '96. Proceedings of the Fifth International Symposium on Information Science, pp. 25-31. Konstanz, DE, Universitäts Verlag Konstanz (1996)

21. Fuhr, N.: A Probability Ranking Principle for Interactive Information Retrieval. Information Retrieval. 11, 251-265 (2008).

22. Dumais, S.X., Belkin, N.J.: The TREC Interactive Tracks: Putting the User into Search. In: Voorhees, E.M., Harman, D.E. (eds.) TREC, Experiment and Evaluation in Information Systems, pp. 123-152. Cambridge, MA, MIT Press (2005)

23. Järvelin, K., Price, S.L., Delcambre, L.M.L., Lykke Nielsen, M.: Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In: Macdonald, C., et al. (eds.) ECIR 2008, LNCS 4956, pp. 4–15. Springer-Verlag, Heidelberg Berlin (2008)

24. Su, L.: Evaluation Measures for Interactive Information Retrieval. Information Processing and Management, 34, 557-579 (1998)

25. Tagliacozzo, R.: Estimating the Satisfaction of Information Users. Bulletin of the Medical Library Association, 65, 243-249 (1977)

26. Oddy, R.N.: Information Retrieval through Man-Machine Dialogue. Journal of Documentation, 33, 1-14 (1977)

27. Cooper, W.S.: On Selecting a Measure of Retrieval Effectiveness. Journal of the American Society for Information Science, 24, 87-100

28. Belkin, N.J., Bierig, R., Cole, M.: Is Relevance the Right Criterion for Evaluating Interactive Information Retrieval. In: Proceedings of the ACM SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments. http://research.microsoft.com/~pauben/bbr-workshop (2008)

29. Belkin, N.J., Cole, M., Liu, J.: A Model for Evaluation of Interactive Information Retrieval. In: Proceedings of the ACM SIGIR 2009 Workshop on Understanding the User. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-512/ (2009)

30. Cole, M., Liu, J., Belkin, N.J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., Zhang, X.: Usefulness as the Criterion for Evaluation of Interactive Information Retrieval. In: Proceedings of the third Workshop on Human-Computer Interaction and Information Retrieval. http://cuaslis.org/hcir2009/ (2009)

31. Martyn, J.: Information Needs and Uses. In: Cuadra, C. , Luke, A.W. (eds.) Annual Review of Information Science and Technology. 9, 3-24 (1974)

32. Tague-Sutcliffe, J.: Measuring the Informativeness of a Retrieval Process: In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 23-36. ACM Press, New York (1992)

Address of congratulating author:

Nicholas J. Belkin
Department of Library and Information Science
School of Communication & Information, Rutgers University
4 Huntington Street, New Brunswick, NJ 08901, USA
Email: belkin[at]rutgers.edu

The Cognitive Viewpoint: The Essence of Information Retrieval Interaction

Pia Borlund

Royal School of Library and Information Science, Aalborg, Denmark

Foreword: A tribute to Professor Peter Ingwersen from a former student of his

This paper is in honour of Professor Peter Ingwersen on the occasion of his retirement from the Royal School of Library and Information Science, Denmark.

Personally, I have known Professor Ingwersen since 1993. Meeting him the first time was breath-taking, as he is in many ways an atypical Danish personage. Atypical in the sense of being flamboyant, charismatic, colourful, and proudly (as well as loudly) confident of himself – this said in the most positive sense. In addition, he presents himself with an undeniable enthusiasm (close to love) for Information Science in general, and information retrieval (IR) interaction in particular. This enthusiasm and dedication of his was (is) contagious, and is the very reason why I ended up with an academic career in Information Science. A career which he has strongly supported and helped along by his attention, advices, his generous sharing of his world-wide network of colleagues, and by forming a stimulating research environment with room for exciting, inspiring, and thought-provocative discussions. For all this I am most grateful to Peter!

However, Professor Ingwersen is not only a benefactor and of importance to me, but also to the field of Information Science and the IR community, in that he is the leading proponent of the *cognitive viewpoint*, and tirelessly carries on in the further development and promotion of this viewpoint.

The present paper builds upon a chapter on the introduction to the cognitive viewpoint from my doctoral thesis [11] of which Professor Ingwersen was my supervisor.

1. Introduction

The cognitive viewpoint is user-centred and acknowledges the user's personal perception of the information need, the consequently subjective relevance assessments of information in response to that information need, and the context that surrounds the user, creates the given situation, and shapes the information need. As such the cognitive viewpoint is concerned with the concept of the information need and its formation process as perceived and acted upon by the user – at a more abstract level referred to as the changes or transformations of knowledge structures of the recipient by the act of communication and the processes of perception, evaluation, interpretation, and learning [25]. In essence, the viewpoint is about the user's desire for information, and hence is a platform for authentic information studies of users' retrieval, search and seeking interactions in the process of achieving this goal of desired information.

The objective of the paper is to introduce the cognitive viewpoint within the field of Information Science, and the research area of IR, by outlining the main characteristics of the viewpoint. Hence it is not an ambition of the paper to provide an exhaustive literature review of the cognitive viewpoint. The presentation of the viewpoint and its impact on Information Science and IR is based on the selective, though representative works of the four predominant scholars: B.C. Brookes, N.J. Belkin, M. De Mey, and P. Ingwersen.

1.1 The cognitive viewpoint: a contribution to the field of Information Science

The history of Information Science is characterised by its concern with itself as a discipline and its scope of study. Over the years this has resulted in various proposals of what Information Science should study, and how that should be done. One of the proposed epistemological approaches to Information Science is the cognitive viewpoint [e.g., 3-6; 12-17; 22; 23; 25; 27; 28]. Within the research area of IR the viewpoint was introduced as an alternative to the mainstream system and document-driven IR research tradition [19].

It is impossible to name anyone specifically as the originator of the cognitive viewpoint in Information Science. Belkin [7, p. 11] points out that a number of publications started to appear from the mid-1970s that explicitly called for, or proposed, a cognitive view of Information Science [2; 12-14; 18]. Although they did not have precisely the same definition of what such a view is, or what it entails, there was consensus of the meaning common to them all. Later Wilson summaries the cognitive viewpoint as "...the idea of human perception, cognition, and structures of knowledge" [34, p. 197].

Over the years some of the contributors have achieved the recognition of being personalised with the cognitive viewpoint, e.g., Brookes, Belkin, De Mey, and Ingwersen. Brookes is one of the earliest proponents of the view, and of great inspiration to many others who also advocate for this view. Brookes' contribution, in this context, is his 'fundamental equation of information science' which embodies the explicit form this view takes for him [12-15]. Belkin is personally inspired by

Brookes [7]. Belkin has received great acknowledgment for his proposal of the ASK hypothesis which is to be seen as the result of his definition of the relationships and phenomena with which Information Science should be concerned [2-3; 8; 9]. De Mey is primarily associated with the epistemological aspects of the viewpoint [16]. He is often ascribed as the originator of the viewpoint. Ingwersen has in several cases demonstrated the applicability of the cognitive viewpoint to Information Science either in his own work or by reference to the work of fellow researchers. As such, Ingwersen in recent time is seen as the leading and most active proponent of the cognitive viewpoint to Information Science. The following four sub-sections present the four scholars' basic ideas of the cognitive viewpoint.

1.2 Brookes: The fundamental equation of Information Science

Brookes' contribution to the theoretical development and identification of the scope of the field of Information Science is his proposal of the *fundamental equation of Information Science*. The fundamental equation is a model equation, which expresses how knowledge structures are affected and become modified as the consequence of the intervention of externally added information. From Brookes' point of view the fundamental equation is a tool to the field of Information Science to help in uncovering and understanding the scope of the field. Brookes proposes and discusses in detail his fundamental equation in a series of articles [12-15]. The final form of the equation is published in 1980 [15, p. 131], and is expressed as follows:

$$K\,[S] + \Delta I = K\,[S + \Delta S]$$

The equation "…states in its very general way that the knowledge structure $K\,[S]$ is changed to the new modified structure $K\,[S + \Delta S]$ by the information ΔI, the ΔS indicating the effect of the modification" [15, p. 131]. Brookes comments that the equation implies that information is *that* which modifies what is denoted by $K\,[S]$, which is a knowledge structure; that knowledge and information have the same dimensions; and that information is, as is the knowledge discussed, structured [14; 15]. As pointed out by Belkin [7] this implication demonstrates the power of the cognitive viewpoint with its emphasis on knowledge structures and their interactions with one anther. However, Brookes at the same time as formulating the equation also explicitly says that the fundamental equation does not solve problems for Information Science, but rather poses them. He says "…the interpretation of the fundamental equation is the basic research task of Information Science…" [13, p. 117]. In his own work he attempts to do this. Brookes reaches his aim with the fundamental equation as the equation has over the years have led to various discussions and proposals of clarifications of the scope of the field. Even improvements and adjustments of the equation can be included, such as the suggestions by Ingwersen [23; 25].

Another cognitively based contribution to the field of Information Science as a discipline and to its scope of study is proposed by Belkin.

1.3 Belkin: The relationships and concepts of Information Science – ASK

Belkin contributes to the identification of the scope of the field of Information Science by outlining the problem of Information Science as well as its implications. Belkin takes the fundamental problem of Information Science to be the "…effective transfer of desired information from human generator to human user…" [2, p. 197; 3, p. 187]. According to Belkin [3, p. 187] the problem implies at least the following set of concerns for Information Science:

1. the relationship between information and the generator of that information;
2. the concept of desired information;
3. the relationship between information and user; *and*
4. the concepts of effectiveness of information and of information transfer.

In 1978 Belkin [4, p. 58] adds a fifth set of concern, which is:

5. the information in human, cognitive communication systems.

Like Brookes, Belkin is occupied with the processing of information, though expressing it more specifically than Brookes by restricting his study to the transfer and processing "…from human generator to human user..." [3, p. 187]. Belkin operates with the concept of 'knowledge state' which briefly explained refers to the user's mental model (a world model or image) of him/herself and his/her world of conceptual knowledge and prejudices. In this connection Belkin introduces the concept of an 'anomalous state of knowledge', shortened to the acronym ASK, and commonly known as the ASK hypothesis [5; 8]. He explains how the ASK concept is a synthesis of previous works by, e.g., Taylor [32] and Wersig [33]. Belkin [5, pp. 136-137] describes how an ASK shares characteristics of the 'problematic situation' suggested by Wersig [33], and the need development level one and two outlined by Taylor [32]. An 'anomalous state of knowledge' is a conceptual state, which the user realises is deficient and wishes to correct. For example, the user's recognition of an insufficient knowledge model, which results in a need for information in order to reduce uncertainty or solving a problem. A change in the user's state of knowledge due to the impact of new information is identical to the change of knowledge structure of Brookes.

It is due to the ASK concept that the cognitive viewpoint achieves a breakthrough in IR, and the cognitive revolution becomes a reality [30]. The result of the ASK idea is that the user's information need is seen as a *reflection* of an anomalous state of knowledge. This is a change of the scientific perception of the information need from a static concept (as viewed and applied in the system-driven approach to IR) to a user-individual and potentially dynamic concept (as employed by the cognitive user-oriented approach to IR).

The work by Belkin has had a great impact on past and present research. In addition to the ASK hypothesis, Belkin contributes to the understanding of the concept of information which, like Brookes, he views as a communicated and transformed knowledge state in the form of a structure [2, p. 198; 4, p. 80]. In parallel to Belkin's proposals of the relationships and phenomena to study in Information Science, De Mey successfully frames the philosophical foundation and rationale of the cognitive viewpoint which the ideas and works by Belkin (and Brookes) are based on.

1.4 De Mey: The cognitive paradigm[1]

According to De Mey [16, p. XVI] a strong movement, establishing itself as a cognitive science, is seen within a diversity of fields (e.g., psychology, artificial intelligence (AI), sociology, and anthropology). De Mey suggests that attention is brought to this approach as it might be of use also to the field of Information Science. To De Mey the central point of the cognitive view is "…that any *processing of information*, whether perceptual or symbolic, is *mediated* by a system of categories or concepts which, for the information-processing device, are a *model* of his *world*" [16, pp. XVI-XVII]. In order for De Mey to understand as well as to illustrate the power and impact of the cognitive viewpoint, he adopts the view to AI (more specifically to 'visual perception and language understanding). This leads to the extension of a classification by Michie [29] on the stages through which the thinking on information processing has developed. The four stages are:

"1. [*A monadic stage*] during which information units are handled separately and independently of each other as if they were simple self-contained entities.
2. [*A structural stage*] where the information is seen as a more complex entity consisting of several information units arranged in some specific way.
3. [*A contextual stage*] where in addition to an analysis of the structural organization of the information-bearing unit, there is required information on context to disambiguate the meaning of the message.

[1] De Mey refers to the approach of the cognitive viewpoint as a paradigm; however the meaning of the word is not to be understood as strictly as in the Kuhnean sense of the paradigm concept.

4. [*A cognitive or epistemic stage*] in which information is seen as a supplementary or complementary to a conceptual system that represents the information-processing system's knowledge of its world" [17, p. 49].

The stages are to be seen as evolutionary stages, as each new stage draws upon the features of the foregoing one. De Mey [17, pp. 49-51] explains the implications of the stages: stage 1) implies template matching; stage 2) feature analysis; stage 3) contextual analysis; and stage 4) analysis by synthesis. Gradually, the development goes from sign and object in the message toward world knowledge of the information-processing system. De Mey [17, p. 54] puts it as follows: "From clearly delineated units handled in isolation toward handling information processing in terms of world models". This corresponds to how the cognitive viewpoint is to be seen as an alternative to the traditional system-driven view of information handling and processing. Generally speaking, the fourth and final stage illustrates the level on which most human information processing takes place [25, p. 23] including the processes of the information need formation and development. This is perfectly in line with the works and ideas of Brookes and Belkin, and is also to be seen as the reason why the cognitive approach to Information Science has become so useful to the increasing community of user-centred IR research.

1.5 Ingwersen: The cognitive view as a holistic view

At the time of De Mey's work on the epistemological aspects of the cognitive viewpoint, and Brookes and Belkin's attempt to identify the scope of Information Science, Ingwersen was one of the young researchers who entered the community of user-centred IR research.

The contribution by Ingwersen can be roughly divided into two categories of contributions. The first category contains contributions of works where Ingwersen further develops and adds to the works of his own and fellow scholars [e.g., 23; 25; 27]. The second category covers and refers to the works where Ingwersen demonstrates the applicability of the viewpoint's philosophical framework to cognitive user-centred studies and investigations of information use and transfer [e.g., 22-24; 26].

With reference to the first category, the following are illustrative examples of how Ingwersen has further developed the works by his three fellow scholars. In the section on Brookes' contribution it is briefly mentioned that Ingwersen has suggested improvements of the equation. Ingwersen's suggestions are carried out in regard to the 1977-version of the equation and not the final version otherwise reported on here (Brookes' 1977 equation reads: $[\Delta I] + K \rightarrow [K+\Delta K]$ [14, p. 197]). Ingwersen finds the expression of the 1977-version to be more dynamic. Ingwersen inserts to the equation the element of potential information (pI), and hence the equation reads: $pI \rightarrow \delta I + K(S) \rightarrow K(S + \delta S) \rightarrow pI'$ [23, p. 468]. The idea is that a user's new-generated information might be potential information to others. In 1992 Ingwersen further modifies the equation with the adding of the concept of data or designation (D), which leads to the following expression: $pI \rightarrow D + K(S) \rightarrow K(S) + \delta S \rightarrow pI'$ [25, p. 32]. Hereby representing the system's handling of the user's input.

Ingwersen also adds specifications to the model of the cognitive communication system by Belkin [5, p. 135]. Later the modified model (please see Fig. 1) becomes sort of a trademark of Ingwersen's research and view of IR, as he in several cases uses it to present his holistic view of IR interactions [e.g., 22, p. 171; 23, p. 469; 24, p. 222; 25, pp. 16, 135, 148; 27, p. 9]. Most recently, the model appears in Ingwersen and Järvelin's book [28]. Here it exists in several versions and with various detail levels according to the given focus and objective of illustration of the model.

Another example of further development is Ingwersen's extensions to the MONSTRAT-model by Belkin and colleagues [10], which leads to the more comprehensive Mediator model. The objective of the Mediator model is to be a tool for identification of topical domain, system models, feedback generator, requests, and user characteristics [25, pp. 206-220]. And in relation to De Mey's definition of the cognitive view, Ingwersen emphasises that the world model consists of cognitive structures that are determined by the individual and its social/collective experiences, education, training etc. [22, p. 168]. The modifications and further developments by Ingwersen to the works of his fellow scholars show his holistic view of the cognitive viewpoint to Information Science.

In regard to the second category of contributions by Ingwersen, the applicability of the cognitive viewpoint, one example is the empirical investigation of the transfer processes involved in reference work in public libraries [22]. Another example is the proposal of the cognitively based principle of poly-representation [e.g., 25-27]. The poly-representation principle is an information searching strategy based on the idea of cognitive overlaps. That works by means of the conscious exploration of cognitive inconsistencies of a variety of knowledge representations/ knowledge structures and interpretations by the involved agents in IR. Further, the principle implies that cognitive overlaps of information objects, originating from different interpretations of such objects (i.e., simultaneous use of different methods of knowledge representation, and a variety of different IR techniques of different functional and cognitive origin), may lead to retrieval results that decrease the degree of uncertainty inherent in IR. Recently, Ingwersen has moved into the research area of Informetrics (scientometrics, bibliometrics and webometrics). In these areas he demonstrates the applicability of the poly-representation principle with the merger of different types of knowledge structures/representations in the form of citations seen as evidence of interpretations [e.g., 1; 20; 21; 31]. Another fine example of Ingwersen's demonstration of the applicability of the cognitive viewpoint is the book titled "The turn: Integration of information seeking retrieval in context" that is co-authored with Järvelin [28]. "The turn" aims at integrating research in information seeking and IR by providing a research framework based on the cognitive viewpoint and by posing research questions to be addressed in order to take the IR and information seeking research a step further.

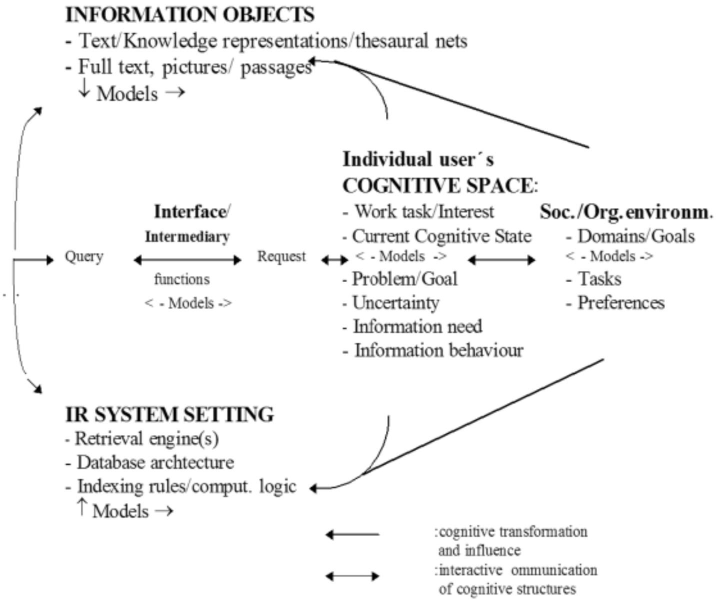


Fig. 1. Ingwersen's holistic cognitive model of IR interaction.

The modifications and further developments by Ingwersen expand as well as specify in detail the conditions of the cognitive viewpoint. In other words, the further developments illustrate Ingwersen's holistic view of the IR interaction scenario within the field of Information Science. With his holistic cognitive view Ingwersen emphasises how each of the involved cognitive agents (e.g., the information generator, the information re-presenter, the intermediary, and the information recipient/user) are of equally importance in order to achieve successful and optimal IR. To Ingwersen, the purpose of IR is to find the vortex of the appropriate harmony among the cognitive agents involved in the IR interaction.

1.6 The developing cognitive viewpoint

The cognitive viewpoint is concerned with the dynamic and interactive processing of information. The viewpoint is based on human involvement, e.g., the generator of information, the intermediary, and the recipient/user of information. The processing of information goes from the generator of the information towards the recipient of the information, with the purpose of causing an effect in the state of knowledge of the recipient. Hence, information is defined as *that* which changes a knowledge state. In addition, each of the involved agents in the information processing process is seen as individual recipients and generators. Individually, the recipients perceive the information according to their own model of the world. The concept of an information need is defined as the outcome of a change in the state of knowledge which results in an 'anomaly state of knowledge' (ASK) [5]. The change that results in an ASK, which is a cognitive development internal to the user/recipient, is happening due to an external situation, e.g., a given work task situation. In other words, an external situation causes a change in the knowledge state and in the knowledge structure of the user/recipient, which results in an ASK. An ASK is the user's recognition of an insufficient knowledge model which results in an information need, for instance, in order to reduce uncertainty. As the result of the impact of further externally added information, e.g., retrieved information, the information need may change or develop over time in order to satisfy the present problem situation as perceived by the recipient. This means that the concept of an information need, within the cognitive viewpoint, is understood as a dynamic and potentially developing concept – as indicated by the cognitive revolution presented by Robertson and Hancock-Beaulieu [30]. Basically, an information need is born out of a situation, and may develop during the process of reaching the requirements of that situation. The user's perception of an information need is thus triggered by the perception and interpretation of a given situation, a problem to be solved or a state of interest to be fulfilled, under influence of the user's current cognitive and emotional state. This state is affected by the cultural and social context within which the user acts.

The works of the four scholars show that the cognitive view to Information Science satisfies a demand of a socio-cognitive oriented approach to IR. The scholars' works define the cognitive viewpoint to be about the processing of information. Quite often this has been (mis-)interpreted in a very narrow way, in terms of a strictly user/recipient viewpoint, concerned with the information processing from the sender to the user/recipient. Brookes' 'fundamental equation' [15] has, for instance, often been understood in this restrained way in spite of his emphasis of the occurrence of cognitive processes "[a]t both ends of the channel..." [14, p. 195]. However, it is central to the viewpoint that both the generation *and* the perception of information are acts of information processing, just as the information processing depends on the actual agent's world model. The latter statement implies that all of the involved agents also function as a recipient applying their own world model. This is due to the viewpoint's basic notion of what information is (the interpretation of 'sense-data'), and its basic notion of what the informing effect on the recipient is (the change of knowledge structures/states). As such, the cognitive viewpoint is holistic by nature, as pointed out by Ingwersen [e.g., 25]. Ingwersen [27, p. 5] concludes, based on the changing roles of the involved agents in the IR scenario, that "[t]his interchange of [generator and recipient] positions makes the viewpoint a forceful theoretical foundation for IR interaction…".

Afterword: The positioning of the Royal School of Library and Information Science, Denmark

As stated in the foreword Professor Ingwersen has been of importance not only to me, but to the field of Information Science and the IR research community, too. And so has he been to the Royal School of Library and Information Science, Denmark. He has, as nobody before him, managed to position the Royal School as the world leading school in Library and Information Science. An achievement he has managed through his continuing advocacy for, and further development of the cognitive viewpoint. Every time he advocates for the cognitive viewpoint, that being via the publishing of journal and conference papers, and books, at presentations or as invited keynote speaker, when submitting research applications, and carrying out research projects, supervising students, and by being a dedicated mentor (to many of us, world wide) he represents and positions the Royal School. The Royal School is indebted to Professor Ingwersen. Hence it is a privilege that Professor Ingwersen continues as Professor Emeritus of the Royal School of Library and Information Science, Denmark.

References

Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404-426.

Belkin, N.J. & Robertson, S.E. (1976). Information science and the phenomenon of information. *Journal of the American Society for Information Science*, (27), 197-204.

Belkin, N.J. (1977). Internal knowledge and external information. In: de Mey, M., Pinxten, R., Poriau, M. and Vandamme, F., eds. *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent, 187-194.

Belkin, N.J. (1978). Progress in documentation: information concepts for information science. *Journal of Documentation*, (43), 55-85.

Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. *The Canadian Journal of Information Science*, (5), 133-143.

Belkin, N.J. (1984). Cognitive models and information transfer. *Social Science Information Studies*, (4), 111-130.

Belkin, N.J. (1990). The cognitive viewpoint in information science. *Journal of Information Science*, (16), 11-15.

Belkin, N.J., Oddy, R. & Brooks, H. (1982). ASK for information retrieval: part I. Background and theory. *Journal of Documentation*, (38)2, 61-71.

Belkin, N.J., Oddy, R. & Brooks, H. (1982). ASK for information retrieval: part II. Results of a design study. *Journal of Documentation*, (38)3, 145-164.

Belkin, N.J., Seeger, T. & Wersig, G. (1983). Distributed expert problem treatment as a model for information system analysis and design. *Journal of Information Science*, (5), 153-167.

Borlund, P. (2000). *Evaluation of Interactive Information Retrieval Systems*. Åbo: Åbo Akademi University Press. Doctoral Thesis, Åbo Akademi University.

Brookes, B.C. (1975). The fundamental problem of information science. In: Horsnell, V., ed. *Informatics 2*. London: Aslib, 42-49.

Brookes, B.C. (1975). The fundamental equation of information science. In: *Problems of Information Science, FID 530*. Moscow: VINITI, 115-130.

Brookes, B.C. (1977). The developing cognitive viewpoint in information science. In: de Mey, M., Pinxten, R., Poriau, M. & Vandamme, F., eds. *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent, 195-203.

Brookes, B.C. (1980). The foundation of information science: part I: philosophical aspects. *Journal of Information Science: Principles and Practice*, 2, 125-133.

De Mey, M. (1977). The cognitive viewpoint: its development and its scope. In: de Mey, M., Pinxten, R., Poriau, M., & Vandamme, F., eds. *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent, xvi-xxxii.

De Mey, M. (1980). The relevance of the cognitive paradigm for information science, In: Harbo, O. & Kajberg, L., eds. *Theory and application of information research. Proceedings of the 2nd International Research forum on Information Science*. London: Mansell, 49-61.

De Mey, M., Pinxten, R., Poriau, M., & Vandamme, F. (Editors). (1977). *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent.

Ellis, D. (1992). The physical and cognitive paradigms in information retrieval research. *Journal of Documentation*, 48(1), 45-64.

Hjortgaard Christensen, F. & Ingwersen, P. (1996). Online citation analysis: a methodological approach. *Scientometrics*, (37)1, 39-62.

Ingwersen, P. & Hjortgaard Christensen, F. (1997). Data set isolation for bibliometric online analyses of research publications: fundamental methodological Issues. *Journal of the American Society for Information Science*, (48)3, 205-217.

Ingwersen, P. (1982). Search procedures in the library analysed from the cognitive point of view. *Journal of Documentation*, 38(3), 165-191.

Ingwersen, P. (1984). A cognitive view of three selected online search facilities. *Online Review*, 8, 465-492.

Ingwersen, P. (1986). Cognitive analysis and the role of the intermediary in information retrieval. In: Davies, R., ed. *Intelligent Information Systems.* Chichester, West Sussex, 1986. England: Horwood, 206-237.

Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham. VII-X; 1-60; 83-156. http://ix.db.dk/ift/litteratur.htm.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In: Croft, W.B. & van Rijsbergen, C.J., eds. *Proceedings of the 17th ACM Sigir Conference on Research and Development in Information Retrieval. Dublin, 1994.* London: Springer Verlag, 101-110.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3-50.

Ingwersen, P. & Järvelin, K. (2005). The Turn: Integration of information seeking retrieval in context. Dordrecht, Netherlands: Springer Verlag.

Michie, D. (1974). *On machine intelligence*. Edinburgh: Edinburgh University Press.

Robertson, S.E. & Hancock-Beaulieu, M.M. (1992). On the evaluation of IR systems. *Information Processing & Management*, 28(4), 457-466.

Skov, M., Larsen, B. & Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match IR. *Information Processing & Management*, 44(5), 1673-1683.

Taylor, R.S. (1968). Question negotiation and information seeking in libraries. *College and Research Libraries*, (29), 178-194.

Wersig, G. (1971). *Information - kommunikation - dokumentation: ein beitrag zur orientierung der informations- dokumentationswissenschaften.* München-Pullach: Verlag Dokumentation Saur KG.

Wilson, T.D. (1984). The cognitive approach to information-seeking behaviour and information use. *Social Science Information Studies*, (4), 197-204.

Address of congratulating author:

Pia Borlund
Royal School of Library and Information Science
Fredrik Bajers Vej 7K
DK-9220 Aalborg East, Denmark
Email: pb[at]iva.dk

Genre Searching: a Pragmatic Approach to Information Retrieval

Luanne Freund

University of British Columbia, Vancouver, Canada

Abstract. This paper explores the idea of a pragmatic approach to information retrieval (IR), drawing upon the case of genre searching as a task-based information seeking strategy within a workplace setting. Making use of genre in IR systems is proposed as a means to strengthen pragmatic communication among the cognitive actors involved by building common ground, supporting joint action and increasing relevance[1].

Keywords: information retrieval, genre, pragmatics

1 Introduction

The keyword matching approach to information retrieval makes use of syntactic and semantic features of texts to predict relevance. However, much of the meaning in human communication is determined outside the realm of the text *per se*, through pragmatics: the use and interpretation of language by individuals in context. While information retrieval research has begun to address aspects of context, user behaviour and information use, dubbed the "IR ecology" by Sparck-Jones [1], and mapped out as a research agenda by Ingwersen and Järvelin in *The Turn* [2], there is still little consideration of pragmatics as a model for IR interactions. One way in which pragmatics is expressed in written communication is through the use of document genres: recognizable categories of texts that share common elements of form, content and communicative function.

This paper will explore the idea of a pragmatic approach to IR, drawing upon the case of genre searching as a task-based information seeking strategy. I propose that making use of genre in IR systems has the potential to strengthen pragmatic communication among the cognitive actors involved. The paper concludes with a brief discussion of how task-genre relationships can be situated within the General Model of Information Seeking and Retrieval (IS&R) proposed by Ingwersen and Järvelin [2].

[1] This paper is adapted from a talk given by the author at the University of Glasgow in October 2008.

2 Pragmatics and Information Retrieval

Pragmatics is the branch of linguistics concerned with language usage and the meaning that is derived from the relationships between the language, its users and the context of use [3]. Two examples of pragmatics at work are the ability to infer meaning from referential pronouns such as *he*, *she* or *it*, and from utterances that carry no implicit meaning, such as "my goodness!" by relying on shared context and mutually accepted patterns of interaction [4]. Cognitive pragmatics studies how hidden meaning is extracted from language through the workings of the mental inference mechanism and Socio-pragmatics studies the ways in which language use and interpretation are shaped by social norms [5].

One of the key ideas in pragmatics is that human communication is a social undertaking, or a Joint Action, according to Clark [6], which is based on a shared intentionality to communicate and a willingness to cooperate in achieving this goal [4]. Simply put, we assume that people are trying to communicate things that are meaningful and we make the effort to infer their intent, even when it is not obvious. A closely related concept is that of common ground: "the sum of mutual, common or joint knowledge, beliefs and suppositions" [6] held between two or more people. Without common ground, it is very difficult to frame meaningful utterances or to infer meaning from them, so a prerequisite to effective communication is the establishment of some shared awareness. Common ground can be established through shared experience, personal acquaintance, or membership in cultural or professional communities [6].

Sperber and Wilson [7] built upon these ideas to develop the communicative relevance principle. The central idea is that communicative acts carry the presumption of relevance: we expect people to communicate things that are informative and in some way connected to the context of what is going on around us, and we cooperate by interpreting what is said accordingly. Because human cognition is tuned to seek out the greatest possible effect for the smallest possible effort, people will tend to settle on the most efficient interpretation that is closely related to the current context. Furthermore, information that is easier to process will have a greater chance of being relevant and producing a "contextual effect," by changing or adding to existing knowledge.

Pragmatics assumes that people have efficient ways of communicating, interpreting and extracting meaning and that they do this by relying heavily upon assumptions based on known or shared context and through situated inference. It provides insight on how people focus their attention, construct meaning and distinguish between relevant and non-relevant information.

Current IR systems do little to support pragmatic channels of communication. Interactions between authors and readers via IR systems tend to be heavily mediated, asynchronous and remote. Document retrieval is carried out by reducing documents and queries to tokens, which are stripped out of their context of creation and use. Similarly, the context of the cognitive actors is usually unknown and uncontrolled, so that common ground cannot be assumed. Drawing upon the discussion above, a pragmatic approach to IR might establish the following goals:

- Build retrieval algorithms that can take into account the functions and uses of information objects;
- Preserve and provide to the searcher evidence of the intents and context of creation of documents;
- Facilitate and make use of context and assumptions that are shared by authors, searchers and system designers;
- Minimize the cognitive effort of searchers when assessing relevance.

3 Genre Searching

One approach to strengthening the pragmatic dimension of information retrieval is to make better use of genre. Document genres are distinctive forms of communicative acts that are recognizable within information use communities [8], [9], [10] and which serve important pragmatic functions [5]. Authors make use of genre conventions to design information objects which can accomplish their goals and can be easily recognized, interpreted and used by their readers. From a socio-pragmatic perspective, we know that members of discourse communities are socialized to create and use specific genres for particular purposes and in certain situations [8] . Within these communities, genres help to establish common ground and shared intentionality between authors and readers. From the cognitive pragmatic perspective, genres are used to trigger sets of assumptions used to infer meaning, prime expectations of relevance, and provide cues to support reading and use.

Unger [5] argues that genres develop initially as cognitive pragmatic phenomena designed to support particular communication goals and later as communities become accustomed to their use, they become established as socio-pragmatic phenomenon. Following Sperber and Wilson, it is possible to claim that genre use increases the relevance of information, because genre helps to make explicit how the document is meant to be used and the standardized features help users process the information with less cognitive effort [10].

Genre-based communication is not well supported in most IR systems to date, although it is an active area of research [11], [12]. Despite the lack of system support, genre searching is a common information seeking strategy. In a study of workplace information behaviour, I found that software engineers sought out particular genres as a shortcut to locate information suited to particular situations and tasks [13]. Fig. 1 places genre searching in the context of a broader model of information behaviour developed for this domain. On the left are the categories of contextual factors found to influence the information access constraints and the information characteristics sought. On the right are the strategies used to find the required information, one of which is genre searching.

To illustrate how this occurs, consider an engineer faced with the task of figuring out how to install a new product. In this case, he is likely to be seeking information with particular characteristics: concrete, specific, sanctioned by the company, and containing instructions. As a member of this information use community, he knows that these characteristics can be found in the product documentation genre, and so sets out to search for that genre. He has expectations of relevance and utility based on his familiarity with the genre, which also allows him to extract meaning efficiently from the document. So, genres, which are informal and constructed categories of information objects within this community, are used as implicit short cuts to finding information objects with characteristic (level of specificity, purpose, etc.) suited to particular situations. Tasks are related to genres because they determine which information characteristics are needed.

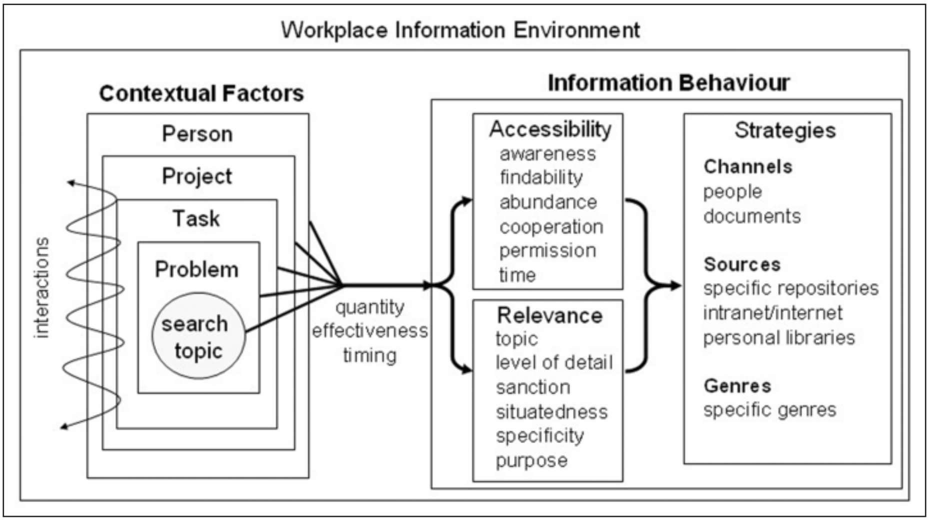


Fig. 1: Model of Contextual Effects on Searching and Selecting Behaviour [13]

Considering this relationship between tasks and genres in the framework of the General Model of IS&R [2] helps to explain it further (Fig.2). On the right of the Fig. 2 is the organizational and domain context out of which tasks arise and are imposed on the cognitive actors (searchers), prompting them to search for information. The genre types emerge out of the same context to add meaning and context to the information objects, which are created by cognitive actors (authors) and added to the information system. Shared intents and purposes motivate both the actors who are engaged in task-performing actions and the actors who are making use of genre to engage in communicative actions. Task – genre associations act as implicit links between information objects and cognitive actors performing tasks, based on the common context out of which they emerge. This implicit relationship can be modeled as an explicit set of associations and embedded in an information system, as we did with the X-Site system developed for this domain [14]; however, other implementations could be developed based on this general model.

It follows, that use of task – genre associations is likely to be most effective when the contexts of document creation and of document use are overlapping. This is the case in this setting, as the cognitive actors are all drawn from the same work group within the company. Similar situations certainly exist within other professional groups and communities of practice. However, when this is not the case, it is possible that bridges could be constructed to map tasks and genres across a contextual divide. Further work will be needed to explore the extent to which genre searching in effective in other types of settings.

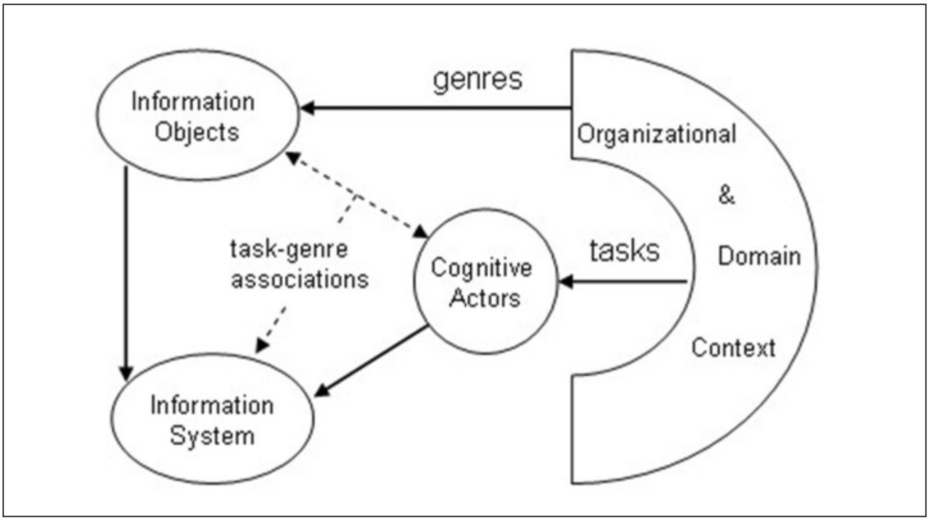


Fig. 2: The task-genre relationship mapped onto Ingwersen and Järvelin's General Model of IS&R (Adapted from [2])

4 Conclusion

The idea of supporting pragmatic communication through an IR system is appealing, but problematic. It is appealing in that it would build more directly on natural human communication behaviours and skills, which are inherently social and cooperative [4]. As Benoît [15] notes, this might help to redress a power imbalance between system designers and system users who tend to have conflicting notions of language and meaning: "To most designers, language is subjected to tests of 'technical rationality.' Word units, or semantic tokens, are taken in groups removed from other contexts, and processed. To users, it is the utterance situated in a relational context" [15]. The very notions of common ground and joint action imply greater agency and awareness on the part of searchers than has been afforded them to date.

On the other hand, the extent to which the preconditions for pragmatic communication exist when communication is asynchronous, indirect and mediated by a retrieval system is not clear. It may not be meaningful to suggest that a shared intentionality exists in this case; however, it offers something to work towards. Some work has been done to use pragmatic principles in Human Computer Interaction design [16], so this is a possible avenue for further exploration in IR.

Genre is clearly an important pragmatic phenomenon that is situated at the intersection of the cognitive and social realms. The concept of genre is intuitive to information seekers and all large, organic document collections contain genres, so there is potential to make better use of them to support IR. As context carriers, genres help to bridge the gap between cognitive authors by establishing common ground and priming expectations of relevance and use strategies. However, there are many challenges in working with genres: they are organic, dynamic and only loosely defined. Ongoing research on genre classification [12] and labeling [11] seeks to address these issues, but there is more work to be done before genre searching becomes a standard feature in IR systems.

References

1. Sparck-Jones, K.: Document retrieval: Shallow data, deep theories; historical reflections, potential directions. In: Sebastiani, F. (ed.): Advances in information retrieval: Proceedings of the 25th European Conference on Information Retrieval (ECIR), Pisa, Italy. Springer, Berlin (2003) 1-11
2. Ingwersen, P., Järvelin, K.: *The turn: Integration of information seeking and retrieval in context*. Springer, Berlin (2005)
3. Levinson, S.C.: *Pragmatics*. University Press, Cambridge, UK (1983)
4. Tomasello, M.: *Origins of Human Communication*. MIT Press, Cambridge, MA (2008)
5. Unger, C.: Cognitive-pragmatic explanations of socio-pragmatic phenomena: the case of genre. EPICS I Symposium., Seville, Spain (2002)
6. Clark, H.H.: *Using Language*. University Press, Cambridge, UK (1996)
7. Sperber, D., Wilson, D.: *Relevance: Communication and Cognition*. Blackwell, Malden, MA (1995)

8. Spinuzzi, C.: *Tracing genres through organizations: a sociocultural approach to information design*. MIT Press, Cambridge, MA (2003)
9. Orlikowski, W.J., Yates, J.: Genre repertoire: the structuring of communicative practices in organizations. *Administrative Science Quarterly* 39 (1994) 541-574
10. Dillon, A., Vaughan, M.: „It‘s the journey and the destination“: shape and the emergent property of genre in evaluating digital documents. *New Review of Multimedia and Hypermedia* 3 (1997) 91-106
11. Rosso, M.: User-based identification of Web genres. *Journal of the American Society for Information Science & Technology* 59 (2008) 1053-1072
12. Santini, M.: Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing & Management* 44 (2008) 702-737
13. Freund, L.: Exploiting task-document relations in support of information retrieval in the workplace. Faculty of Information Studies, Ph.D. University of Toronto (2008)
14. Yeung, P.C.K., Freund, L., Clarke, C.L.A.: X-Site: a workplace search tool for software engineers. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27. ACM, New York, NY (2007) 900
15. Benoît, G.: Critical theory as a foundation for pragmatic information systems design. *Information Research* 6 (2001)
16. Clark, H.H.: Arranging to do things with others. CHI ,96 (1996) 165-167

Address of congratulating author:

LUANNE FREUND
University of British Columbia
Vancouver, Canada
Email: luanne.freund[at]ubc.ca

Towards a Geometrical Cognitive Framework

Ingo Frommholz[1], Keith van Rijsbergen[1], Fabio Crestani[2] & Mounia Lalmas[1]

[1] University of Glasgow, Glasgow, Scotland
[2] University of Lugano, Lugano, Switzerland

Abstract. Ingwersen's cognitive framework is regarded as the beginning of a turn which eventually should bring together classical system-oriented and user-oriented IR communities. One of the consequences of this framework is the polyrepresentation principle. The Logical Uncertainty Principle (LUP) is regarded as a compatible model with the cognitive framework. Recently it was shown how LUP can be expressed using the mathematics of Hilbert spaces. This formalism, which is applied in quantum mechanics, harmonises geometry, probability theory and logics. Apart from being a way to express LUP, a further potential arises from a quantum perspective of IR. We present an interactive framework as an example of a quantum-inspired approach which also supports polyrepresentation.

1 Cognitive Framework in IR

One of the main assumptions behind Ingwersen's cognitive framework for IR interaction, as introduced in [3], is that processing takes place on a symbolic or sign level, whereas communication between humans may in addition take place on a cognitive level. This inevitably leads to a *cognitive "free fall"* as during the translation of a message into signs, any of its presuppositions, meaning and intentionality is constantly lost. Conversely, a human's interpretation of a message may restore the meaning and intention at the cognitive level, but there is increase in uncertainty which is an inherent feature of any communication process and hence also IR.

To tackle this problem, the principle of *polyrepresentation* exploits different cognitive and functional representations within the information and the cognitive space. In an information space, different actors (e.g., author, indexer, user) with different tasks and goals in mind (e.g., tagging, indexing, commenting, reviewing) provide various document representations. In a cognitive space, different representations comprise the user's information need, the problem state, the current cognitive state and the work task. Information needs are often (but not always) unstable and may be ill-defined and only vaguely formulated.

Polyrepresentation makes use of different representations in different spaces, which are all a result of different interpretations by the actors involved. As a consequence, the polyrepresentation hypothesises, that the more representations point to a set of documents, the higher the probability that these documents are relevant. Thus the basic idea of polyrepresentation is to facilitate a multitude of cognitively and functionally different representations, provided by different actors. This is done by determining the so-called *cognitive overlap*, which is the intersection of various sets of documents which are relevant with respect to their single representations.

2 Geometry, Logics and Probabilities in IR

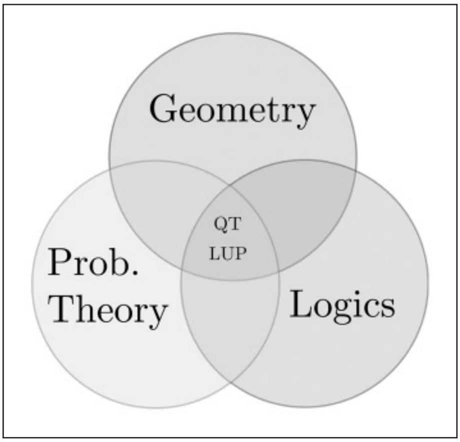


Fig. 1: Probability theory, logics and geometry

Ingwersen's cognitive framework is regarded as the beginning of a turn which eventually should bring together the classical system-oriented and the user-oriented IR communities in order to develop an integrated view of information seeking and retrieval [4]. Such a view needs a strong methodological framework or a new "language" going beyond classical prevalent models. A step towards this goal is the formulation of the *Logical Uncertainty Principle* (LUP) [7]. The LUP starts from the consideration that logic by itself cannot fully model IR. In determining the relevance of a document *d* to a query *q* the success or failure of the logical implication $d \rightarrow q$ is not enough. It is necessary to take into account the *uncertainty* inherent in such an implication. To cope with uncertainty a logic for probabilistic inference was introduced. If $d \rightarrow q$ is uncertain, then we can measure its degree of uncertainty by $\Pr(d \rightarrow q)$. In [7] van Rijsbergen proposed the use of a non-classical conditional logic for IR. This would enable the evaluation of $\Pr(d \rightarrow q)$ using the LUP, that was defined as follows:

> "Given any two sentences *x* and *y* a measure of the uncertainty of $y \rightarrow x$ related to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$."

This principle was the one of first attempts to make an explicit connection between non-classical logics and IR modelling. However, when proposing the above principle, van Rijsbergen was not specific about which logic and which uncertainty theory to use. As a consequence, various logics and uncertainty theories have been proposed and investigated. The choice of the appropriate logic and uncertainty mechanisms has been a main research theme in logical IR modelling leading to a number of different approaches over the years (for a detailed description of some of these models see [1]).

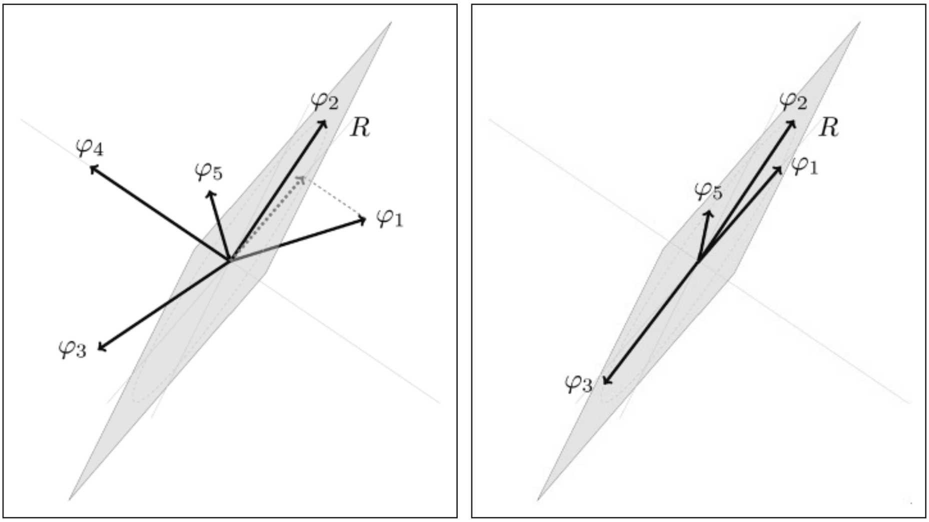
It turns out that the LUP is fully compatible with the cognitive framework. In fact, Ingwersen described uncertainty as "which and how much context needs to be added to retrieve those semantic values which provide the information searched for" [3, p. 36].

A broader view is provided by looking at the mathematics of Hilbert spaces and linear operators, as used in quantum theory (QT). This potentially combines different directions underlying the most important models in IR, namely geometry, logics and probability theory [8], as indicated in Figure 1. On the geometry side, there are vectors, subspaces and projectors. While these can be used to model many of the well-known retrieval approaches like the vector space model, Gleason's Theorem builds a bridge between geometry and a generalisation of probability theory. By applying this theorem, so-called density operators are able to induce a probability distribution on subspaces [8, ch. 6] which gives, if certain conditions (like vectors being normalised) are met, geometric approaches to probabilistic interpretations. Moreover, a conditional logic can be established based on the notion of subspaces [8, ch. 5], which together with Gleason's Theorem gives rise to a geometric interpretation of the LUP, seamlessly combining geometry, probability theory and logics. For example, if d and q can be represented as subspaces, one way of estimating $\Pr(d \rightarrow q)$ is using the trace function to compute an inner product between q and d [8, p. 96].

3 Quantum-Inspired Interactive Polyrepresentation Framework

As an example of a quantum-inspired approach which also facilitates polyrepresentation, we present an interactive geometrical framework called IQIR (Interactive Quantum Information Retrieval) [6]. Its aim is to address certain aspects of the cognitive framework, applying the mathematical formalism known from quantum mechanics, so that it supports interactive retrieval. In terms of the Venn diagram in Fig. 1, the IQIR framework lies in the intersection of probability theory and geometry.

The core assumption of the IQIR framework is that there exists a Hilbert space $\mathcal{H}$ of so-called *pure information needs* (pure INs). This resembles the idea from quantum mechanics, where physical systems are represented as vectors in a Hilbert space. Each such system can be in a certain state, which is indicated by a *state vector* φ, a unit vector in the corresponding Hilbert space. In the IR case, φ represents the system's view of the user's information need. It can be shown that a state vector induces a probability distribution on the subspaces of $\mathcal{H}$ [8]; each such



a) Before observation *b) After observation*

Fig. 2: Example mixed state. The dashed arrow shows the projection of φ_1 *onto* R.

subspace *R* represents an event, for instance the event that we measure a certain physical property (like the location of a particle) or, more IR-related, the event that a document *d* is relevant. Given a state φ, we can now compute $\Pr(R|d, \varphi)$, the probability that *d* is relevant given the current system state, as the square of the length of the orthogonal projection of φ onto the subspace *R*.

Usually a system cannot know precisely the information need, thus there will always be uncertainty about the user's intention (as a result of the cognitive free fall and for other reasons). In quantum mechanics, we face a similar problem as there is often uncertainty about the state a physical system is in, so it is assumed that the system is in a certain state with a given probability. This means that we deal with a set of state vectors and associated probabilities (the so-called *ensemble S* of states); the system is said to be in a *mixed state* if there is more than one possible state vector. An example can be seen in Fig. 2a, where we find five possible state vectors, each of them has a probability assigned that the system is in the respective state (omitted here). Transferred to an IR scenario, the system would assume the user to have one of the 5 different pure information needs each represented by a corresponding state vector. To compute the probability of relevance $\Pr(R|d, S)$, here a generalisation of the law of total probability is applied using the squared length of the orthogonal projection of each state vector onto *R*, which in this example represents document relevance as a 2-dimensional subspace.

As discussed in [3,4], information needs are usually unstable. To reflect this situation, in the IQIR framework, we can borrow another notion from quantum mechanics, the measurement postulate, which says that each observation of an event usually implies a state change. In interactive IR, such events can, for instance, be the submission of query by the user or a relevance judgement. Given that we can describe such an event as a subspace in our Hilbert space, the realisation of the event would cause the ensemble of state vectors to be projected onto the corresponding subspace and to be renormalised. Suppose in the example in Fig. 2a the event that a user judges a document as relevant is represented by the subspace *R*. Fig. 2b describes the situation after this event was observed by the system - the state vectors are now all within the 2-dimensional plane determined by *R*, and one vector, φ_4 even disappeared as it was orthogonal to *R*. Two effects result from this kind of dynamics. First, the system gains more certainty about the user's IN as now all state vectors are bound to a lower-dimensional plane, and some state vectors even disappeared. Second, slight shifts in information needs are supported as well, for instance if the vectors in Fig. 2b are later projected onto another 2-dimensional subspace which is similar to *R*.

The IN space discussed so far is very abstract. There are several ways to create a concrete instantiation of the IN space. For example, in [5] it is shown how a topical IN space can be constructed in a standard term space well-known in IR, where queries are represented as state vectors and documents as subspaces. But we are not bound to topical IN spaces. In fact, to satisfy different aspects of information needs, one may need to refer to non-topical information (like user-given ratings). In [2] it is shown how different topical and non-topical representations can be expressed as IN spaces in the IQIR framework with the aim to support polyrepresentation of documents. These component spaces can be combined into a composite space by means of a tensor product to create the cognitive overlap. This composite space can be regarded as an IN space in its own right as discussed above, with all its properties regarding user interaction. It further allows for the system state to become non-separable; the system is said to be in an *entangled* state then. Entanglement and composite spaces are further important mechanisms borrowed from quantum mechanics. In our case, entanglement can be used to express possible interdependencies between different representations in a polyrepresentation scenario.

Acknowledgements. We thank Benjamin Piwowarski and Birger Larsen for their valuable comments.

References

1. Fabio Crestani, Mounia Lalmas, and Keith van Rijsbergen. *Information Retrieval: Uncertainty and Logics*. Kluwer Academic Publisher, Norwell, MA, USA, 1998.

2. Ingo Frommholz, Birger Larsen, Benjamin Piwowarski, Mounia Lalmas, Peter Ingwersen, and C J van Rijsbergen. Supporting Polyrepresentation in a Quantuminspired Geometrical Retrieval Framework. In *Proc. IIiX 2010*, New Brunswick, 2010.
3. Peter Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52:3–50, 1996.
4. Peter Ingwersen and Kalvero Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
5. Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith van Rijsbergen. Exploring a Multidimensional Representation of Documents and Queries. In: *Proceedings of the 9th RIAO Conference (RIAO 2010)*, Paris, France, 2010.
6. Benjamin Piwowarski and Mounia Lalmas. A Quantum-based Model for Interactive Information Retrieval. In: *Proceeedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR 2009)*, pages 224–231, Cambridge, UK, 2009.
7. C J van Rijsbergen. A Non-Classical Logic for Information Retrieval. *The Computer Journal*, 29:481–485, 1986.
8. C J van Rijsbergen. *The Geometry of Information Retrieval.* Cambridge University Press, New York, NY, USA, 2004.

Addresses of congratulating authors:

Ingo Frommholz
Department of Computing Science
University of Glasgow, Scotland
Email: ingo[at]dcs.gla.ac.uk

Keith van Rijsbergen
Department of Computing Science
University of Glasgow, Scotland
Email: keith[at]dcs.gla.ac.uk

Fabio Crestani
Faculty of Informatics
University of Lugano, Switzerland
Email: Fabio.crestani[at]usi.ch

Mounia Lalmas
Department of Computing Science
University of Glasgow, Scotland
Email: mounia[at]dcs.gla.ac.uk

Selecting Search Keys in IIR Tests: Is There a Label Effect?

Jaana Kekäläinen

University of Tampere, Tampere, Finland

Abstract. In interactive information retrieval experiments, subjects typically test retrieval systems or interfaces. An information need, triggering a search process, is given to the subjects because the test environment and setting may not allow the subjects to search on their own topics. The form of the information need descriptions, often known as tasks, vary according to the level of abstraction. This may affect the search key selection in the experiments and even induce a label effect; namely a detailed search task description may cause a loss in the variability of queries. We analyze queries of 124 subjects of an interactive retrieval experiment and compare their search keys to the words of task descriptions. The results show that there is a substantial overlap and the task description may induce the label effect.

Keywords: Label effect, Interactive information retrieval experiment, Search key selection, Information need, Simulated work task, Search task, Search topic

1. Introduction

Information retrieval research emphasizes evaluation as a part of the research process. This is probably due to the practical nature of the area. In recent years, user-centred evaluation has gained popularity beside the traditional IR evaluation, which is based on test collections without interaction. This trend is most welcome because the traditional evaluation model has neglected many important aspects of information retrieval, or more broadly, information access.

The interactive information retrieval (IIR) evaluation comprises a range from field studies to tests in a laboratory environment; the former type offers more realism, and thus more generalizability, the latter more control on variables, and thus more explanation power. Laboratory tests, or IIR experiments, are typically executed to evaluate retrieval methods, systems or user interfaces, which cannot be tested in an operational environment. Such a test may involve from few to dozens of subjects, search topics or themes and one or several search environments.

In many IIR laboratory test settings, the subjects search information in experimental systems. Query formulation is crucial for the outcome. Therefore the search topic, or the information need behind it, is an important factor. How is the information need mediated to the subjects? In the early days of IIR research the subjects were given topics from non-interactive test collections (e.g. the first interactive track in TREC, see [13]). Later on Borlund and Ingwersen [2], also Borlund [1], suggested a new way to introduce more realism into laboratory type tests. This method, known as the *simulated work task situation*, was soon adopted widely in the interactive information retrieval (IIR) research (see e.g. [5],[6],[14])[1].

The simulated work task is a description of a work situation where the actor has an information need [1], [2]. The purpose of the task is to give the test subject a context for searching instead of giving a plain search topic. The subject should be somewhat familiar with the work task type, so that he can imagine himself in the task situation. The context of the simulated work task adds realism into the search task because it gives situational relevance conditions (see [3], [11]) for searching in addition to topical relevance.

As researches have adapted the simulated work task method in their test settings, there has often been confusion about the work task and the search task, the former being sometimes neglected. This tends to simplify the task to resemble a search topic. In such a case, the subjects may miss the personal interpretation of the context and purpose of the work task, and merely conduct a search based on the search task description. The difference between an information need, a search topic – often referred to as a request – and a query is important regarding the experimental setting. To summarize, in IIR experiments where the subjects search information, an information need or a search topic is given to the subjects in varying formats: as a plain topic, as a search task or as a search task embedded in a simulated work task.

In 1982 Ingwersen introduced the concept *label effect* [9]. In *Information Retrieval Interaction* Ingwersen describes the label effect as follows [8, p. 299]: "The phenomenon that *request* formulations may often consist of one or several concepts which are of a more general nature or out of the context which constitutes the real information need." We stretch the scope of the concept to explore the search key selection in an IIR experiment. More specifically, we hypothesize that the search topic or search task descriptions influence search key selection in the query formulation phase. We call this the label effect although the original concept was devoted to situations where the searcher has his own information need. In IIR experiments the subjects are given the information need and their unfamiliarity or uncertainty with the need leads to choose the search keys from the description.

[1] These are a few examples; the number of studies employing simulated search tasks is much greater.

However, there is a clear analogy to the original meaning: Ingwersen [8, p. 117] states that the label effect is typical for "*muddled topical needs* … i.e. the user wants to explore some new concepts or concept relations *outside known* subject matter." The subjects in IIR experiments are more or less searching outside the known subject matter because the information need is given.

In the present study we analyze queries from an interactive information retrieval experiment and compare the search keys of the queries to the words of the task descriptions. The aim is to find out to what extend the tasks influence the search key selection. In Section 2 our case and analysis methods are introduced. Section 3 presents the results. The last section discusses the findings.

2. Data and methods

We are interested in how an information need is described to the subjects of an IIR experiment and what is the relation between the description and queries formulated by the subjects.

2.1 IIR experiment at INEX 2004

Initiative for the evaluation XML retrieval (see [7], [4]) has attracted research groups around the world to participate in the development of the retrieval methods for structured documents. INEX has several tracks ranging from ad hoc retrieval to interactive retrieval. In the present study, we focus on the 2004 interactive track.

The aim of the track was to investigate how searchers interact with the components of XML documents [12]. This was done through an experiment where subjects were given modified topics from the ad hoc track and asked to search on the topics with a single retrieval system provided by the INEX organizers. Participating research groups enrolled at least eight subjects who were given two topics each. The test collection consists of 12,107 scientific articles from IEEE Computer Society's publications in XML format. The collection affected the topics and also the choice of subjects. During the experimental procedure, the subjects were to answer questionnaires concerning their background, familiarity with the topic, satisfaction with the search results, and some other details of minor concern for the present study.

In the track report the information need descriptions are referred to as 'tasks'. There were four tasks falling into two information need types: background category (B) – classic topical search – and comparison category (C) – search for differences between x and y [12]. The subjects selected one task from both categories. We analyzed the most popular tasks from each category further. The exact formulation of these tasks is given in Appendix. The two tasks differ regarding the specificity of the description. Task C2 is more like a simulated work task whereas task B1 resembles more a search task.

2.2 Analysis of the queries

The queries formulated by the subjects were collected into a log file. We analyzed all the queries of all the participants who had searched on the tasks B1 and C2. The number of subjects who chose the B1 task was 54, and 67 subjects chose the C2 task. Time allocated per task was 30 minutes. The number of queries for B1 was 292, and 460 for C2. Further details about the queries are given in Table 1.

Task	Subjects	Queries	**Queries/ subject** (stdev)	# Search keys	**Search keys/ query** (stdev)	# Search concepts	**Concepts/ query** (stdev)
B1	54	292	5.4 (3.6)	933	3.2 (1.5)	851	2.9 (1.4)
C2	67	460	6.9 (3.5)	1538	3.3 (1.7)	1438	3.1 (1.6)
Both	121	752	6.2 (3.6)	2471	3.3 (1.6)	2289	3.0 (1.6)

Table 1. Details of query data

We adapted the three level model suggested by Järvelin [10] for the comparison of queries and tasks. The three levels of the model are the concept, expression and occurrence level. Of these, we employed occurrences and concepts. Occurrences are character strings separated by spaces; generally they correspond to word forms. For simplicity, we refer to occurrences as search keys (in queries) and words (in tasks). First, all search keys from each query were identified and compared with words appearing in the corresponding task. Then the proportion of the search keys of the query appearing also in the task was calculated. We call this proportion an *overlap* between the query and the task. In other words, let Q be a set of search keys in the query, and T be a set of words in the task. Then, the overlap was calculated as:

$$(|Q \cap T|)/|Q|$$

The overlap is asymmetric because queries have less search keys than tasks have words; thus it is reasonable to calculate the overlap from the perspective of the query.

At the concept level, word form normalization was executed. That is, single and plural forms of a search key/word were conflated (*treatment - treatments*), as well as different tenses (*develop – developed*). Also obvious misspellings were corrected to their proper form (*lanuage – language*), and spelling variations unified (*sideeffect – side effect*). Phrases, marked with quotes in queries, were considered as concepts. Further, search keys and task words were conflated into the same concepts according to the following rules:

1. a synonym for the word appearing in the task (*advantage – benefit*)
2. a derivation of the word appearing in the task (*therapeutic – therapy*).

The overlap of concepts was calculated analogously to the calculation at the occurrence level; only search keys/word sets were replaced by concept sets.

Table 1 shows the total number of search keys and concepts in the queries, as well as the number of keys and concepts per query. The difference between the occurrence and concept level is not great. The identification of concepts in queries is problematic and we did not want to 'over-interpret' the intentions of the subjects (e.g. we considered only phrases marked with quotes), and thus the interpretation is conservative. The number of queries per subject varies from 1 to 21; the average is higher for the C2 task, which is obviously more difficult. The average number of search keys and concepts per query is rather steady.

3. Overlap between search keys and task words

We report the overlap figures at the occurrence and concept levels for all queries, and for the first and last queries. Table 2 shows the asymmetric overlaps between the queries and tasks. The overlap at occurrence level varies from 0.75 to 0.82, which is considerably high. At the concept level the overlap still increases, which is to be expected. Out of 752 queries, the overlap is 1.0 for 404 queries at the occurrence level, and for 548 queries at the concept level. These results provide evidence for the label effect.

Task	# Queries	**Overlap at occur. level** (stdev)	**Overlap at concept level** (stdev)
B1	292	0.75 (0.31)	0.81 (0.30)
C2	460	0.82 (0.25)	0.90 (0.23)
Both	752	0.79 (0.27)	0.87 (0.26)

Table 2. Overlap between queries and tasks at occurrence and concept level.

The overlaps are slightly higher for the task C2 than for the task B1 although C2 has more work task flavour. Obviously, there is not much variation in naming the two basic concepts in C2: *Java* and *Python*. Other concepts are more auxiliary in nature and not always helpful in queries (*development, large application, comparison, efficiency*). In B1, the word given in the task for one of the main concepts, *cybersickness*, is not the only or the best search key for the concept.

The average over all queries favours subjects with many queries. In the course of interaction there might also be changes in the features of the queries. Therefore we analysed the first and last queries of each subject separately. Table 3 shows that there is a change in the course of interaction: the overlap between the last queries and tasks is minor compared to the overlap of the first queries and tasks. The difference in the overlaps of the first and last queries is statistically significant (t-test, p<0.001).

How did the queries evolve? Search keys were deleted, added, misspellings corrected and single search keys combined into phrases. Most interesting here is the adding of new search keys: In the second and later queries the percentage of search keys not present in the task is 22; the percentage of such words in the first queries is 13. Obviously, the seen result documents had an impact on query formulation.

4. The effect of the description

We analyzed the overlap between the search keys of queries and the words of task descriptions, all originating in one of the IIR experiments of INEX. The results reveal that from 75 to 82 % of the search keys of the queries can be found in the task descriptions. There was, however, a difference between the first and last queries: the first queries had more overlap with the task than the last queries.

Task	# Queries	Occurrence level		Concept level	
		First query (stdev)	**Last query** (stdev)	**First query** (stdev)	Last query (stdev)
B1	54	0.88 (0.20)	0.67 (0.34)	0.93 (0.19)	0.76 (0.32)
C2	67	0.88 (0.19)	0.82 (0.26)	0.95 (0.16)	0.86 (0.28)
Both	121	0.88 (0.20)	0.75 (0.31)	0.94 (0.17)	0.82 (0.30)

Table 3. Overlap between first queries, last queries and tasks at occurrence and concept level.

Our case data are several years old, yet the basic experimental setting is typical for IIR experiments: the collection at hand affects search topic selection, probably more than the expertise of the subjects. In such situations, the subjects meet information needs they may not be familiar with. Their natural, and only, starting point for the search process is the task description given to them. Therefore the description is critical for the outcome of the experiment. The information need may be embedded in a search topic, in a search task or in a work task encompassing a search task. These all differ with respect to the context they offer for the subjects to build on. Also, there may be variation in the specificity of the description of the search task embedded in the work task; indeed, the search topic may be stated explicitly or the work task may be given at such an abstraction level that the subject has to create the information need(s).

A counterargument could be that the number of ways any concept can be expressed is limited, and if the main concepts of the topic are described in the work task, their most likely linguistic expressions are already given. Further, one may argue that the search topic has to be more or less fixed in order to restrict too large variation in queries, or searching in general; in other words to control the variables for the sake of the experiment. Yet, if we fix the search topic, do we need subjects? If search keys originate from tasks, why not simulate interaction?

Simulation is tempting for experimenters. However, as our case shows, queries evolve during the interaction to some extent. Simple search key selection from the task description is not enough for simulating interaction. The IIR experiments should do better but the experimental setting has pitfalls: If the task evokes the label effect by encompassing a too enforcing search task description, the subjects are likely to select search keys from the description and act similarly. As a consequence, their queries resemble automatically generated queries, and the experiment outcome is more likely to confirm the traditional, non-interactive laboratory test results. More realistic work tasks with less explicit search tasks may give more reliable information about interaction.

References

1. Borlund, P.: Evaluation of Interactive Information Retrieval Systems. Doctoral thesis. Åbo Akademi University Press, Åbo (2000)
2. Borlund, P., Ingwersen, P.: The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. J. Doc., 53, 225--250 (1997)
3. Cosijn, E., Ingwersen, P.: Dimensions of Relevance. Inf. Process. Manage. 36, 533--550 (2000)
4. Gövert, N., Kazai, G.: Overview of the Initiative for the Evaluation of XML Retrieval (INEX) 2002. In: Proc. of the 1st Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Schloss Dagstuhl, Germany, December 9-11, pp. 1--17 (2002)
5. Harper, D. J., Koychev, I., Sun, Y., Pirie, I. Within-Document Retrieval: A User-Centred Evaluation of Relevance Profiling. Inf. Retr. 7, 265--290 (2004)
6. He, D., Brusilovsky, P., Ahn, J., Grady, J., Farzan, R., Peng, Y., Yang, Y., Rogati, M.: An Evaluation of Adaptive Filtering in the Context of Realistic Task-Based Information Exploration. Inf. Process. Manage. 44, 511--533 (2007)
7. INEX, INitiative for the Evaluation of XML Retrieval, http://inex.is.informatik.uni-duisburg.de/
8. Ingwersen, P.: Information Retrieval Interaction. Taylor Graham, London (1992)
9. Ingwersen, P.: Search Procedures in the Library: Analyzed from the Cognitive Point of View. J. Doc., 38, 165--191 (1982)
10. Järvelin, K.: Merkkijonot, sanat, termit ja käsitteet informaation haussa [Strings, Words, Terms and Concepts in Information Retrieval]. Kirjastotiede ja informatiikka 12, 119--128 (1993)

11. Saracevic, T.: Relevance Reconsidered. In: Information Science: Integration in Perspective. Proc. of the 2nd Conference on Conceptions of Library and Information Science (CoLIS 2), pp. 201--218. The Royal School of Librarianship, Copenhagen (1996)
12. Tombros, A., Larsen, B., Malik, S.: The Interactive Track at INEX 2004. In: Advances in XML Information Retrieval. Proc. of the 3rd INEX Workshop. LNCS, vol. 3493, pp. 410--423. Springer, Heidelberg (2005)
13. TREC-6 Interactive Track Specification, http://trec.nist.gov/data/t6i/trec6spec (1997)
14. Villa, R., Cantador, I., Joho, H., Jose, J.: An Aspectual Interface for Supporting Complex Search Tasks. In: Proc. of the 32nd international ACM SIGIR Conference on Research and Development in information Retrieval, pp. 379--386. ACM, New York (2009)

Appendix

Task B1

You are writing a large article discussing virtual reality (VR) applications and you need to discuss their negative side effects.

What you want to know is the symptoms associated with cybersickness, the amount of users who get them, and the VR situations where they occur. You are not interested in the use of VR in therapeutic treatments unless they discuss VR side effects.

Sample queries

First: `VR cybersickness`
Last: `kennedy "simulator sickness"`

Task C2

You are working on a project to develop a next generation version of a software system. You are trying to decide on the benefits and problems of implementation in a number of programming languages, but particularly Java and Python.

You would like a good comparison of these for application development. You would like to see comparisons of Python and Java for developing large applications. You want to see articles, or parts of articles, that discuss the positive and negative aspects of the languages. Things that discuss either language with respect to application development may be also partially useful to you.

Ideally, you would be looking for items that are discussing both efficiency of development and efficiency of execution time for applications.

Sample queries

First: `java python application development`
Last: `"large application" development python java`

Address of congratulating author:

Jaana Kekäläinen
Department of Information Studies and Interactive Media
University of Tampere, Finland
Email: jaana.kekalainen[at]uta.fi

Search Procedures Revisited

Diane Kelly[1] & Ian Ruthven[2]

[1] University of North Carolina, Chapel Hill, USA
[2] University of Strathclyde, Glasgow, United Kingdom

Introduction

In this paper we pay tribute to our friend, colleague and mentor, Professor Peter Ingwersen, by examining one of our favorite of his papers, *Search Procedures in the Library – Analyzed from the Cognitive Point of View* originally published in Journal of Documentation in 1982 [4]. Like many of Peter's articles it is characterized by a strong theoretical basis that drives and informs empirical investigation, and includes thoughtful discussion of previous research in addition to the research findings.

Search Procedures reflects on a series of studies carried out over a four year period in the late 1970s. It was published at an interesting time for Information Retrieval. Written before Information Retrieval became synonymous with online information seeking it focuses on Information Retrieval within Public Libraries, then the major location for everyday information seeking. While many of his contemporaries focused on information seeking in academic or special library settings, Peter chose instead to focus a setting that was visited by a more diverse set of people with a broader range of information needs.

Search Procedures focuses particularly on the role of the librarian as an intermediary for finding information and the techniques used by intermediaries to understand a library patron's information need. However, already around this time Peter was demonstrating the foresight for which he is known: he predicted (prior to the Internet and Web search engines) that Information Retrieval machinery would become a mainstream technology and that end users would be required to learn how to navigate online searches without the assistance of intermediaries. If Information Retrieval was not to become an elite activity, as he described it in [5], then Information Retrieval interfaces would be required to capture something of the intelligent mediation he investigates in *Search Procedures* or Information Retrieval would become 'a kind of gamble.' [5, p472]. Fortunately, Information Retrieval did not become an elite activity but instead has become one of the most important and popular 'inventions' of the 20th century. Today, information search is a normal part of many people's daily routines and millions of searches are performed daily. While typical search engines are capable of some mediation through features such as spell correction and term suggestion, such mediations are quite rudimentary compared to the kind that Peter studied and are focused primarily on the query and search results, rather than the person and the information need.

In this article we summarize the main arguments of *Search Procedures* and, almost 30 years after it was written, reflect on its continuing value.

Search Procedures

Like many of Peter's articles, *Search Procedures* is informed by the Cognitive View of Information Retrieval. The Cognitive View is based on knowledge structures or individual cognitive models of parts of the world. Peter observed that each individual's image of the world consists of a '*conglomeration of different knowledge structures*' [4, p170]. This observation was to be the basis for his subsequent theory of poly-representation. Peter identified three major knowledge structures pertaining to the library intermediary: (1) structures around the professional library activities, such as knowledge of documents available for access, knowledge of how surrogates are created, knowledge of how to conduct standard search routines; (2) structures that reflect the librarian's conceptual or domain knowledge; and (3) knowledge structures that reflect the librarian's understanding of the library patron's stated information need and problem situation.

The Cognitive View is concerned with how these three knowledge structures can help mediate between the two other important sources of knowledge structures, those of the library patron who requires information and those of the document authors, which are reflected in the material available from within the library. *Search Procedures* investigates how the intermediaries negotiate these knowledge structures.

Employing a variant of the think-aloud protocol, the study investigates the information search procedures of 13 librarians conducting searches on written information requests and 5 non-expert searchers searching on their own information needs. The non-expert searchers conducted their own searches and only consulted with the librarians if they found no relevant material, leading to the negotiations which were studied. Peter uses the term 'search procedures', giving the paper its name, to reflect combinations of search actions that are performed within a problem-solving task as opposed to 'search strategies' which infer some conscious series of actions. The concentration is, therefore, on the unfolding cognitive reasoning involved in the mediation process as well as the behavioral actions that embody such cognitions.

A particular interest in this article was the creation of what Taylor [17] referred to as the 'compromised information need', a representation of the enquirer's information need. As Peter notes, '*the skill of the reference librarian is to work with the enquirer back to the formalized need…possibly even to the conscious need…and then to translate these needs into useful search strategy*' [4, p178]. That is, the process of negotiation is to help turn the enquirer's information need into a form that can be used to search the available information, given knowledge of how the information has been represented in the formal systems. This *labeling effect*, requiring enquirers to verbalize their information need into a search statement that may not reflect accurately their information need, is still the subject of much debate; see for example the recent work by Nicolaisen [10]. In *Search Procedures*, Peter does, however, take the position that the labeling effect can misrepresent the actual information need and the role of the intermediary should be to elicit the true information need by a carefully structured dialogue. The label, Peter emphasizes, may be well outside the context of the searcher's real need and the role of the intermediary is to find the right context. Thus, we see in *Search Procedures* an early recognition of the importance of context, a persistent theme throughout Peter's work.

Search Procedures notes that there is not one single patron-intermediary dialogue that is appropriate for all situations. It may be the case, for example, that the librarian is a domain and search expert and, in this situation, will take the lead in the dialogue with the enquirer filling in details. This type of dialogue is referred to as *asymmetrical*. Alternatively, the librarian may be an expert in search but have low knowledge of the search domain, in which case the dialogue is likely to be more *symmetrical* between the patron and librarian. Interestingly, Peter observes that in some cases librarians engage patrons in asymmetrical dialogue because they have too much confidence in their own understandings of patrons' information needs, essentially short-circuiting the process. This, in particular, is a danger when an emphasis is put on speed and least effort. Peter also observes that '*a conscious effort to keep the negotiation on equal-footing would improve the user's chances to provide useful insertions*' [4, p182].

Search Procedures shows that librarians use both open and closed questions to actively build a conceptual understanding of the enquirer's need with concepts being introduced, analyzed, retained or deleted until a suitable understanding emerges that can be used to interrogate the documents. This is described as a type of problem-solving. A surprising feature of the negotiations studied was the low use of 'open' questions: questions that start with 'Why, How, Where,' which should lead to useful information about the context of the information need. Peter's analysis points to the strengths and weaknesses of open questions within the mediation approach as studied: the low use of open questions can limit the enquirer's ability to introduce new concepts and important situational information, whereas overuse of open questions can risk overloading the librarian's original understanding of the need with too much information.

Far more common were 'closed' questions, which Peter divides into *normal closed questions* and *leading closed questions*. Normal closed questions lead to yes or no responses, while leading closed questions present the librarian's expectations about the searcher's answer. In symmetrical dialogue, closed questions can either confirm the librarian's initial understanding of the enquirer's information need or open up more specific or newer search directions. However, in asymmetrical dialogue when the librarian has a false sense of the enquirer's information need, closed questions can make it difficult to redirect the search.

In the study of librarians searching for the written information requests, Peter points to two search *motives*; what he called 'open search motives' which were actions relating to the discovery of information which could be used to enable the search process and 'fixed search processes' which were actions intended to discover documents themselves. Both motives related to *expectations* of whether a document answering the request may exist at all or what form the answer may be present within documents. Interestingly, different librarians engaged in different modes of operation based on these characteristics: some librarians operated an 'open search mode' where they used heuristic processes to expand their cognitive structures and, then, later in the search, moved towards solving the information problems. In this case the librarians were open to learning throughout the information interaction and often found relevant information later in their searches.

Other librarians were characterized by 'fixed and semi-fixed search modes' where the search routines were more algorithmic in nature with fixed modes describing librarians attempting to immediately retrieve documents that provide answers. This infers a fixed expectation about what answer may be appropriate and in what form the answer will appear. In some cases, semi-fixed modes were employed where librarians would move from fixed mode to open mode for parts of a search, to re-orientate a search, and then move back into a fixed mode. Noticeably open search modes resulted in more search concepts being introduced from the process of searching and less reliance on the written information request.

Revisiting the Findings

How does a study of Public Librarians performed in the late 1970s help Information Retrieval in the 2010s? Firstly, we should note that the relevance of the approaches and debates in *Search Procedures* have persisted since it was published – Michel's [9] paper investigating sources of information used by searchers during cognitive processing, Lin et al.'s paper on the value of reference interviews for understanding information needs [8], Kelly and Fu's work on using open questions to elicit better descriptions of searchers' information needs and contexts [6] and most recently, Nicolaisen's paper on compromised information needs [10] – are only four examples of work that have extended the lines of enquiry developed in *Search Procedures*.

Secondly, the work in *Search Procedures* illustrates the important distinction between what a searcher wants, that is, his information need, and what a searcher says he wants, that is, his search request. Within operational search systems, the consequences of the label effect is often tackled as a *corrective* feature; search engines ask for an initial search request and then offer functions to modify this request to something closer to the original information need using techniques such as relevance feedback or behavior modeling. What we see far less of in current search engine design are good approaches to obtain the best *initial* search request. Modern search engines, in general, support very well what Peter called fixed-mode searching: immediate access to documents that may provide an answer. There are strong reasons for this: a lot of search requests, such as finding a particular website, are naturally fixed-mode searches which do not require complex interactions; it is easier to develop algorithms for such searches; and, perhaps, giving immediate access to documents gives the impression that search success is close at hand.

In *Search Procedures* Peter points to the weaknesses of relying on fixed mode searching. In particular, fixed mode searching can result in relevant results being overlooked, possibly because the searcher has not been sufficiently informed by the search environment to make good relevance choices, and there appears to be a relative over-reliance on the original search request. This latter point is a frequent concern in Interactive IR evaluations where experimental participants often form queries based on written search requests rather than on the information need that lies behind the request.

Peter points to the fixed mode of searching being a mental attitude rather than a criterion for a good search solution. As he notes: '*It is not possible to rank the search modes in order to point out the qualitatively best one. No doubt each serves specific aims—as their dependence on the working domain shows. The main problem seems to be awareness of the search consequences they cause*' [4, p189]. That is, given a search request or information need, there may be a variety of ways in which we can conduct the search, each of which will have consequences in terms of search success or search satisfaction. Search modes are a choice and choices have consequences. Information Retrieval systems have typically not supported a reflective process into the choice or development of search strategies, providing little feedback on alternative search actions or encouraging searchers to consider the quality of their interactions. Or, as Peter concludes '*the user needs more assistance and better consciousness in these tasks*' [4, p189].

Peter further notes that '*For 'intelligent' online assistants employment of some kind of open search mode seems likely to be the most efficient, because it combines heuristic features in the beginning of the search process with more formal solutions later, using the search algorithms built into the IR system in a flexible way*' [4, p189]. This claim is interesting in light of what has occurred in online searching. The majority of search interface design, the 'intelligent' online assistance Peter predicted, has not supported open-searching but fixed-mode searching. The 'intelligence' has taken the form, not of information dialogues, but of complex statistical modeling of interaction to automatically change search procedures or offer limited forms of interactive query support.

Peter's observations about the types of dialogues that occurred – symmetrical and asymmetrical – also demonstrate a problem with the current interactions people have with search engines. By and far, such interactions are asymmetrical, with leading closed questions being presented to searchers in the form of term suggestion, spell correction and other query variants. Peter points out that the problem with asymmetrical negotiation is that "*far too little new information is exchanged because of the mode of questioning applied*" [4, p181]. Leading questions are compared to the type often asked during a patient-doctor interaction, where the doctor does not allow the patient to clarify his answers, but rather looks for complementary information that supports the initial diagnosis. The principle problem with search is that too much faith is put into the initial interpretation of the information need and subsequent interactions are used to force the need to fit this interpretation.

Revisiting the Method

The methods used in *Search Procedures* are also worth revisiting since they demonstrate a patience and carefulness that is often missing from contemporary investigations of information needs and search behaviors. Data were collected in a real-world setting and included the search procedures used by a mixed group: 13 librarians conducting searches on written information requests and 5 patrons searching on their own information needs. Audio recordings of searchers as they thought-aloud formed the principle data for the study. Peter also observed searchers, documenting their behaviors and actions. In addition, searchers engaged in something Peter called *self-confrontation*, where he and the searchers elaborated on the think-aloud recordings by making '*repetitive runs of the recorded tape immediately after recording, adding comments*' [4, p173]. These supplemental activities – observation and self-confrontation – were designed to enhance the accuracy and validity of the think-aloud data and reflect Peter's ever-present concern for capturing a holistic understanding of search behavior. The importance of this micro-level, intensive perspective on search behavior can be lost at a time when studies of massive search log data are common. However, more of these types of studies are needed if current Information Retrieval systems are to be responsive to searchers, the variety needs they bring to systems and the diverse contexts in which these needs arise and are addressed.

Variations of the methods used by Peter were later used in several studies that examined patron-intermediary interactions including studies by Belkin, et al. [1], Kuhlthau, et al. [7], Saracevic, et al. [13], Spink, et al. [15], and Wu and Liu [19]. The method itself was presented in several review articles of methods used in library and information science including Fidel [2], Harter and Hert [3], Nozomi [11], Saracevic [12], and Wang [18].

Conclusion

Recently, we see a *turn* in Information Retrieval and a re-recognition that searches are often complex processes requiring more cognitive support for searchers and more emphasis on understanding information needs. While Peter and other pioneers of interactive IR have claimed all along that queries are often impoverished representations of a searcher's information need, it has taken a while for IR research to catch-up and acknowledge that single queries often do not tell the whole story and very often even tell the wrong story. New developments in intellectual property searching, legal searching, literature based discovery, biomedical searching, and exploratory searching have demonstrated that only supporting 'fixed mode' searching with 'asymmetrical negotiation' is insufficient for complex search tasks and that very often people are trying to do more than find an answer to a routine question or navigate to a popular resource.

Web search engines are Information Retrieval's most visible success story: they are useful, efficient and we struggle to imagine how we coped without them. However, for many search situations we also need Information Retrieval tools that treat us like adults. That is, we need tools for those situations where we know that our information needs are complex and multi-faceted, where we know that we will need to engage in difficult cognitive work and where we do not expect our need to be satisfied within 1 second and with the exact same search results that were presented to the previous searcher who entered a similar query. The strength of *Search Procedures* is it recognizes that complex searching is a norm not an exception and that good design is not necessarily simpler and faster but more integrative, dynamic and symmetrical.

Suchman [14, p316] claimed that *'..interaction between people and computers requires essentially the same interpretative work that characterizes interaction between people'*. If we are interested in designing intelligent, useful tools for complex search problems then we can find guidance in studies of human-human information interaction of the type described by Peter in *Search Procedures*. In this paper, as in so many others by Peter, we see how research conducted from the cognitive perspective can support modern Information Retrieval.

References

1. Belkin, N. J., Brooks, H. M., & Daniels, P. J. Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies, 27*, 2, 1987, 127-144.
2. Fidel, R. Qualitative methods in information retrieval research. *Library & Information Science Research, 15*, 3, 1993, 219-247.
3. Harter, S. P. & Hert, C. A. Evaluation of information retrieval systems: Approaches, issues and methods. *Annual Review of Information Science & Technology, 32*, 1997, 3-94.

4. Ingwersen, P. Search Procedures in the Library – Analysed from the Cognitive Point of View. *Journal of Documentation. 38*, 3, 1982, 165-191.
5. Ingwersen, P. A cognitive view of three selected online search facilities. *Online Review. 8*, 5. 1984, 465-492
6. Kelly, D., & Fu, X. Eliciting better information need descriptions from users of information systems. *Information Processing & Management 43*, 1, 2007. 30-46.
7. Kuhlthau, C. C., Spink, A. & Cool, C. Exploration into stages in the information search process in online information retrieval: Communication between users and intermediaries. *Proceedings of the ASIS Annual Meeting, 29*, 1992, 67-71.
8. Lin, J., Wu, P., and Abels, E. Towards Automatic Facet Analysis and Need Negotiation: Lessons from Mediated Search. *ACM Transactions on Information Systems*, 27, 1, Article 6, 2008, 42 pages.
9. Michel, D. A. What is used during cognitive processing in information retrieval and library searching? 11 sources of search information. *Journal of the American Society for Information Science, 45*, 7, 1994, 498-514.
10. Nicolaisen, J. Compromised need and the label effect: An examination of claims and evidence. *Journal of the American Society for Information Science and Technology*, 60,10, 2009, 2004-2009.
11. Nozomi, I. Approaches in the studies of reference process and its integration: Focusing on the studies of reference interviews. *Library & Information Science, 30*, 1992, 43-58.
12. Saracevic, T. Modeling and measuring user-intermediary-computer interaction in online searching: Design of a study. *Proceedings of the ASIS Annual Meeting, 26*, 1989, 75-80.
13. Saracevic, T., Spink, A. & Wu, M-M. Users and intermediaries in information retrieval: What are they talking about? *Proceedings of the Sixth International User Modeling Conference,* 1997, 43-54.
14. Solomon, P. Conversation in information-seeking contexts: A test of an analytical framework. *Library & Information Science Research, 19*, 3, 1997, 217-248.
15. Spink, A., Goodrum, A., & Robins, D. Elicitation behavior during mediated information retrieval. *Information Processing & Management, 34*, 2-3, 1998, 257-273.
16. Suchman, L. A. Representing practice in cognitive science. *Human Studies.* 11, 1988, 305-325.
17. Taylor, R.S. Question-negotiation and information seeking in libraries. *College and Research Libraries, 29*, 3, 1968, 178-194.
18. Wang, P. L. Methodologies and methods for user behavioral research. *Annual Review of Information Science and Technology, 34*, 1999, 53-99.

19. Wu, M. M. & Liu, Y. H. Intermediary's information seeking, inquiring minds and elicitation styles. *Journal of the American Society for Information Science and Technology, 54*, 12, 2003, 1117-1133.

Addresses of congratulating authors:

Diane Kelly
School of Information & Library Science
University of North Carolina, 100 Manning Hall, CB#3360, Chapel Hill, NC 27599-3360
Email: dianek[at]email.unc.edu

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde, Glasgow, G1 1XH, United Kingdom
Email: ir[at]cis.strath.ac.uk

Simulations as a Means to Address Some Limitations of Laboratory-based IR Evaluation

Heikki Keskustalo & Kalervo Järvelin

University of Tampere, Tampere, Finland

Abstract. We suggest using simulations to address some of the limitations of test collection-based IR evaluation. In the present paper we explore the effectiveness of short query sessions based on a graph-based view of the searching situation where potential queries (query key combinations) constitute the vertexes of a graph G describing each topic. "Session strategies" are rules which determine the acceptable query reformulations. Query reformulations manifest as edges in G, and they express the allowed transitions between the vertexes. Multiple-query topical sessions manifest as paths in G. We present an example of this approach assuming session strategies based on limited query modifications (additions, deletions, or substitutions of few query words). We end by discussing the significance of our approach for IR evaluation.

1 Introduction

In their seminal book *The Turn* Ingwersen and Järvelin point out some of the main problems related to the laboratory-based IR evaluation, including the lack of modeling explicit users and tasks, and the lack of modeling interaction ([1]; see [2] for the original discussion). Recent studies suggest that in real life users typically prefer short queries, try out more than one query if needed [3-7] and often prefer making only small modifications to their queries [3]. Furthermore, even experts encountering the same task may use very different wordings in their searching. They may also consider finding only a few reasonably good documents as success [4]. Users also try to compensate for the performance deficiencies of the systems by adapting their search behavior [5, 7, 8]. The traditional Cranfield-style experiments based on one query per topic are not well-suited to study such behavior.

We suggest using simulations as a solution towards some of the limitations of Cranfield-style experiments discussed in Turn. By simulations we refer to experimentation based on using a symbolic model of a simplified real life search sessions in order to answer research questions. We assume multiple-query search sessions based on alternating querying and browsing phases. In the present paper, in particular, we will simulate search sessions assuming the shortest queries (including several one-word query versions for every topic). We allow several queries for a topic, assume limited modifications to the queries, and define success as being able to find one (highly) relevant document for a topic.

In other words, we restrict our attention to a simulation where short queries are used in various combinations in sessions. We assume that the searcher issues an initial query and inspects some top-N documents retrieved; if an insufficient number of relevant documents are recognized, the user repeatedly launches queries until the information need is satisfied or the user gives up.

The motivation behind our approach is that due to the costs involved during query formulation, the user may optimize the total cost-and-benefit of his sessions by rapidly trying out short queries. In other words, the user is willing to take chances with the quality of the result, and he is prepared to try out several short queries to see if something relevant is to be found.

Formally characterized, we utilize a graph-based approach in test collections explained in Section 3. In the experimental part of the study we will utilize the TREC 7-8 corpus with 41 topics having graded relevance assessments.

Next we will briefly review literature on user behavior and justify our approach. This is followed by defining our research problem. Section 3 explains the graph-based simulation method. Results of our experiments are given in Section 4. Discussion and conclusions are presented in Section 5.

2. The Significance of Multiple-Query Sessions

2.1 User behavior

Searchers behave individually in real life: their information needs may be unclear and dynamic as the users may learn as the session progresses, and the users may switch focus. In practice, a particular searcher may try out several queries during a search session, and different searchers may try out different wordings even when they face the same (well-defined) search task. It may be difficult for the searcher to predict how well a particular query will perform [8] because even assuming that the query does describe the topic well, it may be ambiguous [9] and therefore not retrieve documents serving the particular searcher in his searching context. Therefore, multiple query sessions are commonplace and may be unavoidable in practice in real life.

It has also been observed that real searchers often make use of very short queries and they prefer making small modifications to the previous queries. Jansen and colleagues [3] analyzed transaction logs containing thousands of queries posed by

Internet search service users. They discovered that one in three queries had only one term; two in three had one or two terms. On the average the query length was 2.21 terms per query. The average number of terms used in a query was even smaller, 1.45, in a study by [6] focusing on intranet users. Less than 4 % of the queries in Jansen's study had more than 6 terms. Because very short queries are commonplace, focusing on them in a test collection environment study seems justified.

Real-life searchers also avoid excessive browsing. They may stop browsing if the search result does not look promising almost immediately [10]. The stopping decisions regarding browsing the retrieved document list depend on the search task and the individual [4]. Jansen and colleagues [3] observed that most users did not access results past the first page presenting the top-10 results retrieved. Users may stop the search session after finding one or a few relevant documents. In particular, real searchers very rarely browse the top-1000 documents, although in some cases they do (e.g., patent searchers). Therefore, it is important to study situations where the search is successfully completed after only one or few relevant documents are found.

2.2 Motivation and research question

Generally speaking, valid instruments and study designs used to explain or evaluate some phenomenon should incorporate major factors affecting the phenomenon under study and systematically relate them to each other. We justify our present study design by the following observations. First, in real life users often:

- prefer very short queries (often only 1-2 keys)
- try out more than one query per topic, if needed
- cope by trying out limited modifications to queries
- avoid browsing a long list of documents, and
- stop after finding one or a few relevant documents

In traditional Cranfield-style experiments, it is common to (implicitly) assume fundamentally different kind of user behavior. These studies are typically based on using:

- longer queries (at least somewhat longer, e.g., even title queries typically have more than one word)
- one query per topic (and presenting the results averaged over topics)

Therefore, in the present paper we suggest modeling user behavior, in a test collection, but using:

- the very short queries
- several queries per topic
- limited word-level edit operations to modify queries
- shallow browsing, and
- one or a few (highly) relevant documents as the success criterion

Regarding the first two items, we will construct several alternative one-word query candidates, and slightly longer queries, for each topic. One way to approach searching is to use one-word queries as the starting points for sessions. Regarding the third item we assume that queries are modified by performing limited word additions, deletions, or substitutions.

Regarding the last two items, we assume that if any particular query within a session fails, the user will stop browsing almost immediately (N.B., this makes sense because the simulated user is aware that a short query attempt may very well fail).

If a query is successful, the user will stop searching after finding one (highly) relevant document. We use precision at 5 documents (P@5) as our primary success criterion and experiment with two separate relevance thresholds – liberal and stringent (see Section 2.3) [11].

A successful end result for any search session may require a different number of queries for individual topics. For one topic the first query candidate may be successful – as we will show - while for the next topic additional query candidates may be required.

Research question

Our overall research question in this paper is: *How successful are short queries as sessions when we assume limited query modifications, limited browsing and success defined as being able to find one (highly) relevant document?*

In studying this problem, we will assume that:

- the topical requests remain unchanged during a session - the simulated searcher neither learns nor switches focus during the session;
- the relevance of the documents for the simulated searcher is defined by the recall base of the test collection; and
- the simulated searcher scans the ranked list of documents from the top to bottom – behavior observed via eye-tracking [12].

2.3 The test collection and search engine

We used the reassessed TREC test collection including 41 topics from TREC 7 and TREC 8 ad hoc tracks [11]. The document database contains 528155 documents organized under the retrieval system Lemur. The relevance judgments are done on a four-point scale: (0) irrelevant; (1) marginally relevant: the document only points to the topic but does not contain more or other information than the topic description; (2) fairly relevant: the document contains more information than the topic description but the presentation is not exhaustive; and (3) highly relevant: the document discusses the themes of the topic exhaustively. In the recall base there are on the average 29 marginally relevant documents, 20 fairly relevant documents and 10 highly relevant documents for each topic [11].

2.4 Collecting the query data

All test topics were first analyzed intellectually by two sets of test persons to form query candidate sets. Our intention was to collect a reasonable set of query candidates together with user estimations regarding their appropriateness. During the topic analysis the test persons did not interact with a real system. They probably would have been able to make higher quality queries, if they had had a chance to utilize system feedback. However, this is no limitation to the method described in this paper.

We demonstrate here our graph-based method based on data collected from a group of seven undergraduate information science students. Regarding each topic a printed topic description and a task questionnaire were presented to the test persons. Each person analyzed six topics (one person analyzed five topics) thus 41 topics were analyzed. The users were asked to directly select and to think up good search words from topical descriptions; to create various query candidates; and to evaluate how appropriate the query candidates were.

The test persons were asked to form query versions of various lengths. We used the long query version requested to have three or more words as a starting point: first we selected its first three words A-C for each topic. To get the needed fourth and the fifth word we selected randomly distinct words from the remaining words in the long query version, or, if its words run out, from the other query versions requested from the users. Our goal in using the data collected from the test persons was to define a set of five query words for each topic. The procedure produced some obvious bad keys for topics (see Appendix) but this only makes our argument stronger - if the empirical results show that as sessions these words, tried as various combinations, often produce a rapid success despite some bad keys included.

3. Graph-Based Simulation

Our suggested procedure described next is inspired mainly by two main points: (1) real users cope with short queries, and (2) they prefer small query modification steps. In brief, our graph-based method to study multiple-query session effectiveness in a test collection consists of the following steps:

1. Words are collected to describe the test topics. Sources of data include using topic descriptions of test collections directly; utilizing test persons performing simulated or real tasks, etc. We asked test persons to create realistic content for short topical queries.

2. Query candidates are formed for each topic. We formed all possible word combinations (of 5 word) using the bag of words operator #sum of Lemur. However, queries may have some other structure, e.g., the #and or proximity operators. The basic idea is to create an extensive listing of possible query types (cf. [13]).
3. A search is performed using each query combination for each topic. We used the Lemur retrieval system in our experiment producing a ranked list of retrieved document, but other types of retrieval engines, e.g., Boolean systems, could be utilized.
4. Each distinct query is interpreted as a vertex of a (topical) graph.
5. The effectiveness results (regarding each distinct query) are expressed alongside the vertexes.
6. Sessions are now considered - in retrospect. To do this, we study the properties of the graphs.

To simulate sessions we need to (1) select start vertex; (2) determine the traversal rule(s); (3) define the stopping condition(s), and (4) consider the vertex traversal for each topic. For example

- One-word queries may be considered as start vertexes.
- "One word can be added/deleted/substituted at time" is one example of a traversal rule (a query modification rule).
- "Stop if 1 highly relevant document is found" is an example of a stopping condition.

7. The properties of sessions (paths) can be studied by using various effectiveness metrics.

If all word combinations are formed, their number increases rapidly as the number of keys increases. We limit our experiment to 5 query keys for each topic thus producing 25 graph vertexes.

Vertexes of the graph

In more detail, the simulation process goes as follows. First, the set of vertexes is formed for each topic. We assume unstructured *(#sum)* queries. Each distinct query (query key combination) constitutes one vertex $v_i \in V$ in a directed acyclic graph $G = (V, E)$. The query reformulations are reflected as edges ($e_j \in E$) in G and they express the allowed transitions between the vertexes. Multiple-query topical sessions manifest as paths in G. We have an ordered list of 5 query keys A, B, C, D, E available for each topic in our test data. These five keys produce 25 query combinations. In other words, 32 vertexes of the (topical) query graph are created (31 vertexes if the empty query is excluded). The vertexes are arranged in Table 1 into a diamond-shaped figure so that the number of keys increases in the query combinations from top to bottom.

{}									
A	B	C	D	E					
AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE
ABCD	ABCE	ABDE	ACDE	BCDE					
ABCDE									

Table 1. Query combinations (graph vertexes) arranged by the number of keys.

The figure consists of 6 rows - from top to bottom - one empty query vertex; 5 one-word vertexes; 10 two-word vertexes; 10 three-word vertexes; 5 four-word vertexes and one 5-word vertex. Top-1000 documents are retrieved using each query. For each individual topic the diamond-shaped graph below is formed, and the selected effectiveness values are computed for each vertex. Also the corresponding average figures over 41 topics (liberal relevance threshold) or a subset of 38 topics (stringent relevance threshold) may be computed.

For example, assuming an ordered list of individual query keys A, B, C, D, E, the vertex BC is used to denote an (unstructured) two-word query consisting of the second and the third query key. For example, the query keys A-E for topic #351 constitute an ordered set {petroleum, exploration, south, atlantic, falkland} (see Appendix). In this case the vertex BC corresponds to the query #sum(exploration south).

Edges of the graph

Based on literature, we hypothesize that topical query sessions are often constituted by implicit / educated / learned "moves" between the vertexes. Obviously, the user has to start somehow. We assume that the user proceeds from one vertex and moves into another (creating a directed edge) by applying some acceptable (albeit implicit) rules or heuristics. One such possible user rule would be – based on the principle of least effort - to allow word-edit operations that have a cost of one - compared to the previous query. Such a user would add, delete or edit one word compared to the previous query formulation. In other words, the user tries to cope with a situation by making small, incremental steps.

Session strategies

The success of various query sequences as topical sessions may be analyzed in relation to the start vertexes, the traversal rules, and the stopping condition:

- *Selection of the start vertex.* The effects of selecting the start vertex from some particular level of the graph may be immediately inspected.
- *Traversal rules.* We restrict our attention to consider traversal rules based on small modifications. According to [3] modifications to successive queries are done in small increments; it is common to modify, add or delete a search key.
- *Stopping condition.* As explained, in the present paper we consider the task of finding one (highly) relevant document.

Regarding the graph, we know the exact form of the query in each node (both the "identity" and number of words in it), and its success (measured, e.g., as P@5 using the stringent relevance threshold). We can perform retrospective analyses regarding query sessions after defining the traversal rules (how to move from one node to another) and the stopping condition (what constitutes success). Our purpose is to consider the concept of a session using the data in the graph in retrospect. The vertexes allow us to see what would happen assuming various session strategies and criteria for session success. The graph gives an overview of success assuming different types of queries (e.g., several alternative one-word queries).

4. Results

Next we will discuss three kinds of results. First, we show general results for P@5 values (averaged over topics) using two relevance thresholds (Tables 2-3). The cells in the figure correspond to the query combinations explicated in Table 1. Second, we concentrate on the case of highly relevant documents required. Table 4 shows the share of successful topics, i.e., when a particular query combination was successful in finding a highly relevant document in the top-5.

Last, we will study how successful small query modifications are within sessions (if the current query fails). This analysis needs to be performed topic by topic. Therefore, we first illustrate the results for one topic (Table 5), present the data as a binary phenomenon, and finally present session information for all topics as a binary map (Table 6)

Liberal relevance threshold

In Table 2 following general trend emerges: P@5 gets higher values when we move downwards (i.e., towards the longer queries) and towards left in the graph.

-									
13.7	13.2	3.9	3.9	4.9					
34.2	27.8	26.8	24.4	21.5	19.5	17.1	9.8	10.7	6.8
44.9	40.0	40.1	36.6	34.2	29.3	31.7	27.3	22.9	11.2
48.3	43.9	44.4	35.1	28.8					
49.8									

Table 2. Effectiveness (P@5) (%) averaged over topics (N=41) for the various query combinations (liberal relevance threshold). See Table 1 for the queries in each cell.

On the one hand, it seems that our one-word queries were a "bad call", because even in the best case (the first individual word selected for each topic) the P@5 figure is low (13.7 %). On the other hand it seems that we selected the query keys in the correct order: the first single words selected (the left-most keys) are, on the average, more successful than the last words (P@5 figure 4.9 % for the 5th individual keys). We next repeat the previous experiment but this time accepting only the highly relevant documents as success (Table 3).

Stringent relevance threshold

In Table 3 the same kind of pattern as in Table 2, only weaker, emerges. Again, obviously, basically it seems that we can state, regarding the query length, "the longer the better". Yet, a problem with the numbers in Tables 2 and 3 is that they are impossible to interpret regarding individual topical sessions. Because of this, we will next look at the number of topics for which (at top-5 documents) the queries succeeded. We count the share of topics, out of 38, for which at least one highly relevant document was found in top-5.

Failures become rarer as the queries get longer. This happens rapidly: by using two reasonable keys (e.g., any one of the combinations AB, AC, and AD) the user succeeds for slightly less than half of the topics (failures for 21, 24, and 22 topics corresponding to success in case of 45 %, 37 %, and 42 % of the topics). Interestingly, the distinction between the best 3-word and 4-word queries seems to disappear measured this way, and they are almost as successful as the 5-word queries.

We would like to draw the attention of the reader to the fact that it is not possible to interpret the data in Table 4 much more deeply without considering

-									
7.4	6.8	0.5	1.1	1.6					
13.7	12.1	11.1	10.5	4.7	9.0	7.4	4.2	4.7	2.1
15.8	14.7	16.3	16.8	16.3	10.5	11.6	7.9	10.5	6.3
17.9	16.3	17.9	15.8	11.1					
19.5									

Table 3. Effectiveness (P@5) (%) averaged over topics (N=38) for various queries (stringent relevance threshold).

queries as sequences, and regarding individual topics. For example, one may claim that queries of type E are generally inferior compared to the queries of type A. While this indeed is true, e.g., for the individual topic #351 query A fails but query E succeeds. Also in real life sometimes a (short) query succeeds, sometimes it fails. In that case the user may start reformulating queries. We will next enter into this territory through retrospective session analysis.

-									
24	21	3	5	8					
45	37	42	34	21	26	26	18	16	8
55	53	53	47	53	37	39	26	37	21
55	53	53	53	37					
61									

Table 4. The share of successful topics (%) for which at least one highly relevant document was retrieved at top-5. N=38 topics.

Sessions are next considered as traversals (paths) where the user continues the topical session and launches the next query if and only if any current query fails. We start by showing how to present the success of the component queries for one topic (#351).

Individual query example

Our analysis is limited by the assumption that the user considers only the set of words (5 in our case) available. Although we limit our experiments to 5 words, larger word sets could be used. However, it is not unrealistic to assume that a user may cope in a retrieval situation by indeed using a limited set of query keys. As our results show, if the user is able to invent one or two good keys, (s)he may succeed.

-									
0	0	0	0	0.2					
0	0	0	0.4	0	0	0.4	0	0	0
0	0	0.4	0	0.2	0.2	0	0.2	0.2	0
0	0.4	0.4	0.2	0.2					
0.4									

Table 5. Effectiveness (P@5) (%) for topic #351 ("petroleum exploration south atlantic falkland") measured at stringent relevance threshold, for various query combinations. 14 highly relevant documents exist for the topic. Legend: cells with a value above zero indicate success (+) and zeros indicate failure (-) for any particular query combination.

Table 5 allows studying, in retrospect, the effects of using various session approaches. We may analyze the general level of success through the number of words in queries and traversals via word-level substitution, addition and deletion.

Binary session map

Numbers in the graph vertexes in Table 5 can be interpreted as binary success (e.g., when at least one highly relevant document is found within the top-5, i.e., P@5>0) or failure (otherwise). By labeling the successful vertexes by a plus ('+') sign and the failed vertexes by a minus ('-') sign, information in Table 5 can be expressed in form of a character string:

```
#351 ----+ ---+--+--- --+-++-++- -++++ +
```

To make the diagram readable we arranged it into groups of 5, 10, 10, 5, and 1 symbol, corresponding to query combinations having one, two, three, four, and five query keys. By expressing the topical data this way for every topic a visual map is created. It gives information regarding the query combinations available for topical sessions based on a specific success criterion (Table 6).

#351	----+	---+--+---	--+-++-++-	-++++	+
#353	-----	--------+-	----+----+	---+-	-
#355	+++-+	++++++++++	++++++++++	+++++	+
#358	-----	-+++---+--	---+++---+	---+-	+
#360	-+---	+----++---	-++-----+-	--+--	-
#362	-----	--------+-	----------	-----	-
#364	+----	++++------	++++++----	++++-	+
#365	-+---	+-+-+++---	+++++++++-	+++++	+
#372	-----	-+--+--+--	+--++-+--+	++-++	+
#373	+----	-+++------	+--+++----	++-+-	-
#377	-+---	+---+++---	+++---+++-	+++-+	+
#384	-----	-----+-+--	-+-+--+-+-	+-+++	+
#385	-----	++++------	+++++++-+-	+++++	+
#387	-+---	+-++-++---	+++++++++-	+++--	+
#388	-----	----------	----------	-----	-
#392	-+---	------+---	--+----++-	-----	-
#393	---++	++++-+++++	++++++++++	+++++	+
#396	-+-+-	-++--+++--	++++++--++	++++-	+
#399	-----	----+-----	----------	-----	-
#400	++---	++--------	-++++-----	---+-	-
#402	-----	----------	----------	-----	-
#403	-----	+-+++-----	+++++-++-+	++-++	+
#405	-----	----------	--+-------	-++--	+
#407	+----	++++---++-	++++++++--	++++-	+
#408	-----	----------	------+---	----+	+
#410	+----	+++-------	+++++-----	++++-	+
#415	-----	----------	++----+---	+-+-+	+
#416	+----	-+++------	++-+------	+-++-	+
#418	+----	++++------	++++++----	++++-	+
#420	-----	-++++++---	+++++++++-	+++++	+
#421	-----	+---------	+---------	-----	-
#427	-----	--------++	------+-++	---++	-
#428	-----	+---+-----	+---------	++---	-
#431	+----	+-+-------	+++-++----	+++--	+
#440	-----	----------	----+-----	-----	-
#442	-----	----------	----------	-----	-
#445	-----	+----+----	+++---+-+-	+++-+	+
#448	-----	----------	----------	-----	-

Table 6. Binary session map for 38 topics and all query combinations. Legend: plus ('+') or minus ('-') symbols correspond to the 31 non-empty vertexes in the topical graph, traversed left to right, and rows traversed from top to bottom. Plus indicates success, i.e., P@5 > 0 (stringent relevance threshold) and minus indicates a failure (P@5 = 0).

In Table 6 the very first symbols of each group are especially interesting. For example, the first symbols of the first three groups represent, correspondingly, the queries of type A, AB, ABC. As the test persons were requested to express each topic by using three or more words, these three query types are formed from the very first words (left to right) as listed by the test persons. We will next briefly discuss the properties of one to three word queries in sessions.

One-word queries

Table 6 shows the success of one-word queries (the first group of five symbols in each line) in sessions. We can see that the very first single-word query ('A') succeeded for 9 topics (#355, #364, #373, …) (the first symbol of the first group). Assuming that the user started the session this way and in case of failure continued by trying out the second single-word query ('B')(substitution of the key), it succeeded for 6 additional topics (#360, #365, #377, …) (the second symbol of the first group). Assuming, that the user continued instead by adding one word ('AB'), it succeeded even better, for 10 additional topics (#360, #365, #377, …) (the first symbol of the second group). Obviously, there are limits for this one-word approach as in case of 17 topics out of 38 at least one of the one-word queries succeeded.

Two-word queries

If the session was started by trying out a two-word query (the first two words given by the simulated users: 'AB') it succeeds for 17 topics (#355, #360, #364, …) out of 38. Assuming that the user continues in case of failure by trying out the second two-word query ('AC')(substitution of the second query key), it succeeds for 6 additional topics (#358, #372, #373, …). For 21 topics every one-word query failed, but a successful two-word query can be found for these in 13 cases (#353, #358, #362, …).

Three-word queries

If the session was started by a three-word query (the first three words given by the simulated users: 'ABC') the session immediately succeeds for 21 topics (#355, #364, #365, …) out of 38. Assuming that the user continues, in case of failure, by trying out various substitutions and uses three-word queries extensively, (s)he will succeed for 11 additional topics (#351, #353, #358, …). In other words, at least one of the three-word queries succeeds for 32 topics.

We justify the binary view of success shown in Table 6 by the fact that in real life:

- query sessions have a limited length
- after any query, success or failure may be considered

- success/failure regarding the session may depend on the history of the session, all the retrieved documents collected so far, etc.
- success/failure may not be a binary thing, e.g., the retrieved set of relevant documents may have value of various degrees

Above, we studied a more limited case where:

- sessions have a limited length
- each query within a session succeeds or fails
- the session ends successfully whenever a query succeeds
- the session fails if none of its queries succeeds
- the criterion for binary success is defined as follows: finding one highly relevant document is counted as success (P@5 = 0.2, 0.4, 0.6, 0.8, or 1.0) for any one particular query combination for the topic. Note that the binary success criterion can be defined in many other ways, e.g., as P@10 > 0, using liberal relevance threshold.

Last, we will show the traditional average precision interpretation of the effectiveness of the query combinations (Table 7).

{}									
11.2	7.2	0.8	1.3	1.3					
18.5	15.1	13.1	10.5	7.7	11.2	6.5	4.2	4.3	2.7
19.1	18.7	15.9	15.7	15.3	11.1	12.1	8.1	8.8	5.8
21.1	20.3	17.0	16.4	11.6					
17.9									

Table 7. Non-interpolated average precision (%) for the various query combinations averaged over topics (N=38) (stringent relevance threshold, top-1000 documents retrieved).

Table 7 presents the non-interpolated average precision results based on the top-1000 documents retrieved (stringent relevance threshold). Very short queries appear as inferior compared to the longer queries.

5. Discussion and Conclusions

The list of limitations of Cranfield-style experiments discussed in *The Turn* suggests that the effectiveness of IR methods and systems should be evaluated through several short queries, and assuming multiple-query topical sessions, because such an approach better corresponds to real life IR. We suggested in this paper that a graph-based simulation allows *retrospective analysis of the effectiveness of short-query sessions*. We assumed that a set of alternative queries is available for each topic, and the simulated user may try them in various combinations. The effects of word-level modifications in sessions may be considered systematically (e.g., one-word additions, deletions and substitutions, or more expensive operations) using the graph-based approach.

Note that the shortest queries in our experiment differ from utilizing, e.g., title queries of test collections. In the test data only three topics had a title field containing one word (#364: rabies; #392: robotics; #403: osteoporosis); for 19 topics the title field had two words, and for 19 topics three words. We experimented by trying out, e.g., several one-word queries for each topic. If we use P@5 > 0 as the success criterion (one highly relevant document required), in case of 15 topics (out of 38) success is reached by either the very first one-word query candidate ('A'), or the second ('B'), if the first one failed.

Our approach offers an instrument for comparing IR system performance when we assume input from users who behave by trying out one or more queries, as a sequence, but which may be very short, ambiguous, or both. The graph form allows presenting alternative query versions and considering their systematic modifications. By using a binary success criterion (e.g., P@5 > 0) we may investigate what kind of an IR system should be rewarded. For example, assume an IR system which is able to disambiguate query keys, cluster documents, and offer distinct interpretations for the query key (e.g., jaguar) – to offer one document as a representative for each cluster. The binary success criterion rewards this kind of system, because one correct interpretation in top-5 suffices for success but the system is not rewarded for finding more than one relevant documents (unless the threshold is raised). An IR system performing well – measured this way – is interesting from the user's point of view, because real searchers do use ambiguous words as queries – even as single words. Note that a set of alternative topical queries are needed because in real life the users consider keys from among several alternatives.

Peter Ingwersen [14] identified a phenomenon called the Label Effect. He wrote that searchers tend to act a bit at random, to be uncertain, and not to express everything they know. Instead, searchers express what they assume is enough and/or suitable to the human recipient and/or IR system. They compromise their statements under influence of the current and historic context and situation. In addition, the label effect means that searchers, even with well-defined knowledge of their information problem, tend to label their initial request for information verbally by means of very few (1-3) words or concepts. This description fits well what other studies [3] [6] tell about searcher behavior in the Web or intranets. It also closely matches the simple query session strategies that we propose to simulate in the present paper. In other words, we propose simulation of searching under the label effect.

We focused on retrieval situations where the searchers take their chances by repeatedly trying out short queries. We used a very limited set of query keys in our experiments. However, in the future IR test collections can be extended so that the facets of the test topics and their expressions are suggested by test searchers. Furthermore, the expressions of the facets in the relevant documents can be recognized. This kind of data could be used for more extensive graph-based session simulations. Our initial results indicated that even one-word queries often bring rapid success if they are considered as sequences. We suggest that the effectiveness of IR systems and methods should be compared, in test collections, from this perspective in the future.

Appendix

The five query words corresponding to A, B, C, D, E in Figure 1 are listed below for 41 topics. Due to lemmatization sometimes one user-given key produced more than one word. Due to the limited number of distinct search words given for some topics, some keywords are repeated. For topics #378, #414, and #437 no highly relevant documents exist in the recall base.

```
#351: petroleum, exploration, south, atlantic, falkland
#353: exploration, mine, antarctica, of, research
#355: remote, sense, ocean, radar, aperture
#358: alcohol blood, fatality, accident, drink drunk, drive
#360: drug, legalization, addiction, drug, drug
#362: realize, incident, smuggle, incident, gain
#364: rabies, cure, medication, confirm, confirm
#365: el, nino, flood, drought, warm
#372: native, american, casino, economic, autonomy
#373: encryption, equipment, export, concern, usa
#377: popular, cigar, smoke, night, room
#378: opposite, euro, reason, use, refuse
#384: build, space, station, moon, colonize
#385: hybrid, automobile, engine, gasoline non, engine
#387: radioactive, waste, permanent, handle, handle
#388: biological, organic, soil, use, enhancement
#392: future, robotics, computer, computer, application
#393: mercy, kill, support, euthanasia, euthanasia
```

```
#396: illness, asbestos, air, condition, control
#399: undersea, equipment, oceanographic, vessel, vessel
#400: amazon, rainforest, preserve, america, authority
#402: behavioral, generic, disorder, addiction, alcoholism
#403: elderly, bone, density, osteoporosis, osteoporosis
#405: cosmic, event, appear, unexpected, detect
#407: poach, impact, wildlife, preserve, preserve
#408: tropical, storm, casualty, damage, property
#410: schengen, agreement, border, control, europe
#414: sugar, cuba, import, trade, export
#415: golden, triangle, drug, production, asia
#416: gorge, project, cost, finish, three
#418: quilt, money, income, class, object
#420: carbon, monoxide, poison, poison, poison
#421: industrial, waste, disposal, management, storage
#427: uv, ultraviolet, light, eye, ocular
#428: decline, birth, rate, europe, europe
#431: robotic, technology, application, century, th
#437: deregulation, energy, electric, gas, customer
#440: child, labor, elimination, corporation, government
#442: hero, benefit, act, altruism, altruism
#445: clergy, woman, approval, church, country
#448: shipwreck, sea, weather, storm, ship
```

References

1. Ingwersen, P. and Järvelin, K. (2005) The Turn: Integration of Information Seeking and Retrieval in Context. Heidelberg, Springer, 2005.
2. Kekäläinen, J. and Järvelin, K. (2002) Evaluating information retrieval systems under the challenges of interaction and multi-dimensional dynamic relevance. In CoLIS4, 253-270.
3. Jansen, M. B. M., Spink, A., and Saracevic, T. (2000) Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, Information Processing & Management, 36(2): 207-227.
4. Järvelin, K., Price, S. L., Delcambre, L. M. L., and Nielsen, M. L. (2008) Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, in Proc. ECIR'08, 4-15.
5. Smith, C. L. and Kantor, P. B. (2008) User Adaptation: Good Results from Poor Systems, in Proc. ACM SIGIR'08, 147-154.

6. Stenmark, D. (2008) Identifying Clusters of User Behavior in Intranet Search Engine Log Files. Journal of the American Society for Information Science and Technology, 59(14): 2232-2243.

7. Turpin, A. and Hersh, W. (2001) Why Batch and User Evaluations Do Not Give the Same Results, in Proc. ACM SIGIR'01, 225-231.

8. Swanson, D. (1977) Information Retrieval as a Trial-and-Error Process. Library Quarterly, 47(2): 128-148.

9. Sanderson, M. (2008) Ambiguous Queries: Test Collections Need More Sense, in Proc. ACM SIGIR'08, 499-506.

10. Lorigo, L., Haridasan, M., Brynjarsdottir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., and Pan, B. (2008) Eye Tracking and Online Search: Lessons Learned and Challenges Ahead. Journal of the American Society for Information Science and Technology, 59(7): 1041-1052.

11. Sormunen, E. (2002) Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In Proc. ACM SIGIR '02, 324-330.

12. Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005) Accurately Interpreting Clickthrough Data as Implicit Feedback, in Proc. ACM SIGIR'05, 154-161.

13. Pirkola, A. and Keskustalo, H. (1999) The Effects of Translation Method, Conjunction, and Facet Structure on Concept-Based Cross-Language Queries. Finnish Information Studies 13, Tampere, 1999. 40 p.

14. Ingwersen, P. (1982) Search procedures in the library analyzed from the cognitive point of view. Journal of Documentation, 38(3): 165-191.

Addresses of congratulating authors:

Heikki Keskustalo
Department of Information Studies and Interactive Media
FI-33014 University of Tampere, Finland
Email: heikki.keskustalo[at]uta.fi

Kalervo Järvelin
Department of Information Studies and Interactive Media
FI-33014 University of Tampere, Finland
Email: kalervo.jarvelin[at]uta.fi

Using Thesauri in Enterprise Settings: Indexing or Query Expansion?

Marianne Lykke[1] & Anna Gjerluf Eslau[2]

[1] Royal School of Library and Information Science, Aalborg, Denmark
[2] H. Lundbeck A/S, Valby, Denmark

Abstract. The paper investigates empirically two basic approaches how to use a thesaurus in information retrieval. The study is an experimental retrieval test comparing the performance of three search strategies: searching by controlled metadata derived from domain-specific thesaurus, searching by natural language terms, and natural language searching using domain-specific thesaurus for query expansion. The comparison shows that the performance is lower for searching based on controlled metadata compared to searching based on natural regarding recall as well as precision. Higher performance of the expanded queries indicate that it might be sufficient to base subject retrieval on natural language queries enhanced by a domain-specific thesaurus, or to base metadata indexing on rule-based automatic categorization using thesaural information.

1. Introduction

Controlled metadata has several roles in enterprise information systems. Metadata enhances the retrieval performance, provides a way of managing the electronic digital objects, help to determine the authenticity of data, and is the key to interoperability (Hunter, 2003).

Subject metadata may be applied manually by human examination, or the assignment may be partially or fully automated. Another solution to describe the features of documents is to extract terms from the documents algorithmically. These two basic approaches to represent the content, meaning, and purpose of documents are often called human, intellectual indexing and automatic, computer-based indexing (Lancaster, 2003). By human assignment indexing an indexer analyses the text and assigns metadata terms to represent the content. By automatic indexing words and phrases naturally appearing in the text are extracted and used to represent the content of the text. Thesauri are frequently used to support both indexing and retrieval methods. The metadata terms used in human, assigned indexing are commonly drawn from some form of controlled vocabulary such as thesauri to control synonyms and homonyms and ease the burden of searching. Natural language searching based on automatic, extracted indexing can be considerably improved through searching aids like thesauri by suggesting additional search terms, especially synonyms and narrower terms (Bates, 1986).

Research comparing the strengths and weaknesses of these two basic indexing and searching methods fail to provide conclusive results in relation to retrieval effectiveness (Anderson & Pérez-Carballo, 2001ab). There is a general recognition that the two approaches should be used in combination to obtain the best retrieval performance.

In this study we seek to expand upon previous investigations about the two basic approaches. We want to explore the performance in the context of subject retrieval in a work-place retrieval system. The empirical setting is a pharmaceutical research and production enterprise. Specifically, we want to provide information about the use of a thesaurus to improve, respectively, retrieval based on human assignment of subject metadata and to expand natural language queries. The study seeks to address the questions:

- How does a tailored, domain-specific thesaurus perform in information retrieval when used for automatic query expansion of natural language queries?
- How does a tailored, domain-specific thesaurus perform in information retrieval when used for retrieval based on human assigned metadata

The paper is structured as follows: section 2 provides background information about the two approaches to using a thesaurus in information retrieval and rationale for the study, section 3 presents the case study, section 4 the methodology and the last sections present and discuss the findings, followed by concluding remarks and recommendations for further research.

2. Rationale and related research

Only few studies compared the performance of humanly assigned metadata and automatically extracted keywords in the context of work-place retrieval systems. Stephenson (1999) investigated the effectiveness of metadata on a US agency's public access web site in answering 24 known-item queries formulated from real reference questions posed by members of the public to EPA librarians. She found that only eight queries retrieved a responsive answer when the metadata repository alone was searched, compared with 22 for full-text search. Precision was also found to be higher for full-text search. In a recent study Hawking & Zobel (2007) evaluated subject metadata on two Australian institutional web sites. The study showed that subject and description metadata performed worse compared to other types of natural language indexing terms: terms from the full text, title terms, and anchor text terms. Subject metadata only outperformed terms of the URL's. An additional analysis of a subset of queries showed that for metadata to be useful in search, it needs to be accurate, and to add something to the data that cannot be deduced from the visible text. The authors question whether it is possible to obtain appropriate metadata, whether subject metadata can be any more specific than text content or anchor text.

The fact that computers are clearly faster and cheaper compared with human indexers that, in turn, are costly and not necessarily produce better indexing, suggests that it is central still to explore the qualities of the two indexing methods; especially as the conditions for human indexing are changing with the introduction of networked information systems such as digital libraries, intranets and electronic document management systems (EDMS). In networked systems the indexing task is frequently distributed and shared by people (e.g. the author or assistant) with varied experience and knowledge about indexing.

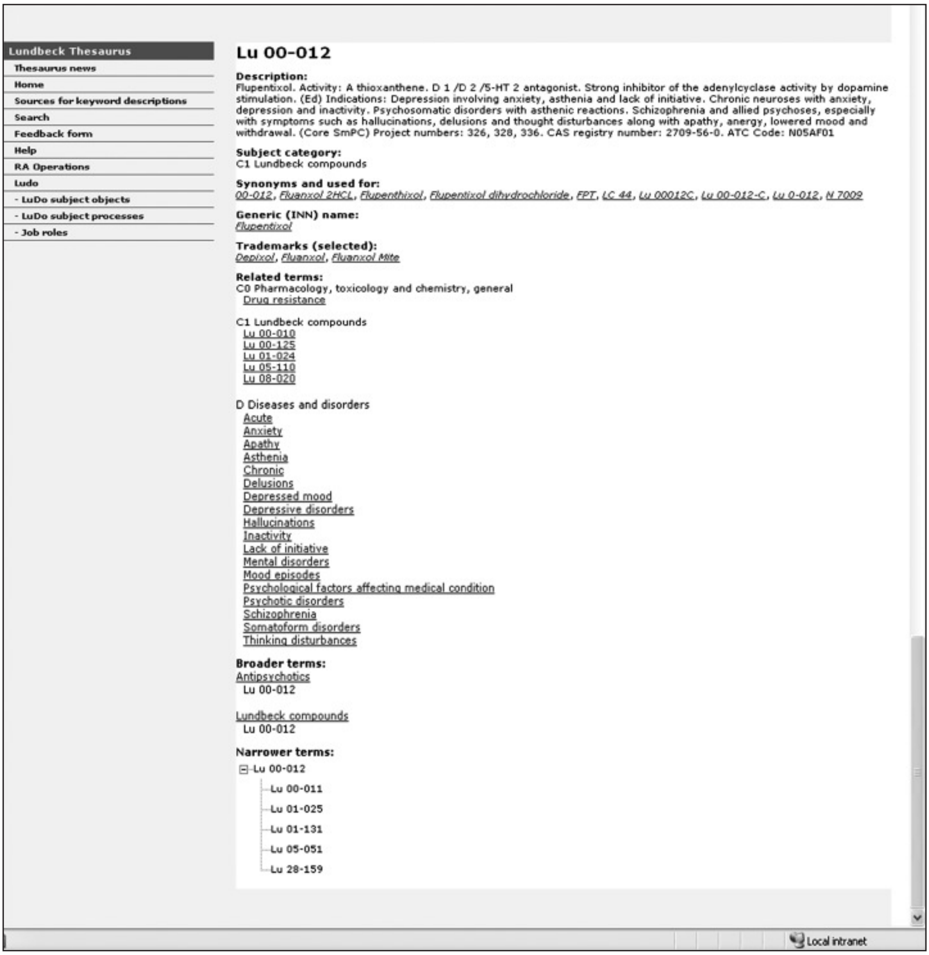
The most important disadvantage of automatic indexing is that it shifts more intellectual effort on the searcher compared with indexing using a controlled vocabulary (Lancaster, 2003). One of the means to meet these semantic problems and enhance searching based on automatic indexing is to expand the search query with additional search terms, e.g. synonym variations or other terms that are closely related to the search topic. Thesauri are one of the tools that are often used to assist searchers (Bates, 1986). A number of studies have investigated the thesaurus as a source for expansion of queries (Kristensen & Järvelin, 1990; Kristensen, 1993; Spink, 1994; Kekäläinen & Järvelin, 1998, 2000; Greenberg, 2001a, 2001b; Nielsen, 2004; Skov, Larsen & Ingwersen, 2006), Tudhope, Binding, Blocks & Cunliffe, 2006). These studies demonstrate the usefulness of thesauri both in term of providing users with alternative search terms and in obtaining improved retrieval performance. None of the previous studies compared retrieval performance of searches based on, respectively, humanly assigned subject metadata and computer extracted index terms expanded with thesaurus terms, but they demonstrate the potential of using a thesaurus for query expansion. Based on the findings it seems obvious to put the question whether thesauri are better used as support for retrieval than for human indexing that is costly, time-consuming, and by its nature, subjective and variable.

3. A pharmaceutical EDMS as a case study

The contextual framework of the study is the EDMS of a pharmaceutical research and production enterprise. The EDMS provides access to a range of document types, and metadata indexing is targeted to the information tasks of a well-defined set of knowledge workers within the department of research and development.

A varied group of indexers carry out the metadata assignment: authors, librarians, assistants, and research staff that act as coordinators for a group of researchers. All indexers have been trained, but the skills of the indexers are varied. Due to demands from regulatory authorities, recall is the most important performance measure. The indexing should retrieve as many relevant documents as possible; support retrieval of documents across units, work tasks, and document types.

The corporate thesaurus was established in 2001, and contains 16.000 terms, of which 5600 constitute the preferred, controlled terms (concepts) that are used for human indexing. The thesaurus provides semantic information about hierarchically broader and narrower terms, related terms, synonyms variations, and definition. The thesaurus structure is shown in Figure 1.



Lu 00-012

Description:
Flupentixol. Activity: A thioxanthene. D 1 /D 2 /5-HT 2 antagonist. Strong inhibitor of the adenylcyclase activity by dopamine stimulation. (Ed) Indications: Depression involving anxiety, asthenia and lack of initiative. Chronic neuroses with anxiety, depression and inactivity. Psychosomatic disorders with asthenic reactions. Schizophrenia and allied psychoses, especially with symptoms such as hallucinations, delusions and thought disturbances along with apathy, anergy, lowered mood and withdrawal. (Core SmPC) Project numbers: 326, 328, 336. CAS registry number: 2709-56-0. ATC Code: N05AF01

Subject category:
C1 Lundbeck compounds

Synonyms and used for:
00-012, *Fluanxol 2HCL*, *Flupenthixol*, *Flupentixol dihydrochloride*, *FPT*, *LC 44*, *Lu 00012C*, *Lu 00-012-C*, *Lu 0-012*, *N 7009*

Generic (INN) name:
Flupentixol

Trademarks (selected):
Depixol, *Fluanxol*, *Fluanxol Mite*

Related terms:
C0 Pharmacology, toxicology and chemistry, general
Drug resistance

C1 Lundbeck compounds
Lu 00-010
Lu 00-125
Lu 01-024
Lu 05-110
Lu 08-020

D Diseases and disorders
Acute
Anxiety
Apathy
Asthenia
Chronic
Delusions
Depressed mood
Depressive disorders
Hallucinations
Inactivity
Lack of initiative
Mental disorders
Mood episodes
Psychological factors affecting medical condition
Psychotic disorders
Schizophrenia
Somatoform disorders
Thinking disturbances

Broader terms:
Antipsychotics
Lu 00-012

Lundbeck compounds
Lu 00-012

Narrower terms:
⊟ Lu 00-012
- Lu 00-011
- Lu 01-025
- Lu 01-131
- Lu 05-051
- Lu 28-159

Figure 1: Record from the corporate thesaurus.

4. Research methodology

The evaluation framework was experimental. We developed ten realistic search tasks, standardized using the methodology by Borlund (2003). The researchers carried out the test searches, using a TREC style single query search strategy (Ingwersen & Järvelin, 2005). The queries were formulated with use of the terms that the original searcher in real life used to formulate the first search query. Each search session consisted of one search query that was put forward to the retrieval system with no further interaction between the test searcher and the retrieval system. We used the Verity K2 search engine for the full-text searches (IDOL K2, 2007), and performed the metadata searches in a Documentum based EDMS (EMC Software, 2006). The test collection consisted of 25,384 documents: some born in the EDMS, others scanned into the system and processed by optical character recognition (OCR). We calculated precision and recall in order to measure the retrieval performance.

4.1.1. Comparison of three search strategies

We searched the search scenarios by use of three different search strategies:

1. **Metadata search strategy, based on human, controlled indexing.** The searcher used controlled subject metadata from the corporate thesaurus to search the retrieval system.
2. **Simple natural language search strategy, based on automatic indexing**. The searcher used terms from the original search query to search the retrieval system.
3. **Advanced natural language search strategy, based on automatic indexing and with use of corporate thesaurus for query expansion**. The searcher expanded the simple natural language search query with additional terms drawn from the corporate thesaurus. The query was expanded by synonyms and narrower terms (including their synonyms) to the original search terms

We transformed the search jobs into search queries by dividing them into appropriate search concepts (facets). The facets were combined by the Boolean operator AND, and the expansion terms for each facet were combined by OR.

4.1.2. Data gathering

Precision and recall measured retrieval performance. Relative recall was calculated, based on the union search result of each search job (Kristensen, 1993). Calculation was based on 4-scaled relevance assessments (Sormunen, 2002). To avoid subjective judgments, we asked the relevance assessors to assess the documents retrieved according to the work task situation and the indicative request.

5. Findings

On average the test searches retrieved 77.5 unique documents per search jobs, ranging from 18 to 207 documents, (Table 1). Table 1 shows the number of documents retrieved for each of the ten search scenarios. The list divides the documents according to relevance score.

	Number of documents retrieved											
	SJ1	SJ2	SJ3	SJ4	SJ5	SJ6	SJ7	SJ8	SJ9	SJ10	All	All %
3 Highly relevant	8	3	12	11	18	1	0	16	0	7	76	10
2 Fairly relevant	1	2	6	45	7	3	1	9	8	2	84	11
1 Marginally relevant	10	6	2	105	3	4	3	7	0	19	159	20
0 Irrelevant	59	86	43	46	77	10	72	20	25	17	455	59
Total	78	97	63	207	105	18	76	52	33	45	774	100

Table 1: Number and relevance of documents retrieved for the ten search jobs (SJ1-10)

Table 2 and Table 3 summarise the results concerning search performance. Concerning recall the expanded natural language search performed best, followed by the simple natural language search. This applies to seven out of the ten search jobs. In four cases the search strategy based on metadata resulted in zero hits.

In one case, search scenario 7, the metadata search obtained better recall compared with the expanded search. In two other cases, the metadata search obtained better recall, compared with the simple natural language search strategy. In all cases this was due to the fact that the Verity K2 does not recognize search terms with slashes as prefix, e.g. ‘drug therapy/arrhythmia’.

The results reflect previous findings that none of the basic approaches how to apply a thesaurus in information retrieval perform convincingly better. In this case typographical formats challenge the performance of automatic indexing, as the search engine is sensitive to typographical formatting. Additionally, the OCR processing caused noise in retrieval by translating terms wrongly, for instance ‘lu 00-012’ to ‘lu 00-011’.

In general, the expanded search queries perform best. This is not surprising, but it is thought provoking, because the indexing policy, the metadata scheme, the indexing checklist, and the corporate thesaurus, are tailored specifically to meet information tasks as the ones investigated. The indexing policy instructs indexers specifically by the tailored checklist to index the subjects appearing in the ten search scenarios. The subjects represent the primary focus of pharmaceutical research, and to index these topics should be straightforward.

With respect to precision the simple natural language search strategy performed best in seven cases out of ten. In most cases precision is low, only in three cases was precision 50% or more. The metadata searches resulted in the lowest precision.

	Recall (%)									
Search strategy	SJ1	SJ2	SJ3	SJ4	SJ5	SJ6	SJ7	SJ8	SJ9	SJ10
Search stategy1, metadata	0	0	0	33	29	61	100	1	0	45
Search strategy2, NL	42	52	88	38	79	54	39	3	12	7
Search strategy3, NL expan.	100	95	100	88	89	100	39	100	100	77

Table 2: Recall measures per search job

	Precision (%)									
Search strategy	SJ1	SJ2	SJ3	SJ4	SJ5	SJ6	SJ7	SJ8	SJ9	SJ10
Search stategy1, metadata	0	0	0	39	15	24	3	33	0	30
Search strategy2, NL	12	13	34	50	43	29	1	33	68	50
Search strategy3, NL expan.	16	7	26	38	30	24	1	47	16	34

Table 3: Precision measures per search job

	Documents of relevance 3 (%)									
	SJ1	SJ2	SJ3	SJ4	SJ5	SJ6	SJ7	SJ8	SJ9	SJ10
Search strategy	n=8	n=3	n=12	n=11	n=18	n=1	n=0	n=16	n=0	n=7
Search stategy1, metadata	0	0	0	18	17	100	-	0	-	43
Search strategy2, NL	62	67	92	45	83	100	-	0	-	14
Search strategy3, NL expan.	100	100	100	100	94	100	-	100	-	71

Table 4: Documents of relevance 3 retrieved per search strategy

This finding is remarkable, as human indexers should be better at weighting the significance of subjects, and be more able to distinguish between important and peripheral compared with computers that base significance on term frequency. This is especially true in the context of enterprise retrieval, where retrieval is embedded in and targeted to specific information tasks. Compared with retrieval systems with a more general, broader scope it should be easier for a human indexer to interpret and relate document content to work domain characteristics. These results supports findings by Stephenson (1999); in the US agency case 16 out of 24 metadata searches resulted in zero hits and in low precision for the rest of the searches.

Additionally, it is unexpected that the expanded search strategy retrieves the largest number of highly relevant documents, with relevance score 3 (Table 4). It seems natural that the exhaustive expanded search retrieves the largest number of documents, and also relevant documents, but the fact that the expanded search retrieves more of the highly relevant documents is discouraging. From an indexing point of view, we could expect that the human indexing retrieves specifically the limited number of highly relevant documents.

6. Conclusion and future research

Being a case study these findings do not provide any conclusive results as to whether a thesaurus is better used as support for expanded natural language searches compared to controlled metadata searching. The study indicates that it is not easier to obtain high quality in human indexing in EDMS and workplace environments. In general, the findings do not change the general opinion that the two approaches should be used in combination. However, the high performance of the expanded natural language searches indicate that it might be feasible to base subject retrieval on automatic indexing enhanced by a domain-specific thesaurus during searching. The expanded searches provided high recall, including the highly relevant documents of relevance score 3.

In general, the precision is low and the expanded retrieval strategy could be enriched by some form of ranking. The present results are based on structured concept-based queries and inclusion of all search aspects, but still precision is not satisfactory. Freund & Toms (2005) suggests ranking documents according to contextual factors rather than uniquely basing it on frequency and position of search terms. Going through comments from the relevance assessors we can observe contextual relationships between information task, document type, source, and study techniques. We will pursue this finding in future research to see if contextual factors can improve ranking or precision of expanded searches. As human indexing is costly, it could be useful and productive to use the human indexer to assign other types of metadata such as contextual metadata, and leave the subject indexing to the computer.

The thesaurus used for query expansion is tailored to meet the needs of the test environment. Domain-specific thesauri are costly to produce and maintain, and another issue to study is to compare retrieval performance of query expansion with use of a tailored, manually constructed thesaurus and a statistically constructed thesaurus, as suggested by Kekäläinen & Järvelin (1998, 2000).

A third issue to investigate is whether the semantic information of the thesaurus may be used as basis for automatic assignment of metadata terms. Investigations of rule-based automatic categorization shows promising result, but an important drawback is the work involved in generating categorizations rules (Golub, 2006). The present findings suggest that it might be possible to base the generation of categorization rules on semantic data from the domain-specific thesaurus. Compared with automatic extracted indexing that requires that the searcher knows the words to use, automatic categorization techniques assign metadata terms from an existing vocabulary. In some instances, it is easier to find information about a particular subject if you know the search terms on beforehand, or can see it in the context of related information (Bates, 1986). This is an important advantage of using controlled vocabularies for information retrieval and stress the importance of developing the automatic, rule-based categorization techniques.

Acknowledgements. We thank the relevance assessors for assistance in making relevance judgments. We also thank Per Christensen and Frans la Cour from Metier for granting us a temporary license to use Ultraseek and for technical support.

References

Anderson, J. D. & Pérez-Carballo, J. (2001a). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management*, 37. 231–254.

Anderson, J. D. & Pérez-Carballo, J. (2001b). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II. Maschine indexing, and the allocation of human versus maschine effort. *Information Processing & Management*, 37. 255–277.

Bates, M. J. (1986). Subject access in online catalogs: a design model. *Journal of The American Society for Information Science*, 37(6), 357-376.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), paper no. 152.

Delphi Group (2004). *Information intelligence: content classification and the enterprise taxonomy practice*. Boston (MA): Delphi Group.

EMC Software (2007). Retrieved July 2, 2007, from http://www.software.emc.com/

Freund, L.; Toms, E. G. & Waterhouse, J. (2005). Modelling the information behaviour of software engineers using a work – task framework. In Grove, A (ed.). *Proceedings 68th Annual Meeting of the American Society for Information Science and Technology*, Charlotte, NC, October 28 – November 3, 2005.

Golub. K. 2006. Automated subject classification of textual Web pages, based on a controlled vocabulary: challenges and recommendations. *New Review of Hypermedia and Multimedia*, 12(1), 11-27

Greenberg, J. (2001a). Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology,* 52 (5). 402 – 415.

Greenberg, J. (2001b). Optimal query expansion (QE) processing methods via semantically encoded strutures thesauri terminology. *Journal of the American Society for Information Science and Technology,* 52(6). 487–498.

Hawking, D. & Zobel, J. (2007). Does topic metadata help with web search. *Journal of the American Society for Information Science and Technology.* 58(5). 613 – 628.

Hunter, J. (2003). Working towards MetaUtopia: A survey of current metadata research. *Library Trends*, 52 (2).

IDOL K2 (2007). Retrieved July 2, 2007, from http://www.autonomy.com/content/Products/IDOL_K2/ .

Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context.* Dordrecht: Springer.

Kekäläinen, J. K. & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Devlopment in Information Retrieval* (ACM SIGIR '98), Melbourne, Australia, August 24-28, 1998. New York (NY): ACM Press. 130-137.

Kekäläinen, J. K. & Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance inprobabilistic text retrieval. *Information Retrieval,* (1). 329-344.

Kristensen, J. (1993). Expanding end-users" query statements for free text searching with a search-aid thesaurus. *Information Processing & Management.* 29(6). 733–44.

Kristensen, J. & Järvelin, K. (1990). The effectiveness of a searching thesaurus in free-text searching of full-text database. *International classification.* 17(2). 77–84.

Lancaster, W. (2003). *Indexing and abstracting in theory and practice.* London: Facet publishing.

Moens, M-F. (2000). *Automatic indexing and abstracting of documents.* Boston: Kluwer.

Nielsen, M. L. (2004). Task-based evaluation of associative thesaurus in real life environment. In Schamber, L & Barry, C L (eds.), Proceedings of the ASIST 2004 Annual Meeting; "Managing and Enhancing Information: Cultures and Conflicts", Providence, Rhode Island, November 13 – 18. 437-447.

Skov, M., Larsen, B. & Ingwersen, P. (2006). Inter and intra-document contexts applied in polyrepresentation. In Ruthven, I, Borlund, P, Ingwersen, P, Belkin, N, Tombros, T, & Vakkari, P (eds.), *Proceedings of the first IIiX Symposium on Information Interaction in Context*, Royal School of Library and Information Science, Copenhagen, Denmark, October 18-20, 2006. 163-170.

Sormunen, E. (2002). Liberal Relevance Criteria of TREC – Counting on Negligible Documents? In *Proceedings of the Twenty-Fifth Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* Tampere, Finland, August 11-15, 2002. 324-330.

Spink, A. (1994). Term relevance feedback and mediated database searching: implications for information retrieval practice and systems design. *Information Processing and management,* 31(2). 161–171.

Stephenson, L. S. L. (1999). An assessment of the effectiveness of metadata as a tool for electronic resource discovery. Unpublished Master's Thesis. Scholl of Information and Lirary Science, Unoversity of North Carolina at Chapel Hill. Retrieved June 5, from ils.unc.edu/MSpapers/2511.pdf

Tudhope, D., Binding, C., Blocks, D. & Cunliffe, D. 2006. Query expansion via conceptual distance in thesaurus indexed collections. *Journal of Documentation*, 62(4), 509-533.

Addresses of congratulating authors:

Marianne Lykke
Royal School of Library and Information Science
Fredrik Bajers Vej 7K, DK-9220 Aalborg East, Denmark
Email: mln[at]iva.dk

Anna Gjerluf Eslau
H. Lundbeck A/S, Ottiliavej 9, DK-2500 Valby, Denmark
Email: age[at]lundbeck.com

Polyrepresentation and Interaction

Ryen W. White

Microsoft Research, Redmond, USA

Abstract. Information seeking is traditionally conducted in environments where search results or recommendations are primarily selected and presented independent of context. The principle of *polyrepresentation* (Ingwersen, 1994) suggests that Information Retrieval (IR) systems should provide and use different cognitive structures during acts of communication to reduce the uncertainty associated with interactive IR. This interaction can occur between a user and a search system or more broadly between people and resources as they explore document collections. In this paper we describe two research projects, each strongly related to polyrepresentation and each demonstrating the potential of polyrepresentation to enhance search interaction. First we describe the creation and evaluation of *content-rich* search interfaces that can display multiple representations of the retrieved documents simultaneously on the results interface. Then we describe research on leveraging overlap between multiple contextual sources to boost performance in a Web page recommendation setting, where pages are suggested to users as they navigate the Web.

Introduction

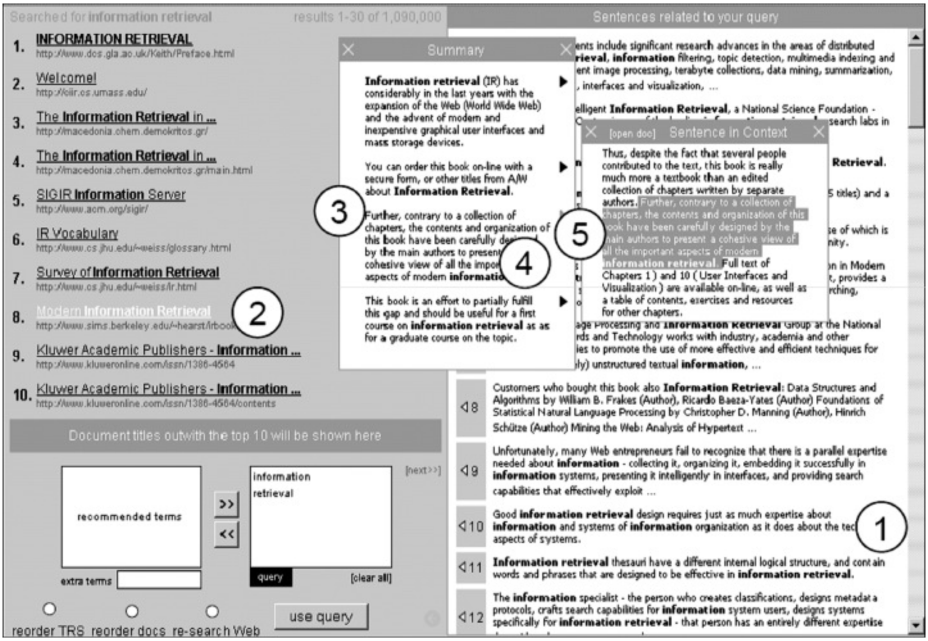
Ingwersen's theory of polyrepresentation (first published in 1994 and in fully-expanded form in 1996) suggests that overlaps in users' information needs and overlaps between document representations can improve information retrieval (IR) effectiveness. The theory fundamentally altered how the IR community regarded information redundancy and helped underscore the importance of cognitive processes in information seeking. The cognitive structures around which polyrepresentation is based are manifestations of human cognition, reflection or ideas. In IR they are typically transformations generated by a variety of human actors with a variety of *cognitive origins*. Author text, including document titles and their full-text are representations of cognitive structures intended to be communicated. These portions of text, have different *functional origins*; they have the same cognitive origin but were created in a different way or for a different purpose.

Polyrepresentative theory has been implemented through plausible inference techniques applied on networks of document representations (Turtle & Croft, 1990), or across networks of citations where those who cite documents have unique cognitive structures (Larsen & Ingwersen, 2002). More recently, researchers have applied the theory to context modeling (Skov, Larsen & Ingwersen, 2006), data fusion (Larsen, Ingwersen & Lund, 2009), and geometric retrieval (Frommholz et al., 2010). While such research has benefit to searchers, more work is needed to investigate the value of polyrepresentative principles in interactive settings. Belkin and colleagues (1993) established that the polyrepresentative extraction of information needs is potentially more effective than eliciting the solitary, isolated query statements gathered by most IR systems. In a similar way, offering thesauri (Jones et al., 1995) and clarification forms (Kelly et al., 2005) during query formulation have been shown to lead to more effective query statements.

In this paper, we describe some of our research on leveraging the principles of polyrepresentation to support more effective search interaction. In particular, we target two of the theory's defining elements: overlap between document representations and overlap between users' information needs. We first describe a polyrepresentative search interface capable of showing multiple representations of top-ranked search results and a series of studies of this approach with human participants and user simulations. We then describe a large-scale log-based analysis of search behavior in which we examined the use of multiple document contexts for a user on a Web page – task, historic, social, collection, and interaction – and combinations of these contexts to support user interest modeling in a recommendation setting. Both of the studies described use sources with multiple cognitive and functional origins, and both studies demonstrate the potential of polyrepresentation for enhancing search- or recommendation-related interaction.

Content-Rich Search Interfaces

Search result presentation plays an important role in influencing search interaction and ultimately, search success. Polyrepresentation suggests that representations of different cognitive structures should be offered to searchers, and used by them during their interaction with IR systems. We developed polyrepresentative search environments comprising multiple representations (or views) on each of the most highly-ranked Web documents (see White, 2004 for a detailed summary). Many of the systems we developed monitored user interactions with the elements in those search interfaces (document representations of differing granularity) and used that interaction to construct models of user interests through implicit relevance feedback. A screenshot of an example search environment is shown in Figure 1. As well as being represented by their full-text, documents are also represented by a number of smaller, query-relevant representations, created at retrieval time. These comprise the title (2) and a query-biased summary of the document (3) (White, Jose & Ruthven, 2003). A list of sentences extracted from the top-ranked documents retrieved scored in relation to the query, called *Top-Ranking Sentences* (TRS), include sentences from each document (1). Each sentence included in the top-ranking sentence list is a representation of the document, as is each sentence in the summary (4). Finally, for each summary sentence there is an associated sentence in the context it occurs in the document (i.e., with the preceding and following sentence from the full-text) (5).



Searched for **information retrieval** results 1-30 of 1,090,000

1. **INFORMATION RETRIEVAL**
   http://www.dcs.gla.ac.uk/Keith/Preface.html
2. Welcome!
   http://ciir.cs.umass.edu/
3. The **Information Retrieval** in ...
   http://macedonia.chem.demokritos.gr/
4. The **Information Retrieval** in ...
   http://macedonia.chem.demokritos.gr/main.html
5. SIGIR **Information** Server
   http://www.acm.org/sigir/
6. IR Vocabulary
   http://www.cs.jhu.edu/~weiss/glossary.html
7. Survey of **Information Retrieval**
   http://www.cs.jhu.edu/~weiss/ir.html
8. Modern **Information Retrieval**
   http://www.sims.berkeley.edu/~hearst/irbook
9. Kluwer Academic Publishers - **Information ...**
   http://www.kluweronline.com/issn/1386-4564
10. Kluwer Academic Publishers - **Information ...**
    http://www.kluweronline.com/issn/1386-4564/contents

Document titles outwith the top 10 will be shown here

Summary

Information retrieval (IR) has considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphical user interfaces and mass storage devices.

You can order this book on-line with a secure form, or other titles from A/W about **Information Retrieval**.

Further, contrary to a collection of chapters, the contents and organization of this book have been carefully de... by the main authors to present... cohesive view of all the impor... aspects of modern **informatio**...

This book is an effort to partially fulfill this gap and should be useful for a first course on **information retrieval** as as for a graduate course on the topic.

[open doc] Sentence in Context

Thus, despite the fact that several people contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern **information retrieval**. Full text of Chapters 1) and 10 (User Interfaces and Visualization) are available on-line, as well as a table of contents, exercises and resources for other chapters.

Sentences related to your query

...ents include significant research advances in the areas of distributed ...**rieval**, **information** filtering, topic detection, multimedia indexing and ...ent image processing, terabyte collections, data mining, summarization, ..., interfaces and visualization, ...

...elligent **Information Retrieval**, a National Science Foundation - ...search labs in

...**Retrieval**.

...5 titles) and a

...se of which is ...nity.

...on in Modern ...t, provides a ...rching,

...age Processing and **Information Retrieval** Group at the National ...rds and Technology works with industry, academia and other ...ies to promote the use of more effective and efficient techniques for ...ly) unstructured textual **information**, ...

◁8 Customers who bought this book also **Information Retrieval**: Data Structures and Algorithms by William B. Frakes (Author), Ricardo Baeza-Yates (Author) Foundations of Statistical Natural Language Processing by Christopher D. Manning (Author), Hinrich Schütze (Author) Mining the Web: Analysis of Hypertext ...

◁9 Unfortunately, many Web entrepreneurs fail to recognize that there is a parallel expertise needed about **information** - collecting it, organizing it, embedding it successfully in **information** systems, presenting it intelligently in interfaces, and providing search capabilities that effectively exploit ...

◁10 Good **information retrieval** design requires just as much expertise about **information** and systems of **information** organization as it does about the tec... aspects of systems.

◁11 **Information retrieval** thesauri have a different internal logical structure, and contain words and phrases that are designed to be effective in **information retrieval**.

◁12 The **information** specialist - the person who creates classifications, designs metadata protocols, crafts search capabilities for **information** system users, designs systems specifically for **information retrieval** - that person has an entirely different expertise

Figure 1. Content-rich search interface.

The document representations were arranged in interactive *relevance paths* (the order of which is denoted by the numbers in Figure 1), and encouraged interaction with the content of the retrieved document set. We call this approach *content-driven information seeking* (CDIS) since it is the content of the retrieved documents that drives the information-seeking process (White, Jose & Ruthven, 2005). This is in contrast to query-driven information-seeking, where searchers proactively seek information through the query they provide. Typically Web-search systems use lists of document surrogates to present their search results. This forces searchers to make two steps when assessing document relevance; first assess the surrogate, then perhaps peruse and assess the document (Paice, 1990). Such systems enforce a *pull* information seeking strategy, where searchers are proactive in locating potentially relevant information from within documents. In CDIS, it is the system that acts proactively, presenting the searcher with potentially relevant sentences taken from the document set at retrieval-time. The system uses a *push* approach, where potentially useful information is extracted from each document and proactively pushed to the searcher at the results interface. Searchers have to spend less time *locating* potentially useful information. As the users explore the top-ranked search results through this interface, the system uses their interaction to make suggestions about additional query terms that may be appropriate to add to the original query, or retrieval strategies related to the estimated level of change in their information needs during the search session. Depending on the amount of divergence from the original request the system estimated, it would either take no action, recommend that the user reorder top-ranking sentences extracted from the top documents, reorder the top-ranked search results, or if the estimated change in need was sufficient, then re-search the Web.

We performed five user studies on variants of this interface, involving over 150 participants over the course of three years. Each user study targeted a particular aspect of the interface, from the use of document representations to facilitate more effective information access (White, Ruthven & Jose, 2005), to different amounts of user control over aspects of the search process (relevance indication, query formulation, and action selection) (White & Ruthven, 2006). The findings of our research suggested that users found these content-rich interfaces useful for tasks that were exploratory in nature (i.e., where they needed to gather background information on a particular topic or gather sufficient information to enable them to make a decision about the best course of action). However, the interfaces were not as effective in known-item searches where users had to find a specific piece of information. In addition, users wanted to retain control over the strategic aspects of their search such as the decisions to conduct new searches, but were willing to delegate control for less severe interface actions to the system. A number of our studies compared this interface with a traditional Web search interface. The findings showed that searchers benefited from the additional information both in terms of subjective measures such as task success and more objective measures such as task completion time.

In addition to the user studies, we also developed simulations of searchers' interaction behavior with this environment that afforded us greater control over experimental variables. Simulations provide a way to evaluate systems, interaction metaphors, and suchlike prior to system building. We used these searcher simulations to select the best performing implicit relevance feedback algorithms from a set of alternatives (White et al., 2005) and in re-designing the layout of the interface to maximize implicit relevance feedback performance (White, 2006). User simulations are a potentially powerful technique for assessing aspects systems without the need for user studies and prior to large-scale deployment. An alternative source of information on system effectiveness are interaction logs that record users' interactions with deployed systems. In the next section we describe research on leveraging log data and reference information (from hyperlinks) to model and use another important aspect of polyrepresentation: context.

Modeling Context Combinations

A key aspect of the principle of polyrepresentation is the use of *cognitive overlap* between multiple contextual elements to strengthen the relevance signal of certain items. We have studied the use of these overlaps to support more effective modeling of user's short-, medium-, and long-term search interests in a Web page recommendation setting (White, Bailey & Chen, 2009). Although our study is aimed at providing better Web page recommendations for users engaged in browsing activity, the findings could also potentially improve the design of context-sensitive search applications. The situation we addressed was given that a user is on a Web page, , predict their future interests using context for that page. We developed user interest models based on and the five sources of contextual information used in our study. The sources were chosen based on elements of a nested model of context stratification proposed by Ingwersen and Järvelin (2005). The dimensions of that model represent the main contextual influences affecting users engaged in information behavior: (i) *object structures*::signs (i.e., discrete units of meaning), page features, and cognitive structures (user); (ii) *inter-object contexts or structures*: between-object relations such as hyperlinks or citations; (iii) *interaction*: evidence of interaction behavior during the search session; (iv) *social, systemic, domain-work task*: peer group, retrieval system (systemic), real work or daily-life tasks; (v) *economic techno-, physical-, and societal context*: prevailing infrastructures that influence all elements in the nested model of context, and; (vi) *historic*: the experiences of the cognitive actor (user) that affect how they perceive and interpret situations. The context stratification is illustrated in Figure 2, with the user at a given Web page, at the core of the model, and with the dimensions used in our study underlined and shown in boldface. The dimensions not chosen (e.g., intra-object structures, signs, and emotions) could not accurately be modeled in a log-based study since we lacked access to Web page content (only their URLs), the user's cognitive and affective state at session time, or infrastructure details.

Understanding which sources and source combinations best predict future user interests is critical for the development of effective Web page recommendation methods. We studied the value of the current page and five distinct context sources in predicting future interests at different temporal granularities. The contexts were interpreted given the log and link data available in our study as: (i) *interaction*: recent interaction behavior preceding the current Web page; (ii) *collection*: pages with hyperlinks to the current page;

(iii) *task*: pages related to the current page by sharing the same search engine queries; (iv) *historic*: the long-term interests for the current user, and; (v) *social*: the combined interests of other users that also visit the current page. This was the first study to systematically assess contextual variants for user interest modeling. Interests were modeled as a distribution of Open Directory Project (ODP, dmoz.org) category labels. Perhaps more interestingly from a polyrepresentative perspective, we also studied the use of overlap between multiple sources as a stronger source of contextual signal.

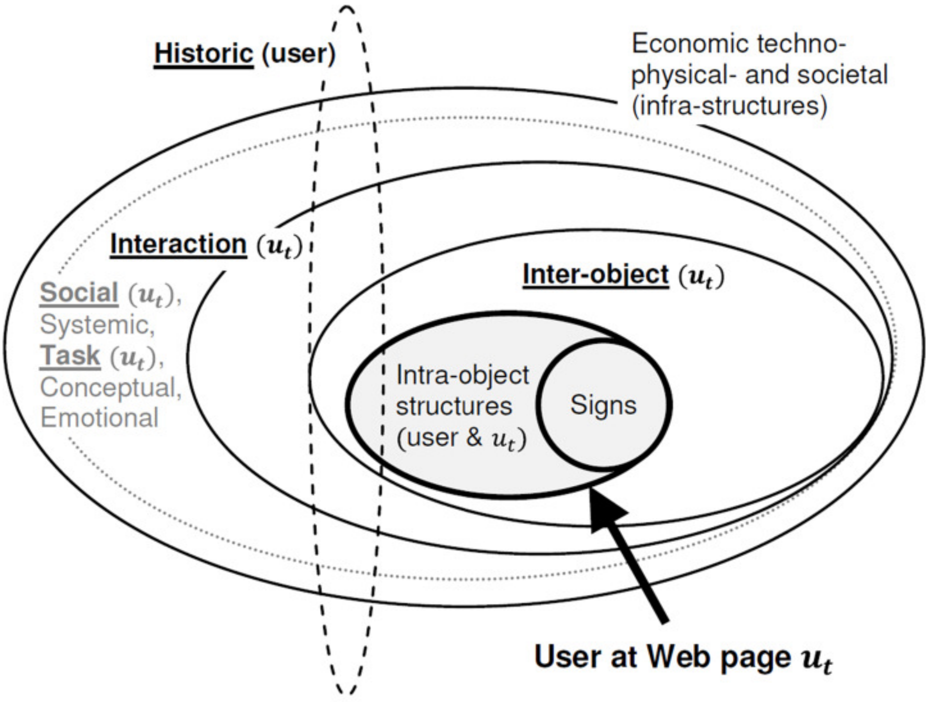

Historic (user)
Economic techno-physical- and societal (infra-structures)
Interaction (u_t)
Social (u_t), Systemic, Task (u_t), Conceptual, Emotional
Inter-object (u_t)
Intra-object structures (user & u_t)
Signs
User at Web page u_t

Figure 2. The nested model of context stratification for information seeking and retrieval (based on Ingwersen and Järvelin, 2005).

The findings from our study suggested that: *interaction context* most accurately predicts short-term future interests (within the next hour), *task context* most accurately predicts medium-term interests (within the next day), and *historic context* most accurately predicts long-term future interests (within the next week). We also systematically varied the combinations of contexts used, such that over 50 context combinations were tested. We selected the ODP category labels and their respective frequencies *for labels that appeared in all relevant interest models*; giving us the overlap between context sources. The findings of our analysis show that using a combination of multiple context sources leads to more accurate future predictions in the short-, medium-, and long-term. For each time duration, there exists at least one context combination that significantly outperforms all contexts in isolation; this supports the principle of polyrepresentation. In addition, our findings demonstrate that certain contexts are required to obtain high prediction accuracy (current page and *interaction context* in short-term predictions, *task context* in medium-term predictions, and *social context* and *historic context* in long-term predictions).

Conclusions and Future Directions

It is clear that Ingwersen's seminal work on polyrepresentation has inspired a significant amount of information-seeking research. In this paper we have summarized some of our work on leveraging the principles of polyrepresentation in the design and evaluation of search and recommendation systems. The breadth of the research described emphasizes the range of research areas and application domains that can benefit from extant and future polyrepresentation research. To realize the potential of polyrepresentation, more research is needed on ways to more completely model it in IR, on evaluation methodologies capable of measuring the potential benefit of polyrepresentative search interfaces prior to costly development and deployment, and on tools to elicit polyrepresentative representations of information needs from searchers either explicitly or implicitly. Of particular interest to us is how log data on existing search interfaces can be leveraged to build accurate simulations of peoples' search behavior that can be useful for evaluating aspects of new polyrepresentative search interfaces in the formative stages of their design.

References

Belkin, N.J., Cool, C., Croft, W.B. and Callan, J.P. (1993). The effect of multiple query representations on information retrieval system performance. In *Proc. SIGIR*, 339-346.

Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P. and Van Rijsbergen, C.J. (2010). Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proc. IIiX*, in press.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proc. SIGIR*, 101-110.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, 52, 3-50.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J. and Walker, S. (1995). Interactive thesaurus navigation: intelligence rules ok. *JASIST*, 46(1): 52-59.

Kelly, D., Dollu, V.D. and Fu, X. (2005). The loquacious user: a document-independent source of terms for query expansion. In *Proc. SIGIR*, 457-464

Larsen, B. and Ingwersen, P. (2002). The boomerang effect: retrieving scientific documents via the network of references and citations. In *Proc. SIGIR*, 397-398.

Larsen, B., Ingwersen, P. and Lund, B. (2009). Data fusion according to the principle of polyrepresentation. *JASIST*, 60(4): 646-654.

Paice, C.D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1): 171-186.

Skov, M., Larsen, B. and Ingwersen, P. (2006). Inter and intra-document contexts applied in polyrepresentation. In *Proc. IIiX*, 97-101

Turtle, H. and Croft, W.B. (1990). Inference networks for document retrieval. In *Proc. SIGIR*, 1-24.

White, R.W. (2004). *Implicit Feedback for Interactive Information Retrieval*. Unpublished doctoral dissertation, Department of Computing Science, University of Glasgow.

White, R.W. (2006). Using searcher simulations to redesign a polyrepresentative implicit feedback interface. *Information Processing and Management*, 42(5): 1185-1202.

White, R.W., Jose, J.M. and Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management*,39(5): 707-733

White, R.W., Ruthven, I., Jose, J.M. and Van Rijsbergen, C.J. (2005). Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3): 325-361.

White, R.W., Jose, J.M. and Ruthven, I. (2005). Using top-ranking sentences to facilitate effective information access. *JASIST*, 56(10): 1113-1125.

White, R.W. and Ruthven, I. (2006). A study of interface support mechanisms for interactive information retrieval. *JASIST*, 57(7): 933-948.

White, R.W., Bailey, P. and Chen, L. (2009). Predicting User Interests from Contextual Information. In *Proc. SIGIR*, 363-370.

Address of congratulating author:

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA USA 98052
Email: ryenw[at]microsoft.com

Information Retrieval and Chemoinformatics: Is there a Bibliometric Link?

Peter Willett

University of Sheffield, Sheffield, United Kingdom

Abstract. There are obvious links between the database processing required for information retrieval (IR) and for chemoinformatics, and this might have been expected to result in a fair degree of cross-citation between the *Journal of Chemical Information and Modeling*, the core journal for chemoinformatics, and leading IR journals. While this was true in the early days of chemoinformatics, current IR researchers would appear to take no account of developments in chemoinformatics, or *vice versa*.

Keywords: Chemoinformatics, Database searching, Information retrieval.

1. Introduction

I first made Peter Ingwersen's acquaintance when attending the annual ACM SIGIR (for Special Interest Group on Information Retrieval) conferences entitled Research and Development in Information Retrieval, and we subsequently worked together on many occasions as members of the conference organising or programme committees. The conferences have always included a series of workshops and seminars. For several years, Peter and I collaborated on a seminar that provided an introduction to information retrieval (hereafter IR) for those delegates new to the field, he covering the cognitive aspects of information retrieval whilst I covered the more computational aspects of the subject. Much of the material presented in these enjoyable sessions was included in a subsequent review [1]. A further recollection from SIGIR days is the 1992 conference. This was held in Copenhagen during the closing stages of that year's European Football Cup, with the conference dinner coinciding with the competition final. Each course of the dinner was accompanied by an update from Peter on the progress of the match, the last of which was suitably celebratory as Denmark won 2-0 against the Cup favourites, Germany.

Whilst being heavily involved in IR during the early part of my career, I had also developed strong interests in chemoinformatics (*vide infra*) and my research efforts became increasingly focused in the latter area, with the result that I ceased to be active in IR by the late Nineties. I do, however, maintain a watching brief since I believe that there are links – actual or potential – between some aspects of IR and some aspects of chemoinformatics. In this note, I shall explore the extent of these links, both from my personal viewpoint as a researcher and from a bibliometric viewpoint as evidenced by a brief inspection of the citation linkages that exist between the literatures of the two subjects.

2. Chemoinformatics

Contributors to this festschrift will be familiar with IR, but chemoinformatics is probably much less familiar, so I shall begin with a brief introduction to the subject before assessing the extent of its overlap with IR [2, 3].

Chemoinformatics has as its principal focus the processing of information about the structures of chemical molecules, in much the same way as bioinformatics focuses on the sequences and structures of biological macromolecules, and many of the techniques that have been developed reflect the very close relationship that chemoinformatics has with the pharmaceutical industry. Thus, much work over the years has focused on the specific task of identifying novel molecules with biological activities that address therapeutic needs, e.g., lowering blood pressure or cholesterol levels, shrinking specific types of tumour, or alleviating the effects of stomach ulcers. Typical chemoinformatics applications include methods for: searching databases of molecules and databases of reactions; deducing statistical relationships between the structures of molecules and their chemical and biological properties; suggesting novel chemical syntheses; predicting the shape and the thermodynamic stability of molecules; designing cost-effective biological testing strategies; exploring the interactions between a potential drug and a biological receptor site; and deducing the identity of an unknown molecule from its spectrum, *inter alia* [4, 5]. While many of these topics have a strong chemical and/or biological focus that is far removed from even the broadest definition of IR, there are substantial similarities in the methods that are used for database searching, as I shall now exemplify.

Text databases can be queried in three ways: an exact match search for a specific document; a partial match search for all those documents containing a Boolean combination of search terms; and a best match search for those documents that are most similar to the query statement. Current chemoinformatics systems provide the same three modes of access to chemical databases, some of which contain millions or even tens of millions of different molecules. Searches can hence be carried out for specific molecules, for all molecules containing some particular substructural pattern (e.g., a benzodiazepine ring system), or for those molecules most similar to a given molecule (e.g., molecules structurally related to an existing drug).

Molecules are stored in a chemical database as labelled graphs, in which the nodes and edges of a graph are used to denote the atoms and bonds of a molecule (or the atoms and inter-atomic distances when representing a 3D molecule), and graph representations are used for many applications in chemoinformatics. However, while graphs provide an exact representation of molecular topology, many of the search algorithms for processing graphs have running times that are factorial in the numbers of atoms involved. Extensive use is hence made (either as an alternative or as a complement) of a simpler representation in which each molecule is represented by a binary string, called a 'fingerprint', that encodes the presence or absence of a few hundreds of substructural fragments in a data structure that is analogous to the manner in which a text signature describes the presence of words in a document. The fragments encoded in a fingerprint hence provide a summary representation of a molecule's structure in just the same way as a few selected keywords provide a summary representation of the full text of a document.

At the heart of IR is the ability to identify those few molecules in a database that are relevant to a user's query. In just the same way as a document is relevant or non-relevant to a query, so a molecule is either active or inactive in some particular biological test, and identifying molecules with a given activity is one of the principal functions of chemoinformatics systems. The performance of chemical searching methods can hence be evaluated using performance measures that mirror closely those used to evaluate the effectiveness of IR systems [6]. Moreover, in just the same way as the IR community has made use of test-collections ever since the early Cranfield experiments almost a half-century ago, so the chemoinformatics community has recently started to use analogous datasets containing sets of molecules that are known to be active or inactive in a biological test. For example, the National Cancer Institute distributes a database of *circa* 40K molecules that have been tested for HIV-1 activity as part of the US government's anti-AIDS programme.

A final area of commonality is the fact that the well-known Cluster Hypothesis, which states that similar documents tend to be relevant to the same requests and which underlies the use of clustering methods in IR [7, 8], has a direct equivalent in chemoinformatics. This is the Similar Property Principle [9], which states that similar molecules tend to exhibit the same biological properties. The equivalence means that just as there has been much interest in IR in identifying groups of similar molecules using the methods of cluster analysis, so there have been extensive studies (and substantial industrial applications) of methods for grouping similar molecules; indeed, my group's extensive work on the clustering of chemical databases [10] was driven in large part by earlier studies of the use of such techniques in the IR context [8].

The similarities that I have noted mean that many approaches that are applicable in the IR context are also potentially applicable in chemoinformatics, and *vice versa*. Thus, in a paper published in 2000, I summarised work in Sheffield on chemical applications of data fusion, of relevance feedback weights and of Zipfian word-frequency distributions, and on an IR application of work on measures of chemical similarity [11]. I argued than, and continue to believe now, that the two research communities have much to learn from each other: in the remainder of this note I report a small-scale bibliometric study that seeks to ascertain whether this belief is shared by others in the two communities.

3. Citation Links

Material on chemoinformatics is scattered very broadly across the chemical literature, as might be expected given the central role that the computer plays in any modern-day scientific discipline. However, a recent bibliometric analysis [12] showed that the subject's core literature comprises just four journals: the *Journal of Chemical Information and Modeling*, the *Journal of Computer-Aided Molecular Design*, the *Journal of Molecular Graphics and Modelling*, and *QSAR & Combinatorial Science*.

Of these, the first journal is by far the most important and I have hence investigated the extent to which there are citation links between it and five of the key periodicals in IR, specifically the *Annual Review of Information Science and*

	Articles	Citing articles	ARIST	IPM	JASIST	JD	JIS	% of citing articles
JCD	900	2139	69	33	96	34	27	12.1
JCICS	3432	25524	43	30	46	16	34	0.7
JCIM	1320	5251	2	0	0	0	0	0.0

Table 1. Citations from IR journals to the Journal of Chemical Information and Modeling

	Articles	Citing articles	JCD	JCICS	JCIM	% of citing articles
ARIST	525	3423	4	16	6	0.8
IPM	3124	12336	16	56	5	0.6
JASIST	5553	21176	105	70	10	0.9
JD	3577	7985	16	8	0	0.3
JIS	1939	5510	0	34	4	0.7

Table 2. Citations to IR journals from the Journal of Chemical Information and Modeling

Technology (*ARIST*), *Information Processing and Management* (*IPM*), the *Journal of of the American Society for Information Science and Technology* (*JASIST*), the *Journal of Documentation* (*JD*), and the *Journal of Information Science* (*JIS*).

The *Journal of Chemical Information and Modeling* celebrates its 50th anniversary in 2010: it started life in 1961 as the *Journal of Chemical Documentation*, became the *Journal of Chemical Information and Computer Sciences* in 1975, and adopted its current title in 2005; for brevity, these three names will be referred to as *JCD*, *JCICS* and *JCIM* respectively in what follows, with the *Journal of Chemical Information and Modeling* (or just 'the journal') referring to the journal as a whole. Citation searches were carried out in May 2010 using the Web of Science (*Science Citation Index Expanded*, *Social Sciences Citation Index*, *Arts and Humanities Citation Index* and *Conference Proceedings Citation Index - Science*), and the results are summarized in Tables 1 and 2.

For each of the three parts of *Journal of Chemical Information and Modeling*, Table 1 lists the number of articles published in the journal, the total number of articles that cited the journal, and then the numbers of articles that cited the journal in *ARIST*, *IPM*, *JASIST*, *JD* and *JIS*, with the final column listing the percentages of the citations to the journal that appeared in the five IR periodicals. The reader should note that 'article' here is taken to include all forms of item published in the journal, including not just articles as such but also reviews, editorial material etc. Table 2 contains the analogous data for the searches in the reverse direction, i.e., citations from the *Journal of Chemical Information and Modeling* to the five IR periodicals. The citation counts for the five IR periodicals include citations to/from previous manifestations of these periodicals: *IPM* was formerly *Information Storage and Retrieval*; *JASIST* was formerly *American Documentation* and then *Journal of the American Society for Information Science*; and *JIS* was formerly the *Information Scientist*.

We consider first the citations to the *Journal of Chemical Information and Modeling* in Table 1. For the citations to *JCD* (first row of Table 1), the five IR periodicals are at positions 2 (*JASIST*), 3 (*ARIST*), 4 (*JD*), 12 (*IPM*) and 14 (*JIS*) when the periodicals are ranked in decreasing order of numbers of citations to *JCD*. The journal has thus been highly cited in the IR literature, and inspection of this ranked list reveals other well-known LIS publications at positions 5 (*Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy*), 6 (*Special Libraries*), 8 (*Nachrichten fur Dokumentation*) and 16 (*Aslib Proceedings*). It is hence clear that the library and information science (LIS) community was well aware of chemoinformatics at this early stage of its development (although the subject was not known by that name till the late Nineties [2, 12]). Indeed, the final column of Table 1 shows that almost one-eighth of all the citations to *JCD* appeared in the five IR periodicals. A very different picture is revealed when we consider the citations to *JCICS* and *JCIM*. There are notably fewer citations to *JCICS* than there were to *JCD*: the highest-ranked IR publication, *JASIST* (and its predecessors), is at position 75 in the ranked list and the other journals are ranked even lower. There has hence been a massive diminution in the importance attached by the IR community to research in chemoinformatics (although some part of this relative decrease is undoubtedly due to the much increased recognition of chemoinformatics by the larger chemical community, as evidenced by the very many chemical journals citing *JCICS*). The situation is still starker with *JCIM*, which appears to have been totally ignored by researchers in IR with the sole exception of two citations from *ARIST*. The much greater number of IR (and LIS more generally) citations to *JCD* than to *JCICS* is particularly noteworthy given the much smaller number of articles available for citing in the former: 900 articles spread over 14 volumes for *JCD* as against 3432 articles over 30 volumes for *JCICS*.

A rather less extreme change over time is observed when we consider the citations from the journal to the five IR periodicals, as detailed in Table 2. All of these periodicals are cited to some extent, with *JASIST* being by far the most popular source with *Journal of Chemical Information and Modeling* authors; however, citations from the journal account for less than one-percent of the total citations for all of the five IR periodicals. In all, the 900 *JCD* articles contained 141 citations to these periodicals; the corresponding figures for *JCICS* and *JCIM* are 3432 articles and 184 citations, and 1320 articles and 25 citations, respectively. A comparison of the *JCD* and *JCICS* figures suggest that the chemoinformatics community took rather less account of IR research than they did in the early days of the subject; however, the drop-off in communication between the two research communities is less marked than in Table 1, and there is evidence for at least some continuing interest in the subject as reflected in the 25 citations from *JCIM*.

Inspection of the IR articles citing the journal or *vice versa* provides some insights to the reasons for citing. Chemistry has always been one of the most information-rich disciplines, and it has thus traditionally been in the vanguard of those seeking to apply technological developments to information processing functions, such as abstracting and indexing, database creation, and online searching *inter alia*. This has been the case not just with Chemical Abstracts Service, the principal bibliographical database system for the chemical sciences, but also with companies in the pharmaceutical and related industries, and academic chemical librarians. Thus, many of the early citations to *JCD* related to the pioneering attempts to computerize information functions that were then being made in chemistry and that would subsequently be applied in other subject domains, i.e., these developments were of interest not only to chemical information specialists but also to the LIS community more generally. As chemoinformatics has developed, the journal's articles have demonstrated a steady trend away from traditional chemical documentation and towards topics that are much more closely related to drug discovery and computer science. Thus, each issue of *JCIM* now contains five sections entitled Chemical; Information, Computational Chemistry, Computational Biochemistry, Pharmaceutical Modeling, and Bioinformatics: of these, only the first is likely to be of much interest to the IR community.

Analysis of the citing authors shows that there have been just five authors who have cited the *Journal of Chemical Information and Modeling* five or more times in articles published in the five IR periodicals; remarkably, no less than three of these are associated with the author's academic department. Thus, in decreasing order of number of citations these authors are: myself; Michael Lynch, now an Emeritus Professor of the department; David Bawden, previously a research student here but now Professor at the City University Department of Information Science in London; Timothy Craven, whose work on nested phrase indexing was closely related to early studies of articulated subject indexing carried out by Chemical Abstracts Service; and Gene Garfield, the founder not only of the world's citation indices but also of *Index Chemicus*, one of the first databases in synthetic organic chemistry. Willett, Lynch and Garfield are also amongst the five most frequent citers from the journal to the IR literature: the other two are John Barnard, another Sheffield research student and subsequently the founder of a chemical software company, and Charles Bernier, who worked at Chemical Abstracts Service in the early days of their computerization programme before moving to the School of Library and Information Science at the State University of New York. The pool of authors who cite across the two literatures is hence extremely limited.

4. Conclusions

The bibliometric data shows clearly that the IR research community (as exemplified by the authors of articles in our five chosen IR periodicals) was well aware of chemoinformatics when that subject first emerged, but that this awareness subsequently decreased to the extent that there is almost no current interest in it. At the same time, although to a lesser extent, there is reduced take-up of IR research by the chemoinformatics community (as exemplified by the authors of articles in the *Journal of Chemical Information and Modeling*).

It is sincerely to be hoped that this situation changes: that it should change is demonstrated by two current areas of interest. One of the most important research areas in chemoinformatics is that of 'virtual screening' [13, 14]: the ranking of a database of previously untested molecules in order of decreasing probability of activity, so that synthesis and biological testing can be focused on those few molecules that are most likely to exhibit the activity of interest. The analogy to IR models that rank documents in order of decreasing probability of relevance is obvious, and there have been a very small number of chemoinformatics studies that have applied IR techniques to the virtual screening context. Examples include the use of fingerprint weighting schemes that are based on probabilistic relevance weighting [15] and the use of rankings based on Bayesian inference networks [16]. However, the work to date has been limited and the wealth of IR models now available suggests that others might also be applicable in the chemoinformatics domain. Conversely, there is much current interest in IR in the concept of 'diversity', i.e., the identification of subsets of a search output that relate to the same topic or sub-topic [17, 18]; use could surely be made of the extensive chemoinformatics studies of molecular diversity analysis that have been carried out on methods for choosing structurally diverse sets of molecules [19, 20]. There are doubtless other areas that would profit by taking account of work done elsewhere.

In conclusion, there have been strong links between chemoinformatics and IR in the past: it is to be hoped that this starts to be the case again so that each subject can profit from future research developments in the other.

5. References

1. Ingwersen, P., Willett, P.: An Introduction to Algorithmic and Cognitive Approaches for Information Retrieval. Libri 45, 160--177 (1995)
2. Willett, P.: From Chemical Documentation to Chemoinformatics: Fifty Years of Chemical Information Science. Journal of Information Science 34, 4767--499 (2008)
3. Gasteiger, J.: The Central Role of Chemoinformatics. Chemometrics and Intelligent Laboratory Systems 82, 200--209 (2006)
4. Gasteiger, J., Engel, T. (eds.): Chemoinformatics: A Textbook. Wiley-VCH, Weinheim (2003)
5. Leach, A.R., Gillet, V.J.: An Introduction to Chemoinformatics. Kluwer, Dordrecht (2007)
6. Willett, P.: The Evaluation of Molecular Similarity and Molecular Diversity Methods Using Biological Activity Data. Methods in Molecular Biology 275, 51--63 (2004)
7. van Rijsbergen, C.J.: Information Retrieval. Butterworth, London (1979)
8. Willett, P.: Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing and Management 24, 577--597 (1988)
9. Johnson, M.A., Maggiora, G.M. (eds.): Concepts and Applications of Molecular Similarity. John Wiley, New York (1990)
10. Willett, P.: Similarity and Clustering in Chemical Information Systems. Research Studies Press, Letchworth (1987)

11. Willett, P.: Textual and Chemical Information Retrieval: Different Applications but Similar Algorithms. Information Research, Vol. 5 (2000) at URL http://InformationR.net/ir/5-2/infres52.html
12. Willett, P.: A Bibliometric Analysis of Chemoinformatics. Aslib Proceedings 60, 4--17 (2008)
13. Böhm, H.-J., Schneider, G. (eds.): Virtual Screening for Bioactive Molecules. Wiley-VCH, Weinheim (2000)
14. Alvarez, J., Shoichet, B. (eds.): Virtual Screening in Drug Discovery. CRC Press, Boca Raton (2005)
15. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A.: New Methods for Ligand-Based Virtual Screening: Use of Data-Fusion and Machine-Learning Techniques to Enhance the Effectiveness of Similarity Searching. Journal of Chemical Information and Modeling 46, 462--470 (2006)
16. Abdo, A., Salim, N.: Similarity-Based Virtual Screening with a Bayesian Inference Network. ChemMedChem 4, 210--218. (2009)
17. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Buttcher, S., MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation. In: Proceedings of the 31st ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 659--666. ACM, New York (2008)
18. Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., Paramita, M.: Multiple Approaches to Analysing Query Diversity. In: Proceedings of the 32nd ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 734--735. ACM, New York (2009)
19. Willett, P. (ed.): Computational Methods for the Analysis of Molecular Diversity. Kluwer, Dordrecht (1997)
20. Dean, P.M., Lewis, R.A. (eds.): Molecular Diversity in Drug Design. Kluwer, Amsterdam (1999)

Address of congratulating author:

Peter Willett
Information School
University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom
Email: p.willett[at]sheffield.ac.uk

Informetrics

The WIF of Peter Ingwersen's website

Judit Bar-Ilan

Bar-Ilan University, Israel

Abstract: Peter Ingwersen was among the first to apply bibliometric measures to the Web. In a seminal paper in 1998 [1], Peter defined the Web Impact Factor, the Web parallel of the well-known Journal Impact Factor. In this short paper in honor of Peter we study the Web Impact Factor of Peter's website, http://www.db.dk/pi/. In addition, we also introduce a new h-type index for websites, and compute it for Peter Ingwersen's website.

1 Introduction

When I was invited to contribute to this festschrift, it was clear to me that I was going to write something about the Web Impact Factor (WIF). Peter contributed to information science in a number of areas, but for me his most significant contribution was the simple and elegant definition of the Web Impact Factor. He is one of the pioneers of webometrics, and with this seminal paper he helped to establish this new subfield of informetrics.

Probably the first paper applying bibliometric methods to the Web was published by Ray Larson in 1996 [2], where he applied co-citation analysis methods to websites. Other early webometric publications included the Almind and Ingwersen [3] paper, where the term "webometrics" was coined, an early paper by Ronald Rousseau on 'sitations' [4], my paper analyzing the growth characteristics of messages in Usenet newsgroups [5] and a paper studying forms of mention on the Web by Blaise Cronin and colleagues [6]. The WIF was quickly picked up by Alistair Smith [7] and Mike Thelwall [8, 9]. Mike Thelwall further developed the WIF concept in a number of papers, e.g. [10, 11].

2 Terminology

Webometrics, in the Almind and Ingwersen paper [3] is defined as "research of all network-based communication using informetric or other quantitative measures" (p.404). In a later paper, Björneborn and Ingwersen [12] limit the term "webometrics" to "[t]he study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (p. 1217), and use the term "cybermetrics" for studying the quantitative aspects of the whole Internet drawing on informetrics methods. In spite of the suggested change of terminology, the term webometrics is still often used in place of cybermetrics.

The WIF also has several different definitions. Peter Ingwersen defined three versions of it in his original paper, and Mike Thelwall proposed later on several variants, e.g. the Research WIF [10]. The definitions proposed by Ingwersen were WIF, external-WIF and self-link WIF, where the WIF of a site, *s* is:

$$WIF(s) = \frac{\#\ \textit{web pages that link at least once to a page in s}}{\#\ \textit{web pages in s}} \qquad (1)$$

For the external WIF, only those pages not belonging to the given site with links to the given site are counted, and self-link WIF takes into account link pages only from within the site. Because internal links are often used for navigational purposes, the external WIF is probably the best measure among the three for evaluating the visibility of a website. In the following we will calculate the external WIF of Peter Ingwersen's website (http://www.db.dk/pi/) and analyze some characteristics of the linking webpages.

Seemingly, nowadays informetric papers "must" include some version of the *h-index* as well [13]. We shall develop an h-index for single sites along the lines of Schubert's h-index for single papers [14], where the *h-index* of a paper, *p*, is defined as

$$h(p) = \max_h \exists\, h \text{ citing papers of } p \text{ that received } h \text{ citations or more.} \qquad (2)$$

Thus, the *hw-index* of a web site *s* is defined by us as:

$$hw(s) = \max_h \begin{array}{c}\exists\, h \text{ pages with links to pages of } s \text{ such that each such} \\ \text{page has } h \text{ or more pages linking to it}\end{array} \qquad (3)$$

3 Data collection

Link data was collected by the Yahoo! Siteexplorer application (http://siteexplorer.search.yahoo.com/) on May 8, 2010. All linking pages were downloaded in order to allow further analysis. In addition, for each linking page we recorded the number of pages linking to it, again using Yahoo!'s Siteexplorer. All the data were collected on May 8, 2010. For the pages linking to Peter Ingwersen's webiste, all links whether external or internal to the domain on which the linking page resides were counted.

4 The WIF of the site

Here we compute the external WIF for Peter Ingwersen's website. The website is small, it is comprised of two pages only: http://www.db.dk/pi/ and http://www.db.dk/pi/iri/ only. Thus the denominator for the WIF computation is 2. The number of pages with inlinks to one of these two pages that were identified by the Yahoo! Siteexplorer on May 8, 2010. Note, that it is of importance to state the exact date of the data collections, since there are slight fluctuations in the number of inlinks reported by Yahoo! [15].

Altogether 197 linking pages were identified, 131 external links to http://www.db.dk/pi/ and 86 links to http://www.db.dk/pi/iri/. There were 20 pages with links to both pages of the site. Thus the external-WIF of Peter's website is 98.5. If we compare this value to the values reported in the original Ingwersen paper on the WIF [1], this value is considerably higher than all of the values reported (the highest WIF in the original paper was 51 for the website of the journal Nature. This comparison is at most of anecdotal value, since it is not reasonable to compare WIF values that were calculated in 1998 with those calculated in 2010. As of May 25, 2010, the number of pages with external inlinks to www.nature.com is 2,486,587, while the reported size of the website is 3,599,788, thus the current external WIF of Nature is 0.69.

The top-level domains and the hosts occurring more than 3 times among the linking pages are listed in Tables 1 and 2 respectively. Interesting to note the large number of inlinks coming from blogs.

Country or top level domain	Country or top level code	Number of linking pages
.com	com	38
Denmark	dk	28
Germany	.de	24
.org	org	20
Finland	.fi	16
.edu	edu	13
.net	net	10
Argentina	.ar	7
Czech Republic	.cz	7
Spain	.es	6
UK	.uk	6

Table 1. Top-level and country domains occuring more than 3 times

Domain	Nr. of linking pages	Domain details
www.db.dk	11	Royal School of Library and Information Science
www.soegning.dk	11	Danish portal/search engine
www.bibliometria.com	8	Spanish blog on bibliometrics
www.abo.fi	7	Åbo Akademi University, Finland
comminfo.rutgers.edu	6	Rutgers, School of Communication and Information
www.cadius.org	6	Blog of a Spanish speaking community of professionals from the fields of interaction design, usability and information architecture
invisibleweblog.blogspot.com	5	Yazdan Mansourian's blog about the Invisible Web and information availability on the Web
community.livejournal.com	4	Document retrieval blog in Russian
hosting.zkm.de	4	ICIE website – an academic website on information ethics
library2pointoh.fi	4	Library 2.0 blog from Åbo Akademi
weblog.ib.hu-berlin.de	4	Blog of the Institute of Library and Information Science at the Humboldt University in Berlin
www.ikaros.cz	4	Ikaros, Czech electronic magazine on the information society

Table 2. Hosts of the linking pages occurring more than 3 times

5 The Web *h*-index of the site

Peter's website has an *hw*-index of 32, which sounds highly respectable, but because we do not have comparative data from websites of other information scientists, it is difficult to say whether this number is high or low. Again, the prominence of blogging sites is very high. Table 3 displays the 32 pages "responsible" for the hw-index of the site. Note that probably Yahoo! Siteexplorer ignores the characters following the ? sign in the URLs, thus three pages from cadius.org appear in the list with exactly the same number of linking pages.

6 Conclusion

In this short paper we calculated the WIF and the newly introduced hw-index of Peter's website. The applicability of the *hw*-index should be further studied.

URL	# of linking pages	Page type
http://invisibleweblog.blogspot.com/2005_11_01_archive.html	908	blog
http://irsweb.blogspot.com/2005_02_01_archive.html	780	blog
http://www.cadius.org/weblog/index.php?cat=46	413	blog
http://www.cadius.org/weblog/index.php?cat=40	413	blog
http://www.cadius.org/weblog/index.php?cat=58	413	blog
http://www.cadius.org/weblog/index.php?cat=50	413	blog
http://www.cadius.org/weblog/index.php?m=200502	412	blog
http://www.bibliometria.com/enlaces	307	blog
http://nouruzi.persianblog.ir/1385/2/	253	blog
http://tati.sappho.net/?m=200503	250	blog
http://en.wikipedia.org/wiki/Human%E2%80%93computer_information_retrieval	186	wiki
http://www.scit.wlv.ac.uk/~cm1993/mycv.html	171	personal page
http://www.huomah.com/Search-Engines/Algorithm-Matters/SEO-Higher-learning.html	157	blog
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/i/Ingwersen:Peter.html	155	database
http://www-csli.stanford.edu/~hinrich/information-retrieval.html	140	resource list
http://library2pointoh.fi/	138	blog
http://masao.jpn.org/d/2006-06.html	132	blog
http://icie.zkm.de/join	117	organization
http://www.webindicators.org/	95	project
http://www.sigir.org/resources.html	91	resource list
http://weblog.ib.hu-berlin.de/?m=200410	63	blog
http://comminfo.rutgers.edu/component/option,com_courses/task,view/sch,17/cur,610/num,551/Itemid,54/	62	course
http://www.db.dk/blar	55	personal page
http://vistoyleido.blogspot.com/2004_10_01_archive.html	44	blog
http://bido.blogspot.com/2004_08_01_archive.html	42	blog
http://ketabnama.blogfa.com/8503.aspx	41	resource list
http://ketabnama.blogfa.com/cat-19.aspx	41	resource list
http://library2pointoh.fi/2009/05/30/library-20-emancipated/	40	blog
http://academic.research.microsoft.com/Author/50345.aspx	38	database
http://library2pointoh.fi/2009/05/	36	blog
http://sunsite.informatik.rwth-aachen.de/dblp/db/indices/a-tree/i/Ingwersen:Peter.html	35	database
http://ir.dcs.qmul.ac.uk/index.php?option=com_content&task=view&id=23&Itemid=44	32	resource list

Table 3. The link pages contributing to the hw*-index*

References

Ingwersen, P.: The calculation of Web impact factors. Journal of Documentation 54, 236—243 (1998).

Larson, R. R.: Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace. in Hardin, S. (Ed.) Proceedings of the 59th Annual Meeting of the American Society for Information Science, pp. 71-78. Information Today, Medford, NJ, (1996). http://sherlock.berkeley.edu/asis96/asis96.html

Almind, T. C., Ingwersen, P.: Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. Journal of Documentation 53, 404-426 (1997).

Rousseau, R.: Sitations: An exploratory study. Cybermetrics 1, issue 1, paper 1 (1997). http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Bar-Ilan J.: The 'madcow disease", Usenet newsgroups and bibliometric laws. Scientometrics 39, 29-55 (1997).

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., Callahan, E.: Invoked on the Web. Journal of the American Society for Information Science 49, 1319-1328 (1998).

Smith, A.: The tale of two Web spaces: Comparing sites using web impact factors. Journal of Documentation 55, 577-592 (1999).

Thelwall, M.: Web impact factors and search engine coverage. Journal of Documentation 56, 185-189 (2000).

Thelwall, M.: Results from a Web impact factor crawler. Journal of Documentation 57, 177-191 (2001).

Thelwall, M.: Extracting macroscopic information from Web links. Journal of the American Society for Information Science and Technology 52, 1157-1168 (2001).

Thelwall, M.: Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites. Journal of the American Society for Information Science and Technology 53, 995-1005 (2002).

Björneborn, L., Ingwersen, P.: Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology 54, 1216-1227 (2004).

Hirsch, J. E.: An index to quantify an individual's scientific research output. PNAS 102, 16569-16572

Schubert, A.: Using the h-index to assess single publications. Scientometrics 78, 559-565.

Bar-Ilan, J.: Expectations versus reality - Search engine features needed for Web research at mid 2005. Cybermetrics 9, issue 1, paper 2 (2005). http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

Address of congratulating author:

Judit Bar-Ilan
Department of Information Science
Bar-Ilan University, Israel
Email: barilaj[at]mail.biu.ac.il

Web Impact Factors – A Significant Contribution to Webometric Research

Kim Holmberg

Åbo Akademi University, Åbo, Finland

Abstract. Ingwersen (1998) developed the Web Impact Factor as a measure of impact or visibility of websites. After that several researchers have tested the idea in various contexts and developed new versions of Web Impact Factors. Different Web Impact Factors have not proven to be as accurate or as useful as the early studies had hoped. Yet the significance of the seminal paper on Web Impact Factors by Ingwersen (1998) was great for webometric research. Webometric research can in fact be said to have started with Ingwersen (1998) and the Web Impact Factor. This chapter reviews earlier research on Web Impact Factors and discusses the usefulness of them and the impact the invention of Web Impact Factors had on webometric research.

1 Introduction

Roughly at the same time Brin and Page (1998), Kleinberg (1999) and Ingwersen (1998) all published their papers on methods to quantitatively measure the web and all papers used hyperlinks as a data source. Brin and Page's (1998) paper became to be the backbone of search engine Google, while Kleinberg (1999) presented an alternative way of using hyperlinks to rank results in search engines. Both papers showed a method to rank websites according to their assumed relevance or impact by counting hyperlinks. Ingwersen (1998) took another approach and showed a method to measure the impact or visibility of websites or areas of the web.

Ingwersen's (1998) Web Impact Factors (WIF) are closely related to Journal Impact Factors. The Journal Impact Factor is the ratio of citations to a journal divided by the number of articles in that journal over certain specified time periods (Garfield, 2005), while WIFs are calculated by dividing the number of inlinks to a certain website with the number of pages on the website. Later, other versions of WIFs such as using the number of staff as denominator (Thelwall, 2002a), has also been introduced and tested in various settings.

Scientific journals that receive more citations to fewer articles are considered to be high impact journals and publishing in them is usually seen as more valuable than publishing in low impact journals. The Web Impact Factor builds on the same idea and looks at hyperlinks technically as citations or recognitions between different websites or different web pages and it was therefore thought to be a measure of impact or the relative visibility a website had on the web (Ingwersen, 1998). A WIF measures the relative visibility of a website by showing how many inlinks a single web page, a website or an area of the web receives. A higher WIF would mean that the site is receiving more inlinks with fewer pages, or attracting more inlinks with less effort. Using WIFs different websites could be compared and the high impact websites or those with greatest influence could be discovered.

2 Use of Web Impact Factors

Ingwersen (1998) defined three different types of Web Impact Factors: internal WIF, external WIF and overall WIF. In the internal WIF only internal hyperlinks within a website are used as the denominator, while all external hyperlinks are used as the denominator for the external WIF. The overall WIF combines both of these hyperlink counts and uses all inlinks. Smith (1999b) argued that external inlinks are the most indicative ones when counting Web Impact Factors, because internal hyperlinks within a website are often navigational hyperlinks. The idea was that external links could be used to measure a website's impact or relative visibility compared with other websites, while internal links are usually made because of some web design decisions and may therefore not indicate visibility. However, both inlinks and outlinks have been suggested to be useful measures of different features of websites or areas of the web. Following the ideas in the seminal paper by Ingwersen (1998), Chu, He and Thelwall (2002) stated that inlinks could be used as a measure of visibility and outlinks could be used as a measure of luminosity of websites.

Because of the technical similarities between citations and hyperlinks and between Journal Impact Factors and Web Impact Factors, many researchers have tried to find correlations between hyperlinks and research productivity or performance. Thomas and Willett (2000) investigated link counts and research performance at department level of universities but did not find any significant correlation. This suggested to the authors that webometric research methods were not accurate enough to be used at the low individual department level. There may be a lot of noise in the link data (e.g. dead links, links in link lists, etc), which for larger data sets would not have as great impact as for smaller data sets. In smaller data sets the noise could skew the data significantly and result in inaccurate results. Both Smith (1999b) and Ingwersen (1998) suggested earlier that Web Impact Factors may be more reliable and useful when applied on larger organizations' websites or larger portions of the web, as the results from smaller units may not be as reliable.

Web Impact Factors have been used to study Australasian web structures (Smith, 1999a) and significant correlations have been discovered between link counts and research ratings in universities in the UK and Australia (Smith & Thelwall, 2002; Thelwall, 2001a; 2002c). In the UK Web Impact Factors calculated from different sources (links from pages on .edu, .ac.uk, .uk domains and the entire web) have been found to correlate with research ratings (Thelwall, 2002b). Methods for data collection developed and improved over time and later Li, Thelwall, Musgrove and Wilkinson (2003) found that in contrast to Thomas and Willett (2000) hyperlink counts at a lower, departmental level correlated with research ratings. The results showed that hyperlink counts can reflect research at the lower departmental level.

Researchers have also developed new variations and modifications of the original Web Impact Factors. Thelwall (2001a) used four different versions of the Web Impact Factor and showed that the WIF delivering the best correlation between link counts and research ratings was the one where links to research related pages and faculty numbers was used. With the tools available filtering research related pages from all the other pages collected may however not be very cost-effective. Another modification, used mainly in studies about the impact of academic websites, was to divide the external links with the number of fulltime staff at the target university (Thelwall, 2002a). Using the number of fulltime staff was thought to give a better indicator of the size of the university than the number of web pages. After all, some web design decisions can have a huge impact on the amount of pages a university has. Li, Thelwall, Musgrove and Wilkinson (2003) also found that WIFs calculated with the number of staff correlated significantly with research ratings in the UK.

Thelwall (2003b) developed two new metrics based on the Web Impact Factor (Ingwersen, 1998). These were Web Use Factor (WUF) and Web Connectivity Factor (WCF). Instead of using the inlink counts, Web Use Factor uses outlink counts and divides these with the fulltime staff of the organization, which were universities in this case. The Web Use Factor measures to what extent links are created out from the university or, in other words, to what extent the Web is used by university staff. The Web Connectivity Factor is calculated by dividing the total number of interlinking links between pairs of universities with the number of fulltime staff. The Web Connectivity Factor should therefore be high for universities that both use and provide more information on the Web (i.e. create and receive many links). Both metrics were found to correlate with research productivity at the institutional level.

Later researchers turned their focus on the motivations for creating hyperlinks (e.g. Thelwall, 2002d; 2003a) and because of the scarcity of links to academic papers both Thelwall (2003b) and Li (2003) suggests that the WIF is measuring the reputations of universities and scholars rather than quality of their research publications. Studies on linking motivations showed how varied the reasons for creating links were and how only a portion of the motivations were related to research or quality of the research at the target universities.

WIFs have also been calculated for some special case of websites, e.g. municipal websites (Holmberg, 2009). Holmberg (2009) calculated different WIFs using a) external inlinks and b) interlinking links between municipalities and dividing them with the 1) number of web pages a municipal website had and with the 2) number of population. These measures were based on the traditional WIF by Ingwersen (1998) and the WIF used by Thelwall (2002a). The motivations for creating the two different types of links were assumed to be different and hence the different WIFs were hypothesized to show different tendencies and measure different aspects of the municipal websites, but all the calculated WIFs showed the same tendency, that smaller municipalities had higher WIFs. As the majority of the inlinks to the municipalities came from link lists of various lengths, it had to be assumed that the municipal websites would have received these links no matter how large or small their websites were. As all the researched municipalities received a certain amount of inlinks, smaller websites simply had higher WIFs because smaller municipalities had smaller websites. Therefore it was concluded that WIFs calculated for municipal websites are foremost indicators of the size of the municipalities and their websites, not quality or value of their websites or activities.

3 Usefulness of Web Impact Factors

Earlier research has not been able to produce very consequent or useful results from Web Impact Factors, which has caused some criticism against the reliability and usefulness of WIFs. The criticism mainly derives from the data collection methods using search engines. Citation counts and impact factors can be manipulated (Gorman, 2005), just like results of search engines, and that is one of the reasons why the use of search engines to collect link data for Web Impact Factor calculations has been questioned (Bar-Ilan, 2002). Search engines also do not cover the whole Web, search engines are biased (Mowshowitz & Kawaguchi, 2002; 2005) and they may be quite easily manipulated (Schwartz, 1998). It is even possible that search engines censor the results (Goldsmith & Wu, 2008). Also, while Journal Impact Factors cover citations made at one point in time to articles made at another point in time, WIFs give a snapshot of a single moment of the search engines database, which may explain why some earlier research about WIFs have in some cases come to different results. Thelwall (2000) concludes that the coverage of search engines is so uneven that using them in Web Impact Factor calculations may give misleading results. Thelwall (2001b) used a web crawler to collect data for calculation of WIFs for universities in the UK. These early results suggested that with certain restrictions, WIFs could in fact be counted reliably but that they do not correlate with research ratings due to the vast variety of material published on universities' web spaces.

The original idea of calculating Web Impact Factors derived from the technical similarities between citations and hyperlinks, however, this is where the similarities between citations and hyperlinks end. While scientific publications usually go through a peer review process to guarantee the quality of the publication, there is no quality control on the web as anyone can publish whatever text or hyperlinks they want. Presumably there are many more different reasons to create hyperlinks than there are reasons to create citations. Links are created for a wide variety of different reasons and different organizations or other areas of the web are probably linked to for different reasons. Municipal websites in Finland are probably not linked to for the same reasons as universities in Denmark. Hence their visibilities cannot really be compared and such comparison would probably not even be a useful one to make.

It is also important to remember that a publication or a document on the web can be a single web page or it can be divided into several web pages, which means that simple web design decisions may have great impact on counting the WIFs. Even with some concerns and criticism, there seems to be some patterns in the link data that correlate with some offline phenomenon, and although WIFs have not proved to be as useful as initially thought, the Web Impact Factors and the idea of counting hyperlinks can be said to have started webometric research.

4 Discussion

Web Impact Factors were originally thought to be a measure of impact a website had on the web, but later research has shown that WIFs are not very reliable or useful measures. Using search engines to collect the number of links may give unreliable numbers, as search engines may e.g. change their ranking algorithms and even hold back some of the results (Goldsmith & Wu, 2008). Although WIFs have been shown to be an unreliable way to measure impact or influence on the Web, the invention of the WIFs had a significant impact on the birth of webometric research. Web Impact Factors showed that there may be some patterns in hyperlinks and with that, they opened the path for many researchers to use hyperlinks as a data source. Links have since then established their role as a measure of visibility on the web and e.g. as a indication of connections, cooperation and even competition.

Today webometric research has grown beyond just counting hyperlinks. Webometric research is adopting methods from different disciplines (e.g. social network analysis) and with the rise and present popularity of various social media Web Impact Factors may take new forms in counting visibility or impact on the web.

Simply the amount of followers one has on Twitter tells something about the popularity, but a Twitter Impact Factor calculated by dividing the number of followers with the number of tweets one has posted could show who has gained the greatest popularity with least amount of effort. Counting inlinks to blogs could tell something about the general popularity of the blogs, but calculating a Blog Impact Factor by dividing the number of inlinks by the number of postings in a blog, could reveal some information about which blog has gained the greatest visibility with the least amount of effort, which could be a useful measure when comparing different blogs and analyzing their performance.

Although the seminal paper on Web Impact Factors by Ingwersen (1998) did not lead to established use of WIFs as a measure of websites' impact, the paper lead to something more important: the birth of a new research field, webometrics. This contribution to webometric research is greater than the one of reliable and useful WIFs could ever have been.

5 References

Bar-Ilan, J. (2002). How much information do search engines disclose on the links to a web page? A longitudinal case study of the 'cybermetrics' home page. *Journal of Information Science*, vol. 28, no. 6, pp. 455-466.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *WWW7 / Computer Networks and ISDN systems*, vol. 3, no. 1-7, pp. 107-117. Retrieved February 22, 2008, from http://infolab.stanford.edu/~backrub/google.html.

Chu, H., He, S. & Thelwall, M. (2002). Library and information science schools in Canada and USA: A webometric perspective. *Journal of Education for Library and Information Science*, vol. 43, no. 2.

Garfield, E. (2005). The Agony and the Ecstasy - The History and Meaning of the Journal Impact Factor. Talk presented at *International Congress on Peer Review And Biomedical Publication*, Chicago, September 16, 2005. Retrieved June 3, 2009, from http://garfield.library.upenn.edu/papers/jifchicago2005.pdf.

Goldsmith, J. & Wu, T. (2008). *Who controls the Internet? Illusions of borderless world.* OUP USA, 2008.

Gorman, G.E. (2005). How do we count our chickens? Or do citation counts count? *Online Information Review*, vol. 29, no. 6, pp. 581-584.

Holmberg, K. (2009). Webometric Network Analysis - mapping cooperation and geopolitical connections between local government administration on the Web. Dissertation. Åbo: Åbo Akademi UP, 2009. (Available online: Permalink to the publication: http://urn.fi/URN:ISBN:978-951-765-511-8).

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, vol. 54, no. 2, pp. 236-243.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, vol. 46, no. 5, pp. 604-632.

Li, X. (2003). A review of the development and application of the Web impact factor. *Online Information Review*, vol. 27, no. 6, pp. 407-417.

Li, X., Thelwall, M., Musgrove, P. & Wilkinson, D. (2003). The relationship between the WIFs or inlinks of Computer Science Departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics*, vol. 57, no. 2, pp. 239-255.

Mowshowitz, A. & Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing & Management,* vol. 38, pp. 141-156.

Mowshowitz, A. & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing & Management*, vol. 41, pp. 1193-1205.

Schwartz, C. (1998). Web Search Engines. *Journal of American Society for Information Science*, vol. 49, no. 11, pp. 973-982.

Smith, A.G. (1999a). ANZAC Webometrics: exploring Australasian Web structures. In *Proceedings of Information Online and On Disc 99*, Sydney, Australia, 19-21 January 1999.

Smith, A. (1999b). A tale of two Web spaces: Comparing sites using Web impact factors. *Journal of Documentation*, vol. 55, no. 5, pp. 577-592.

Smith, A. & Thelwall, M. (2002). Web impact factors for Australasian Universities. *Scientometrics*, vol. 54, no. 3, pp. 363-380.

Thelwall, M. (2000). Web impact factors and search engine coverage. *Journal of Documentation*, vol. 56, no. 2, pp. 185-189.

Thelwall, M. (2001a). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, vol. 52, no. 13, pp. 1157-1168.

Thelwall, M. (2001b). Results from a Web impact factor crawler. *Journal of Documentation*, vol. 57, no. 2, pp. 177-191.

Thelwall, M. (2002a). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites. *Journal of the American Society for Information Science and Technology*, vol. 53, no. 12, pp. 995-1005.

Thelwall, M. (2002b). A comparison of sources of links for academic Web impact factor calculations. *Journal of Documentation*, vol. 58, no. 1, pp. 60-72.

Thelwall, M. (2002c). A research and institutional size-based model for national university Web site interlinking. *Journal of Documentation*, vol. 58, no. 6, pp. 683-694.

Thelwall, M. (2002d). The top 100 linked-to pages on UK university web sites: high inlink counts are not usually associated with quality scholarly content. *Journal of Information Science*, vol. 28, no. 6, pp. 483-491.

Thelwall, M. (2003a). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, vol. 8, no. 3. Retrieved May 25, 2009, http://informationr.net/ir/8-3/paper151.html?text=1.

Thelwall, M. (2003b). Web use and peer interconnectivity metrics for academic Web sites. *Journal of Information Science*, vol. 29, no. 1, pp. 1-10.

Thomas, O. & Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science*, vol. 26, no. 6, pp. 421-428.

Address of congratulating author:

Kim Holmberg
Department of Information Studies, Åbo Akademi University
Fänriksgatan 3, 20500 Åbo, Finland
Email: kim.holmberg[at]abo.fi

Citation Journal Impact Factor as a Measure of Research Quality

Isabel Iribarren-Maestro[1,2] & Elías Sanz-Casado[1]

[1] Carlos III University of Madrid, Getafe, Spain
[2] University of Navarra, Pamplona, Spain

Abstract. The aim of this paper is to determine if the scientific papers published in high impact journals are not only of sufficient "a priori" quality to be accepted by such journals, but acquire further impact because they are cited by journals with at least a similar impact. A normalized impact factor (NIF) is proposed as a measure to compare the visibility of research conducted by different university departments working in different disciplines. This analysis is supplemented by a study of the distribution by quartiles of the journals involved. In addition, the quality of the journals citing the papers in the sample selected for the study is evaluated to determine research prestige. The feasibility of using citation journal impact in the evaluation as an incentive for research quality is posed in this work. This paper will be of interest to those institutions interested in quality research evaluation as well as those involved in science policy.

Key Words. University Departments' evaluation; Visibility; Citation Analysis; Impact Factor

1. Introduction

The development of a method for comparing institutions working in different areas is essential to the analysis of a university environment such as Spain's, with 48 public and 28 private universities. These institutions are evaluated as a whole on a regular basis, often to rank them on the grounds of their scientific activity [1, 2, 3]. Such exercises fail to take the enormous differences among universities into consideration, however. The factor with the greatest effect on the results of such global evaluations is more than likely the area of specialization, for scientists' habits and research results vary substantially from one subject matter to another. In a recent analysis of scientific activity in all Spanish public universities (measured as papers published and cited) on the basis of subject diversity, publishing profiles were found to differ depending on the degree of specialization [4].

This aspect has been studied together with the Impact Factor (IF) indicator. Traditionally, IF has been considered as an impact or quality measure, and all its advantages and limitations have been reviewed in the following studies [5, 6, 7, 8, 9, 10, 11].

In order to solve the IF limitations as well as the bias of comparing universities with different subject profiles, the solution proposed in the present study is to normalize the IF values provided by the ISI by applying a normalized impact factor (NIF). This indicator could then be used to analyze the impact of each department area's output, conduct an inter-area comparison and evaluate the overall impact of the university as a whole. Many "normalized" impact factors are available for scientific literature, such as those developed by the Centre for Science and Technology Studies (CWTS), discussed by Moed [12] in his latest book on citation analysis.

Journal positions by quartiles within the subject areas listed by Journal Citation Reports (JCR) were also analyzed to supplement the NIF information. These two analyses are complementary because IF distribution varies widely across disciplines: i.e., one subject area may have IF values with a very low standard deviation, indicating that they are concentrated around a central value, with all the journals exhibiting a similar NIF, whereas in others the standard deviation may be high, a reflection of substantial differences between NIF values. This makes the information provided by quartile distribution on a journal's position with respect to other journals dealing with the same subject matter particularly useful.

Nonetheless, while quality analysis based solely on a publishing journal's impact factor limits the conclusions that can be drawn respecting its popularity, it furnishes little information on its prestige. So, the quality of the journals where papers are cited (the "citing journals") is also a factor to be considered, according to some authors who think that the impact of the periodicals where a paper is cited should be considered along with the number of times it is cited [13, 14].

Consequently, this study aims to compare the impact both of a sample of papers and of the journals, in which they are cited, on the assumption that measuring the quality of the journals in which papers are cited is an optimal indicator for analyzing the quality of such papers.

The underlying premise is that scientific papers published in high impact journals are both of sufficient "a priori" quality to be accepted by such journals and acquire further impact because the journals where they are cited have a similar impact.

Consequently, the primary objective addressed in this study to verify this premise was to test the suitability of measuring the visibility of citing journals as a method for analyzing the visibility of the articles cited. This objective was pursued by focusing on the following more specific targets: on the one hand, to analyze journals where a given Spanish public university publishes its papers to determine both the NIF and their relative position in the JCR listing (quartile occupied in the respective subject area classification for the period studied); and on the other, to analyze the citing journals' NIF and relative position by quartile, likewise in their subject area classification. Finally, the impact and visibility measures of the two series of periodicals were also compared.

2. Methodology

The case study was defined on the basis of the scientific output of ten Carlos III University of Madrid (UC3M) departments that routinely publish in journals included in ISI databases, taking the information required from the Web of Science.

The university's output was retrieved from the "Address" and "Reprint address" fields in the above database. Each department's production and the records on the respective citations were subsequently normalized.

The period covered was from 1997 to 2003 (extended to 2004 for the citations). The units selected were: from the university's Polytechnic School, Mathematics (MATH), Physics (PHY), Materials Science and Engineering and Chemical Engineering (MAT), Electrical, Electronic and Robot Engineering (ELEC), Mechanical Engineering (MECH), Computer Science (COMP), and Communications Technology (COMM), and from its Social and Legal Science Faculty, Economics (ECO), Business Administration (BUS) and Statistics (STAT).

The indicators used in the study were:

- Normalized impact factors (NIFs) for UC3M publishing and citing journals. This indicator is proposed to obtain the mean impact for department output when several areas are covered and relate it to the mean factors for each respective category. An index was calculated to render any journal's impact factor comparable to any other by relating its IF to the mean IF of the category to which it belongs, or to the mean of the mean IFs for several categories in the event of multidisciplinary journals. This indicator, which measures the real difference between an IF and the mean for the category, is unaffected by the concentration or deviation of the category's IF distribution.

  In this procedure, the following formule has been applied to find the NIF:

$$NIF = \frac{UC3MjournalIF}{\frac{\sum jrnIF\_category}{n}}$$

  where n= No. of journals in each category.

  The value found was then used to rescale each IF to the mean IF of the respective subject category; the result was a comparable inter-category IF. For journals having more than one subject category, their IF has been rescaled as follows:

$$NIF = \frac{UC3MjournalIF}{\dfrac{\left(\dfrac{\sum jrnIF\_category_1}{n_1} + \dfrac{\sum jrnIF\_category_2}{n_2} + \ldots + \dfrac{\sum jrnIF\_category_n}{n_n}\right)}{N}}$$

where N = No. of subject categories.

A department's NIF for a specific year has been calculated as the mean NIF for all the papers produced by that area in the year in question.

The interpretation of the indicator is: if NIF>1, the journal had an IF higher than the mean; where NIF<1, its IF was lower than the mean; if NIF=1, the journal's IF concurred with the mean; if NIF=0, either no IF was available for the journal in the JCR for the respective year, or the value was 0. A few remarks on the NIF for the citing journals are in order:

» This analysis excluded both citing articles with no IF and department's output that was not cited, for their inclusion would have distorted the analysis of department prestige, inasmuch as it would have entailed taking "zero" citation NIFs into consideration. These data were analyzed separately so that information on uncited articles would not be lost.

» When an article was cited more than once, the NIF of the citing journals was not averaged; rather, the citations were aggregated: e.g., in the event of papers receiving several citations, instead of averaging them, each citation was considered individually. For this reason, the impact of citing journals carried more weight in articles with a large number of citations than the mean impact of papers with fewer citations. That is to say, account was taken of both the popularity and the prestige of scientific output.

- Relationship between the UC3M NIF and the NIF of its citing records. This indicator relates the impact of scientific output to the impact of the citations, associating the visibility of each department's published papers with the visibility of the journals citing such papers.
- Distribution of UC3M output and the respective citations by JCR quartile. This technique, commonly used in similar studies [15, 16], consists in dividing the list of publications (ranked by IF in descending order) into quartiles to compare journals in terms of their relative positions, regardless of the subject area or speciality involved. Where journals were assigned to more than one subject area and perhaps positioned in different quartiles in each, only the highest ranking quartile was used.
- Relationship between UC3M output quartiles and citation record quartiles. This indicator used percentage and absolute values to compare the impact of the citing journals to the impact of UC3M output. With this approach, the percentage of citations in each quartile was related to the percentage of pa-

pers in the respective quartile for the university as a whole. Correspondence analysis (CA) was used to analyze the relationship between citing and cited quartiles. This method aims to deduce the relationships between different categories by defining their similarities and grouping them accordingly [17]. The correspondence analysis values obtained were plotted on bubble charts where, in addition to the similarities between variables, a third measure is shown, namely the relative weight acquired by each value when analyzed.

3. Results

3.1 General output and visibility data

By way of introduction to the findings on the relationship between publishing and citing journal visibility, Table I gives the data compiled on each UC3M department's scientific output and the respective citations, ordered by percentage of the latter. The percentage data refer to the respective totals (1462 papers analyzed, 4594 citations).

DEPARTMENT	% UC3M PAPERS	% UC3M CITATIONS	UNCITEDNESS RATE (%)	% SELF-CITATIONS	CITATIONS PER PAPER
MATH	22.63	38.47	23.35	29.16	5.33
PHY	14.30	22.56	23.70	21.26	4.95
MAT	11.04	8.93	49.08	46.97	2.53
STAT	9.76	7.91	39.58	16.94	2.54
ECO	11.31	7.11	43.71	9.42	1.97
MECH	5.76	5.12	36.47	24.89	2.79
ELEC	5.83	2.92	44.19	33.82	1.57
COMM	6.91	2.46	58.82	39.47	1.12
BUS	4.20	2.27	41.94	7.62	1.69
COMP	8.27	2.25	67.21	36.54	0.85

Table I. Output by department and distribution of citations

According to Table I, the Mathematics Department (MATH) accounted for the highest percentage of papers published and had the highest percentage of citations. It was also the department with the highest percentage of citations per paper and the smallest percentage of papers not cited. The Physics Department (PHY) ranked second in each of these indicators. In his analysis of the 100 largest European research universities [18], van Raan also found a relationship between high production and low number of uncited papers.

The department with the smallest portion of papers was Business Administration (BUS) with 4.20% of the documents published, followed by Electrical, Electronics and Robot Engineering (ELEC) and Mechanical Engineering (MECH), with 5.83% and 5.76%, respectively.

The smallest proportion of citations was recorded for the Computer Science Department (COMP), which also had a very high percentage (67.21%) of uncited papers. It was, moreover, the only department that had less than one citation per paper (0.85), as the rate for the remaining departments ranged from 1.12 to 5.33.

Despite the large number of non-uncited papers, these data did not differ substantially from Seglen's finding to the effect that over 50% of articles selected at random from the Science Citation Index had not been cited three years after publication [19].

Since the unit analyzed in this study was the department, self-citations were regarded to be a department's citations of its own papers. They were identified as the citation records in which Carlos III University of Madrid or the respective department was among the affiliations listed. Of the 4594 citations referring to UC3M papers, 1260 (27.42%) were included in papers authored by the university's own researchers. According to Table I, the highest percentage of self-citations was recorded for the Materials Science Department (MAT), where they accounted for nearly one half (46.97%) of the area's visibility. The three departments with the lowest self-citation indices, in turn, were: Business Economy (BUS), with 7.62%, Economics (ECO) with 9.42% and Statistics (STAT) with 16.94%.

Overall, the self-citation rate found in this study was lower than found for Spanish output as a whole in 1999, when the figure was 34% [20] and lower also than the 36% reported for Norwegian publications between 1981 and 1986 [21].

3.2 Normalized Impact Factor (NIF) for UC3M output and respective citations

The NIFs were found for the two series analyzed, i.e., journals publishing UC3M papers and the respective citing journals; the values for the period covered in the study are graphed in Figure 1.

The figure shows that the citing journal NIF was higher than the publishing journal figure in all the years analyzed. More specifically, the decline in the UC3M's impact in 2003 was not mirrored by the citing journals' NIF. The mean impact for the UC3M papers across the entire period analyzed was 1.43: i.e., 43% higher than the mean IFs of the journals in the respective categories.

The mean NIFs for the publishing journals were calculated for all years and broken down by unit of study for an exhaustive analysis of each department's impact and visibility. The same methodology was used to find the NIF for each

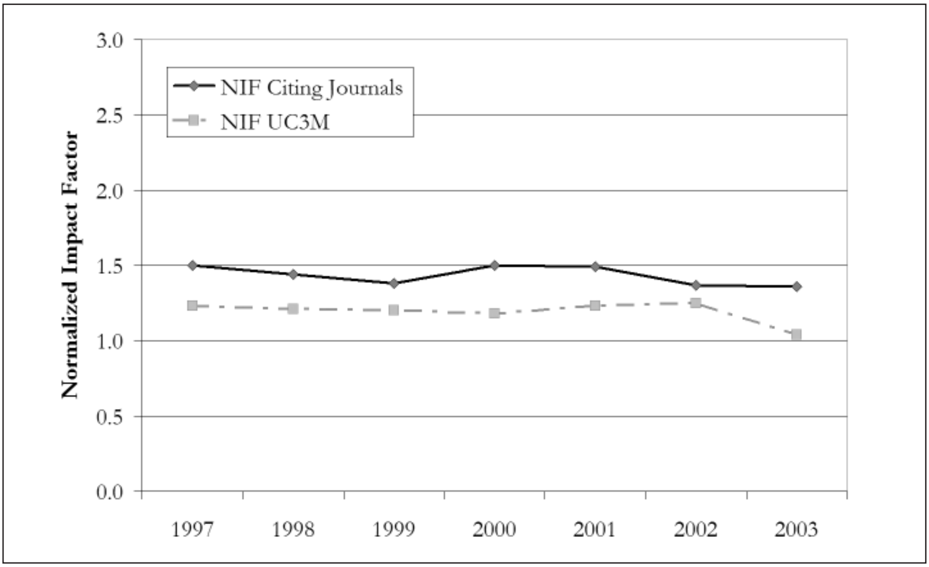


Fig. 1. Comparison of UC3M publishing and citing journal NIFs

department's citing journals. The difference between the two values was then calculated to verify the existence or otherwise of a relationship between publishing and citing journal NIFs. The results are set out in Table II.

Note that the UC3M departments whose citing journals had the highest NIF were the same departments whose papers exhibited the highest impact, namely Physics (PHY) and Mathematics (MATH). Not only did these two units publish in journals with the highest NIF – 1.59 for the former and 1.40 for the latter – ,

DEPART- MENT	NIF FOR AREA/DEPT PAPERS	NIF FOR CITING JOUR.	RATIO (CITING-PUBLIC.)
MATH	1.40	1.65	1.18
PHY	1.59	1.57	0.99
MAT	1.03	1.38	1.34
STAT	0.94	1.02	1.09
ECO	1.02	0.96	0.94
MECH	1.10	1.41	1.28
ELEC	1.12	1.22	1.09
COMM	1.14	1.28	1.12
BUS	0.87	1.13	1.30
COMP	0.81	1.08	1.33

Table II. Difference between citing journal and publishing journal NIF

but their papers were in turn cited in journals with the highest NIF: 1.57 for the Physics (PHY) and 1.65 for the Mathematics (MATH). In this same vein, the only departments having citing journals with a lower impact than their publishing journals were Economics (ECO) and Physics (PHY), although in the case of PHY the ratio between the two indices was 0.99.

The departments showing the greatest variation between the two indicators were Materials Engineering and Science and Chemical Engineering (MAT) and Mechanical Engineering (MECH). In both cases, their papers were cited by journals with a NIF of close to 1.4, but published in journals with a NIF of around 1. Moreover, several departments' publishing and citing journal NIFs barely differed; i.e., they published and were cited in journals with similar visibility. This group included Statistics (STAT), Electrical, Electronic and Robot Engineering (ELEC) and Communications Technology (COMM).

3.3 Relationship between publishing journal quartiles and citing journal quartiles

This section relates the quartiles occupied by the journals publishing university research to the quartiles in which the journals citing these articles are positioned. In this regard, Figure 2 shows each department's percentage output by quartiles, while the quartile distribution of the citing journals is illustrated in Figure 3.

The graph in Figure 2 shows that the departments with the highest proportion of papers in the first quartile were Physics (PHY), Mechanical Engineering (MECH) and Mathematics (MATH), in that order; 70% of the Physics Depart-

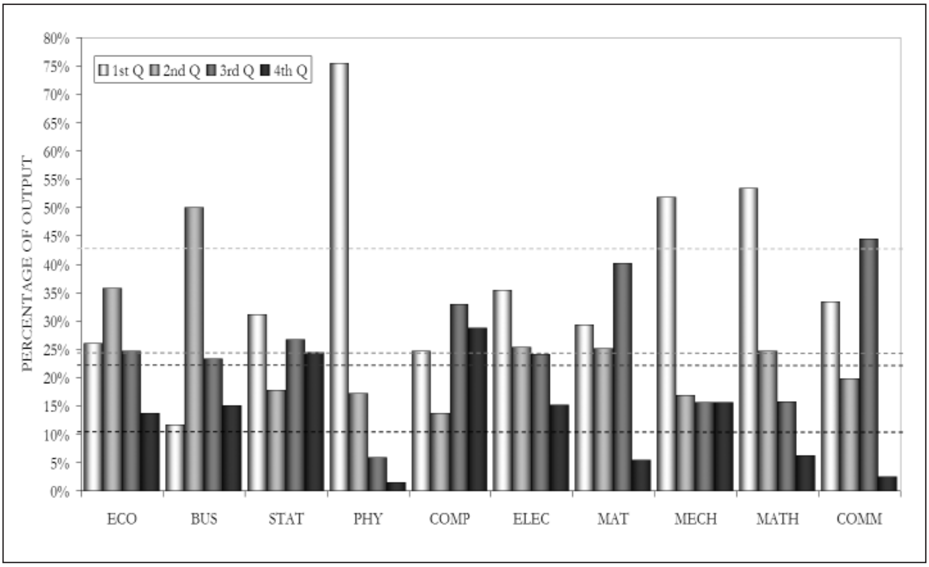


Fig. 2. Department output. Distribution by quartiles

ment output, in fact, was published in first quartile journals. A substantial difference was also observed between the first and second quartile in Mechanical Engineering (MECH), which accounted for 52% and 16% of the department's production, respectively.

Business Administration (BUS) showed low first quartile visibility and was the department with the lowest percentage of publications in this quartile, while half of its output was in the second quartile. Economics (ECO) followed a similar pattern, but with a much larger (double in fact) percentage of first quartile papers and a smaller share of second quartile papers than Business Administration (BUS). The third and fourth quartile percentages were similar for these two social science departments.

Computer Science (COMP) output was concentrated in the third and fourth quartiles, with less than 25% of its papers published in first quartile and less than 15% in second quartile journals.

Figure 3, which gives the quartile distribution of citing journals, shows that Physics (PHY), Mechanical Engineering (MECH) and Mathematics (MATH) had a larger proportion of first quartile citations than the other UC3M departments. Around 70% – 72% for Mechanical Engineering (MECH) – of the references to papers produced by these three departments appeared in first quartile journals.

Other departments in which first quartile citations prevailed were: Computer Science (COMP), Electrical, Electronic and Robot Engineering (ELEC), Materials Science (MAT), Communications Technology (COMM) and Statistics (STAT).

In Economics (ECO), the quartile distribution for citations differed substantially from the overall pattern, for most (30.38%) of its citations was positioned in the third,

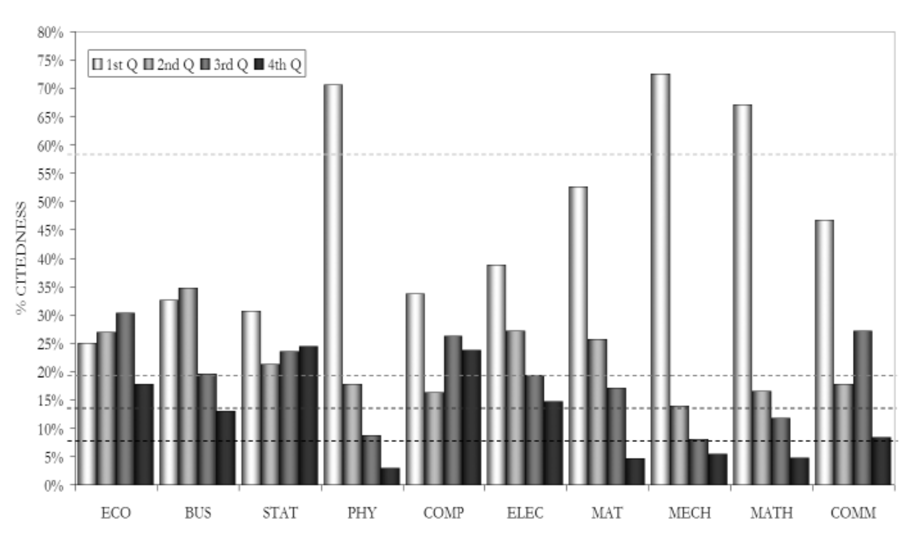


Fig. 3. Department citations. Distribution by quartiles

		CITING JOURNALS			
		1ST Q	2ND Q	3RD Q	4TH Q
UC3M PUBLISHING JOURNALS	1ST QUARTILE	69.87%	15.72%	9.75%	4.65%
	2ND QUARTILE	40.05%	28.15%	19.89%	11.91%
	3RD QUARTILE	24.63%	26.83%	34.88%	13.66%
	4TH QUARTILE	28.37%	23.40%	24.82%	23.40%

Table IV. Relationship between citing journal and publishing journal quartiles

followed by the second (26.90%) and first (25%) quartiles. Most (35%) of the Business Administration (BUS) paper citations, in turn, were found in second quartile journals, although followed closely by first quartile periodicals (32%).

The two variables are analyzed jointly in Table IV, in which the rows denote publishing journal quartiles and the columns citing journal quartiles. The value in each cell indicates the percentage of citations appearing in journals in a given quartile with respect to the total number of citations received by papers published in journals in that quartile.

The chi-square value obtained, throughout the absolute values, 591.95 [v=9; 16.92 at 95% probability], evinced the existence of a correlation between the quartiles of the journals where UC3M researchers publish their papers and the quartiles of the journals where such papers are subsequently cited.

According to Table IV, 69.87% of the citations of university papers published in first quartile journals were found in first, 15.72% in second, 9.75% in third and 4.65% in fourth quartile journals.

The highest proportion of citations of second quartile papers (40.05%) also appeared in first quartile journals, followed in descending order by 2nd, 3rd and 4th quartile citing journals.

Most of the third quartile paper citations (34.88%) appeared in third quartile journals. The next largest proportion of citations of papers in this quartile was found in second quartile journals, followed by first and fourth quartile journals, in that order.

Finally, the citations of papers published in fourth quartile journals were distributed rather evenly across citing journal quartiles, ranging from 23.40% in the 2nd and 4th to 28.37% in the first quartile.

Correspondence analysis explains the relationship between two variables. Here it was used to determine the relationship between the quartile in which each department published its results (small circles with departments' labels and number of quartile in Figure 4) and the quartiles citing its papers (big circles and labels composed by C-citing- and the quartile in Figure 4).

This itemized analysis by department shows that in most cases, when a department published its papers in first quartile journals, its citations were predominantly published in first quartile periodicals. Figure 4 shows how close the Mechanical

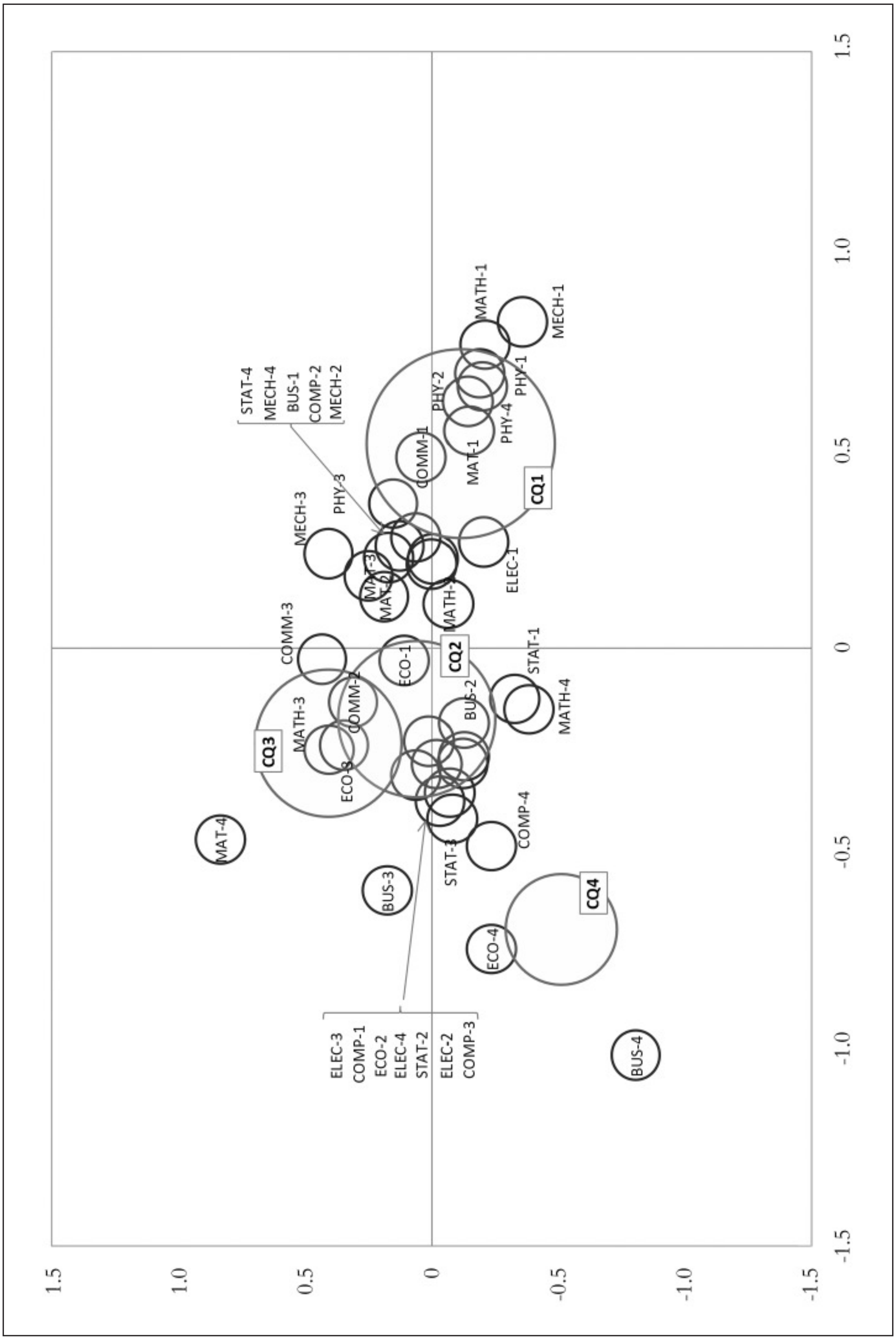


Fig. 4. CA. UC3M production quartiles vs citing journal quartiles

Engineering (MECH), Mathematics (MATH) and Communications Technology (COMM) departments were to the first quartile of citing journals.

Another significant finding was that regardless of the quartile in which they were published, Physics (PHY) and Mechanical Engineering (MECH) papers were primarily cited in first quartile journals. The Economics (ECO) and Business Administration (BUS) department papers, in turn, were cited by journals in the same quartile as the publishing journals.

4. Discussion and Conclusions

The findings of this study show that the methodology proposed is suitable for evaluating institutional quality on the grounds of citing journal impact.

The year-by-year analysis of the findings for the present sample shows that on the whole, UC3M papers were published in journals with a higher than average IF, i.e., a Normalized Impact Factor (NIF) higher than one. Moreover, these papers were cited in journals with a high NIF, on the order of 1.4, in all the years studied. The department breakdown shows the Physics (PHY) and Mathematics (MATH) areas to be particularly prominent in this regard, for while they published in journals with a high NIF, their papers were cited in journals with an even higher factor.

This study revealed that the departments exhibiting the largest difference between publishing and citing journal NIFs were not the ones that published in high impact journals. The reason is obvious, because if an article is published in a journal with a very high impact factor, the possibility of its being cited in journals with even higher IFs is smaller than if it were published in a lower impact journal. Consequently, –while the papers authored by Physics (PHY) and Mathematics (MATH) had the highest impact, they were not the departments with the most favourable difference between citing and publishing journal NIF.

The comparison between the publishing and citing journal quartiles for the various departments showed that the majority of the citations referring to papers published in first quartile appeared in journals in the same quartile. Most of the citations for papers published by the Physics (PHY) and Mechanical Engineering (MECH) departments appeared in first quartile journals, while 68% of the Mathematics (MATH) department citations were also found in the first quartile. The lowest visibility was recorded for Economics (ECO), Statistics (STAT), Business Administration (BUS) and Computer Science (COMP) departments.

In any event, researcher concern about the impact of the journals where they publish may be counterproductive in certain cases, if the journals preferred are not the ones read by the target audience [22]. Indeed, failure to reach the right researchers may determine a smaller number of citations and therefore lower impact. For this reason, researchers should be cautious when choosing the vehicle for transmitting their findings, in addition to seeking publication in high impact journals. The recent trend in scientific evaluation to assess citations makes it preferable to publish in journals whose content and readership are well suited to the type of research addressed.

Lastly, the feasibility of using citing journal impact as an incentive for research quality should be explored. Spanish evaluation agencies, for instance, presently measure researchers' careers in terms of the impact of the journals where they publish their papers; as a result, papers may be published in high impact journals but never cited. In other words, is research quality measured more objectively in terms of the IF of the journal where an article is published or of the number of times it is cited? This study found that departments such as Physics (PHY), Mathematics (MATH) and Mechanical Engineering (MECH), that publish in high impact journals, normally had a higher rate of citations per paper; moreover, their citing journals had a higher impact than the periodicals chosen for publication. Therefore, taking assessment one step further and evaluating the quality of citing journals would not initially appear to jeopardize the sample analyzed. Nonetheless, this practice is regarded to be more suitable to meso- and macro-studies. Inasmuch as citations are sometimes affected by sociological factors, individual researchers may encounter difficulties if their research is assessed on the grounds of citation quality.

Along the lines proposed by Bollen [13], the present paper confirms the premise that even though a given paper may be frequently cited, the quality of such citations may not necessarily be high, whereas other papers may be cited more sparingly, but in high impact journals. This poses the question of whether it is preferable to be profusely cited in mediocre journals or more occasionally in high prestige periodicals. The former option may be a sign of popularity and an indication that the information is being widely used yet the latter is preferable, in principle, for the inference is that papers published in journals with a higher impact are consulted and cited by researchers of greater prestige.

Acknowledgments: We would like to thank Ronald Rousseau for his helpful comments on early draft of this research work.

References

1. Buela Casal, G., Bermúdez, M.P., Sierra, J.C., Quevedo-Blasco, R. & Castro, A.: Ranking de 2008 en productividad de investigación de las universidades públicas españolas. Psichotema, vol. 21, 2, pp. 304-312 (2009)

2. Gómez Caridad, I., Bordons, M., Fernández, M.T., Morillo, F., Structure and research performance of Spanish universities. Scientometrics, vol. 79, 1, pp. 131-146 (2009)

3. Sanz-Casado, E., Iribarren-Maestro, I., García-Zorita, C., Efrain-García, P., Sánchez-Gil, S.: Are Productivity, Impact and Visibility Indicators Appropriate for Measuring the Quality of Research Conducted in Universities? En: LARSEN, B. & LETA, J. Proceedings of the International Conference on Scientometrics and Informetrics, vol.1, pp. 286-290 (2009)

4. García-Zorita, C., Iribarren-Maestro, I., Rousseau, R., Sanz-Casado, E.: Publication and citation inequality in the Spanish University System. En: Larsen, B. & Leta, J. Proceedings of the International Conference on Scientometrics and Informetrics, vol.2, pp. 932-933 (2009)

5. Moed, H. F.: The impact-factors debate: the ISI's uses and limits. Nature, vol. 415, pp. 731-32 (2002)

6. Frandsen, T. F., Rousseau, R., Rowlands, I.: Diffusion Factors. J Doc, vol. 62, 1, pp. 58-72 (2006)

7. Glänzel, W.: Science in Scandinavia: a bibliometric approach. Scientometrics, vol. 48, 2, pp. 121-50 (2000)

8. Glänzel, W., Schubert, A.: Double effort = double impact? A critical view at international co-authorship in Chemistry. Scientometrics, vol. 50, 2, pp. 199-214 (2001)

9. Glänzel, W., Thijs, B.: Does co-authorship inflate the share of self-citations? Scientometrics, vol. 61, 3, pp. 395-404 (2004)

10. Persson, O., Glänzel, W., Danell, R.: Inflactionary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, vol. 60, 3, pp. 421-32 (2004)

11. Garfield, E.: How can impact factors be improved? Brit Med J, 313, pp. 411-413 (1996)

12. Moed, H. F.: Citation Analysis in Research Evaluation. Dordrecht: Springer (2005)

13. Bollen, J., Rodríguez, M. A., Sompel, H. V. d.: Journal Status. Scientometrics, vol. 69, 3, pp. 669-687 (2006)

14. Ball, P.: Prestige is factored into journal ratings. Nature, vol. 439, 16 February, pp. 770-771 (2006)

15. Iribarren-Maestro, I.: Producción científica y visibilidad de los investigadores de la Universidad Carlos III de Madrid en las bases de datos del ISI, 1997-2003 [Doctoral Thesis]. Elías Sanz-Casado (dir.). Getafe: Departamento de Biblioteconomía y Documentación, Universidad Carlos III de Madrid (2006)

16. Gómez Caridad, I., Fernández Muńoz, M. T., Bordons Gangas, M., Morillo Ariza, F.: La producción científica española en Medicina en los años 1994-1999. Rev Clin Esp, vol. 204, 2, pp. 75-88 (2004)

17. Carrasco, J. L., Hernán, M. A.: Estadística multivariante en las ciencias de la vida. Madrid: Cibest; Ciencia 3 (1993)
18. Van Raan, A. F. J.: Bibliometric Statistical Properties of the 100 Largest European Research Universities: Prevalent Scaling Rules in the Science System. Journal of the American Society for Information Science and Technology, vol. 59, 3, pp. 461-75 (2008)
19. Seglen, P. O.: The skewness of science. J Am Soc Inf Sci, vol. 43, 9, pp. 628-38 (1992)
20. Glänzel, W., Thijs, B., Schlemmer, B.: A bibliometric approach to the role of author self-citations in scientific communication. Scientometrics, vol. 59, 1, pp. 63-77 (2004)
21. Aksnes, D. W.: A macro study of self-citation. Scientometrics, vol. 56, 2, pp. 235-46 (2003)
22. Bordons, M. Hacia el reconocimiento internacional de las publicaciones científicas españolas. Revista Española de Cardiología, vol. 57, 9, pp. 799-802 (2004)

Addresses of congratulating authors:

Isabel Iribarren-Maestro[1,2]

[1] Laboratory of Information Metric Studies (LEMI), Librarianship and Information Science Department, Carlos III University of Madrid, C/Madrid, 126, Getafe 28903 Madrid (Spain)
[2] Library, University of Navarra, Apdo. 177, 31080 Pamplona (Spain)
Email: iiribarr[at]bib.uc3m.es

Elías Sanz-Casado
Laboratory of Information Metric Studies (LEMI), Librarianship and Information Science Department, Carlos III University of Madrid, C/Madrid, 126, Getafe 28903 Madrid (Spain)
Email: elias[at]bib.uc3m.es

Amado is Everywhere

Jacqueline Leta

Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro, Brazil.

Abstract. This paper aims to present data on scientific publications from Bahia, a Brazilian state. This is a tribute to Peter Ingwersen, who visited and loved Salvador, Bahia's capital and main city.

Keywords: Bahia, scientific output, Brazilian science

1 Introduction

During the last two decades Brazilian science has enlarged significantly (1). Different from most of developed countries, science in Brazil is extremely concentrated on the public sector, especially universities located in the country's southeast region. According to the Ministry of Science and Technology almost 70% of Brazilian scientists works for the public sector (2). The large concentration of scientists in the public sector pushes these institutions to be the most productive of the country (3).

As for the contribution of Brazilian regions, Leta & Brito presented a detailed scene of the Brazilian states' and regions' scientific productivity (4). Southeast region, where Rio de Janeiro and São Paulo are located, is by far the most productive in terms of scientific publications. According to the authors, Bahia, the largest state located in the northeast region, is among of the twelve most productive Brazilian states, being responsible for 2-3% of the country's international publications.

Hence, although Bahia can not be considered as a core state within Brazilian science system, this short communication aims to present some data on Bahia's science and scientific publications. This is a tribute to Peter Ingwersen, who visited and loved Salvador, Bahia's capital and main city.

2 Bahia's scientific human resources

According to the Brazilian Institute of Geography and Statistics, Bahia is one of largest Brazilian states. It encompasses 564,692,669 Km2 and its population is estimated in 14,600,000, for 2009 (5). With all this size, the state of Bahia can be considered a country. There is an enormous diversity in culture, geography, climate, population, faith.

	2000		2004		2008	
	Bahia	Brazil	Bahia	Brazil	Bahia	Brazil
Research Groups	330	11,760	728	19,470	1,090	22,797
Researchers	2,113	66,804	4,833	119,205	8,307	159,948
Students	1,887	63,512	4,133	113,654	8,737	177,702

Table 1: Number of Bahia's and Brazil's research groups, researchers and students. Source: CNPq, 2010 (7)

Type	Number	% of total in country
Grant – research project	398	3.465
Fellowship – Undergraduate	952	3.72%
Fellowship – Master	280	2.71%
Fellowship – PhD	154	1.70%
Fellowship – Post-Doc	28	2.39%

Table 2: Main grants and fellowships awarded to Bahia's researchers, 2009. Source: CNPq, 2010 (8)

And Salvador, its capital, is the Brazilian city where the Negro's culture and traditions, from those who came to Brazil and slaves, is preserved. Thus, Bahia as well its capital are effervescences of tastes, sounds, parties, people, an unique setting in Brazil. (6)

As for science, data on the three last censuses indicated the number of Bahia's research groups, researchers and students increased 3-fold or more (Table 1). An increase much higher than that observed to Brazil.

Bahia's main research institute is the Federal University of Bahia, known as its short name UFBA. For 2008, UFBA encompassed 348 out of the 1,090 research groups (31.9%), 2,273 out of the 8,307 researchers (27.4%) and 2,526 out of the 8,737 students (28.9%).

According to the National Counsel of Technological and Scientific Development, researchers from the state of Bahia were awarded with more than 4,000 fellowships and research grants in 2009. Table 2 present some of this awards.

3 Bahia's scientific output

Scientific publications written by Brazilian researchers, especially by those from Bahia, were searched in Scopus, by using a simple query string: *Bahia* AND *Brazil* in the address search. The numbers of publication by years as well as details of 2009 publications – subject area – were collected.

Figure 1 presents time trends of Bahia's and Brazil's publication, according to Scopus. In the period, Bahia's publications increased from 38 to 890. Such quantitative increase was followed by an increase in the share of Bahia in Brazil's publication: from 1.3%, in 1990, to 2.4%, in 2009.

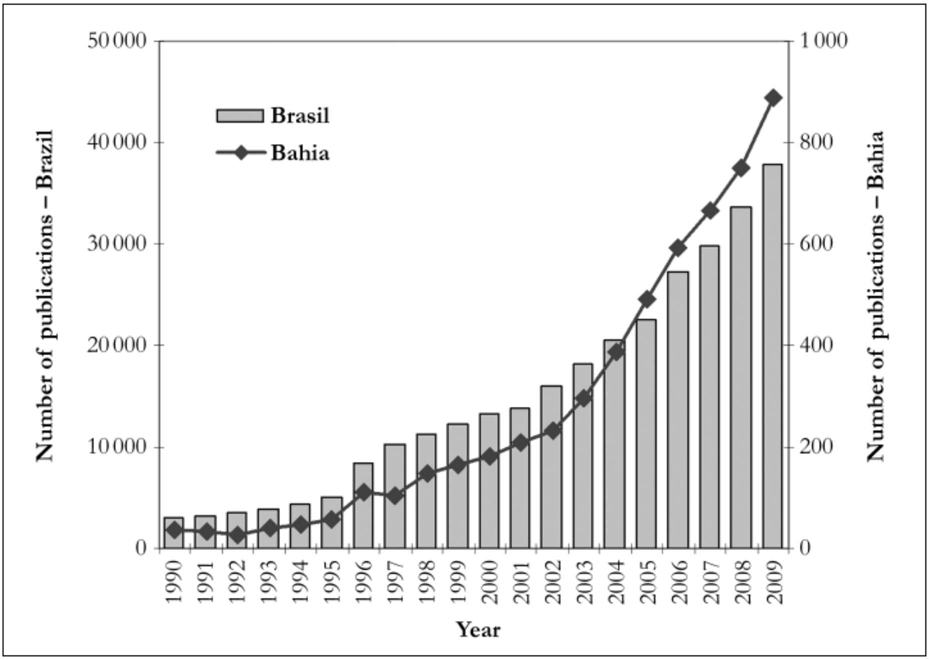


Figure 1: Number of Brazil's and Bahia's scientific publications indexed in SCOPUS, from 1990 to 2009.

As for the main fields the Bahia's publications, it is clear that the field coverage increased substantially in the period (Table 3). Publications in 1990 were related to 11 fields while in 2009 they were related to 28 fields. However, Medicine keeps being the main field in the whole period. More recently, humanities, arts and social sciences do appear as important fields of Bahia's scientific publications.

The large presence of Medicine field in Bahia's publication has to do with the state large tradition in the field. The first Brazilian Medical Scholl was founded in Salvador, Bahia's capital and main city, in 1808, when the whole Portuguese Royal family moved to Brazil. Bahia Surgery School, its former name, was incorporated by UFBA, by the time it was formally founded in 1946. Today, UFBA and its Medicine Faculty are nationally recognized as reference centers in medicine, in terms of both research and services.

4 Conclusion

Although Bahia's scientific output and human resources are enlarging, the state still plays a peripheral role within the country's whole science. Nevertheless, it is unquestionable its role and status recognition in the medical field.

1990	
Fields	Publ.
Medicine	16
Earth & Planetary Sc	9
Materials Sc.	8
Physics&Astronomy	8
Immun& Microbiol	4
Agric & Biol. Sc.	3
Environmental Sc.	3
Bioch, Genetics & Molecular Biol	2
Chemistry	1
Engineering	1
Pharm, Toxicology & Pharmaceutics	1

2000	
Fields	Publ.
Medicine	49
Agric & Biol. Sc	30
Chemistry	28
Physics&Astronomy	28
Immun& Microbiol	18
Earth & Planetary Sc	17
Chemical Eng.	16
Mathematics	16
Bioch, Genetics & Molecular Biol	14
Environmental Sc.	14
Pharm, Toxicology & Pharmaceutics	9
Engineering	8
Materials Science	6
Neuroscience	4
Veterinary	2
Social Sciences	2
Multidisciplinary	2
Computer Science	2
Health Professions	1
Energy	1
Business, Manag	1

2009	
Fields	Publ.
Medicine	330
Agric & Biol. Sc	198
Chemistry	101
Bioch, Genetics & Molecular Biol	99
Physics & Astronomy	68
Computer Science	68
Engineering	60
Immun& Microbiol	54
Chemical Eng.	44
Mathematics	43
Environmental Sc.	42
Earth & Planetary Sc	40
Pharm, Toxicology & Pharmaceutics	31
Materials Science	29
Veterinary	24
Neuroscience	19
Dentistry	17
Nursing	16
Health Professions	10
Multidisciplinary	9
Social Sciences	9
Energy	8
Psychology	7
Decision Sciences	5
Arts and Humanities	5
Business, Manag	3
Econ, Econometrics	1

Table 3: Main fields of Bahia's publications, 1990, 2000 and 2009.

But, more than it: it is unquestionable Bahia's role to keep preserving and alive Negro's traditions, which are the basis of Brazilian identity. Everyone should visit the state and its capital: Peter and Irene visited! And that was the true motivation behind this paper.

The first time I met Peter was at the ISSI Conference in Stockholm, in 2005. Very talkative, lively and funny, Peter had clearly something from Latin Americanness, which made us closer immediately. In fact, it seemed – at least to me – that we knew each other long time ago. During the 2005 Conference, Peter was a key person; he explained to me and gave me tips on how to organize a conference in the field. At that time, I was submitting a proposal for the 2009 Conference to be held in Brazil.

All assistance and collaboration I had from Peter needed, I thought, to be somehow reciprocated. The opportunity came soon after the conference! Peter and Irene were coming to Brazil, to attend an international seminar in Salvador, the Brazilian first capital! As they were going to stay some days more, they needed some tips. Immediately, I wrote a long, long message to both with tips on places to visit, local food and drinks as well as local culture, especially a writer, Jorge Amado (1).

Just one month latter, Peter wrote me a message where he said "both Irene and I wish to thank you very much for your advices and profound information on Brazil and Salvador in particular. We enjoyed very much the entire trip (…) Amado was everywhere (…)".

Hence, this short communication presented a little about science and scientific output published in this part of the country: Bahia. Hope the scientific data touch Peter as much as Pelourinho, Elevador Lacerda, Farol da Barra, Praia do Forte, acarajé, caruru, umbuzada, caipirinha .. did!

References

1. Glänzel, W., Leta,J.,Thijs, B. (2006) Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics, Vol. 67 (1) 67–86.*
2. Ministério de Ciência e Tecnologia, *Indicadores nacionais de ciencia e tecnologia. 2005 Dados sobre recursos humanos*. Data available at: http://www.mct.gov.br/estat/ascavpp/ingles/3_Recursos_Humanos/tabelas/tab3_4_1.htm
3. Leta,J., Glänzel, W., Thijs, B (2006) Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics, Vol. 67 (1) 87–105.*
4. Leta, J., Brito Cruz, C.H. A produção científica brasileira. In: Indicadores de C&T&I no Brasil. Centro de Gestão e Estudos Estratégicos, Brasília: Brasil, 2003. (Report organized by the Brazilian Ministry of Science & Technology)
5. Brazilian Institute of Geography and Statistics. Bahia. Available at: http://www.ibge.gov.br/estadosat/perfil.php?sigla=ba
6. For those interested in Salvador and Brazilian best writers: http://www.emtursa.ba.gov.br/ and http://www.klickescritores.com.br/imortais.htm

7. National Counsel of Technological and Scientific Development. Diretório de Grupos de Pesquisa. Available at: http://dgp.cnpq.br/planotabular/

8. National Counsel of Technological and Scientific Development, Mapas de Investimentos do CNPq. Available at: http://efomento.cnpq.br/efomento/distribuicaoGeografica/distribuicaoGeografica.do?metodo=apresentar

Address of congratulating author:

Jacqueline Leta
Instituto de Bioquímica Médica
Universidade Federal do Rio de Janeiro, Brazil
Email: jleta[at]bioqmed.ufrj.br

On the Interface?

Ed Noyons

Centre for Science & Technology Studies (CWTS), Leiden, The Netherlands

I don't know the story of his life, let alone when he was a child or an adolescent, but for it is clear that Peter has become what he is because of his youth.

Peter must always have been in between choices. Maybe his parents divorsed when he was very young or probably he grew up in a neighbourhood right in between two complete opposite quarters. These two must have had very different cultures. The one to the left was probably high class, while the one to the right were slums. While the one had the advantage of friends who had money, toys, the other had the advantage of fun and real friends. Those opposite hoods you find in many towns and cities. Peter must have lived somewhere in between.

During the early years of his life he must have struggled as to which hood he should go to find his friends. Everyday again, always in doubt: "will I go to the left where I can enjoy luxury, lots of opportunities but boring as hell. Or will I go to the right where I find the friends that really care. Always fun, but no money and too little ambition." Let's just suppose Peter grew up in such a context. Because this must have made him the man he has become, the man we all know.

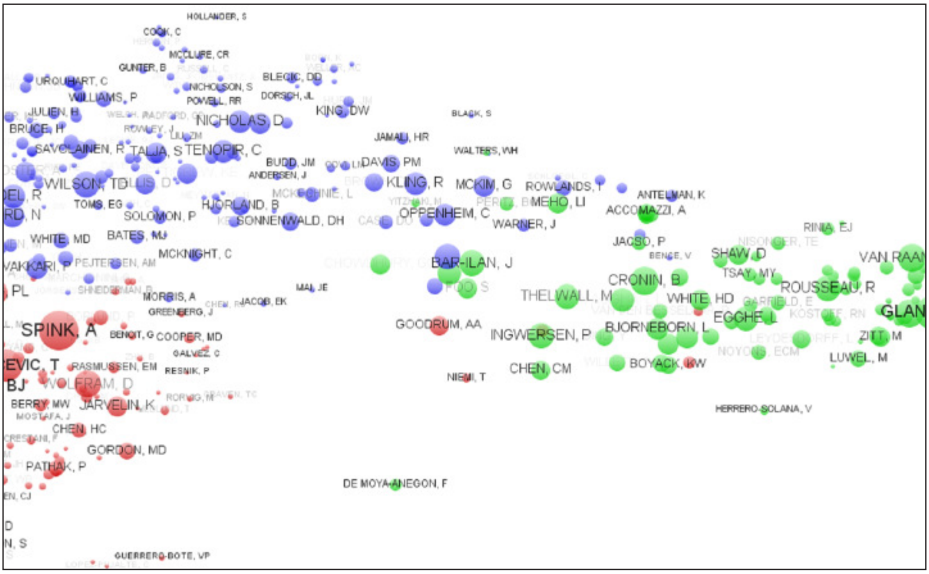


Figure 1: Author co-citation map of Information Science (c.f., www.vosviewer.com/maps)

Because the man we know is a man who knows how to deal with such situations. He has become an expert on the interface. As we all know, he acts on the interface of scientometrics and information retrieval. He had a history in IR for quite some time but that was not satisfactory. He must have thought that is to one-dimensional. He needed to live on the edge, so he flirted with scientometrics. At first with caution to see what was going on there, at ISSI but after a while he was charmed by the empirical work and was particularly working to make a link with IR through mapping techniques. You can check the (co-)author maps that have been created about information science. You will always find Peter somewhere between scientometrics and IR or Librarians.

At this interface I met him and got inspired by him. I experienced him not only as someone on the interface but also as someone who knows very well how to act in this in between area.

I found out how about his ability some ten years ago in Leiden. It was at the celebration diner for my PhD graduation. Peter had been one of my referees during the defence. As usual he had been one of the most prominent referees. He is charming, naughty and funny. What more do you need? Anyway, during the diner we celebrated occasion and during this session waiters would come and ask if you would prefer a glass of red or a glass of white wine. We all recognize this moment of doubt. Some people make their choice on the basis of the food they are having, other choose on the basis of the region or grape of the wine. But we all have this short moment of doubt. Usually in the middle of a conversation a short moment of silence accompanies us. Our brain is stuck in this moment of confusion. While discussing the future of scientometrics, the chances of world peace, German science policy or the French policy on drugs legislation, we have to make a choice between red or white wine.

At the diner we all stop and ponder, but Peter. He has already made his choice. Not for one or the other, but for both. "I'll have a glass of each" were his famous words. Not on the interface actually, but with his foot firmly in both.

The next day I thanked him for being my opponent by offering him a glass of rose, a wine in between red and white. At the interface, I thought. He appreciated the thought but added: "In that case, I will have two glasses." That's our Peter!

Address of congratulating author:

Ed Noyons
Leiden University, NL
Centre for Science & Technology Studies (CWTS)
Willem Einthoven-gebouw, Wassenaarseweg 62A
Postbus 905, 2300 AX Leiden, The Netherlands
Email: noyons[at]cwts.nl

An Overview of Collaboration in Global Warming Research in Africa, 1990-2008[1]

Dennis N. Ocholla[1], Omwoyo Bosire Onyancha[2] & Lyudmila Ocholla[1]

[1] University of Zululand, South Africa
[2] University of South Africa, South Africa

Introduction

Global warming is increasingly becoming a major area of multidisciplinary research right so because of the growing interest and concern of the causes and consequences of the emerging catastrophe that requires proactive intervention before it is too late as attested by some recent studies(Walther et al. 2002; Mathews 2007; Robick et al. 2003; Berger et al. 2005; IPCC, 2007). A recent study by Ocholla and Ocholla(2008) notes that research publications in the domain since 1990 has increased by over 300% and that a total of 116 countries produced one or more publications on global warming, with the USA (2572; 35.7%), England (834; 11.6%) and Japan (546; 7.6%) leading the pack with 3952 (54.85%) publications. In contrast, notes this study, the contribution of African countries to global warming research exists though insignificant, as noted by the participation of 18 (of 53) countries, with South Africa (46), Kenya (14) and Egypt(7) being among the top contributors. Research collaboration between individuals, institutions and countries is increasingly beneficial (Katz & Martin 1997) and inevitable (Rao and Raghavan 2003, 230).This paper presents preliminary findings of an ongoing study on the trends and patterns of collaboration in global warming research in Africa. The study answers the following research questions: Which countries collaborate with the African countries in global warming research? What is the contribution of each of the collaborating country, in terms of the number of co-authored papers, during the study period? What is the comparative regional and international collaboration in global warming research in Africa? What is the degree and strength of collaboration of each country during this study period?

[1] This is a tribute to Peter Ingwersen who is well known to LIS Scholarly community in South Africa for his contribution to informetrics and information retrieval research largely through his initiation, together with others, of Dissanet(see www.dissanet.com) and ProLISSA biennial conferences that have been held in the country since 2002 and become a traditional forum for popularising LIS research in the country. Peter's contribution to research capacity building in the two areas has culminated into Masters and doctorate graduates and the hosting of 13th ISSI conference in Durban, South Africa 4-8th July 2011 where he is also co-chair of the conference organising committee. We are proud to be associated with this humble, distinguished, straight forward/no nonsense scholar.

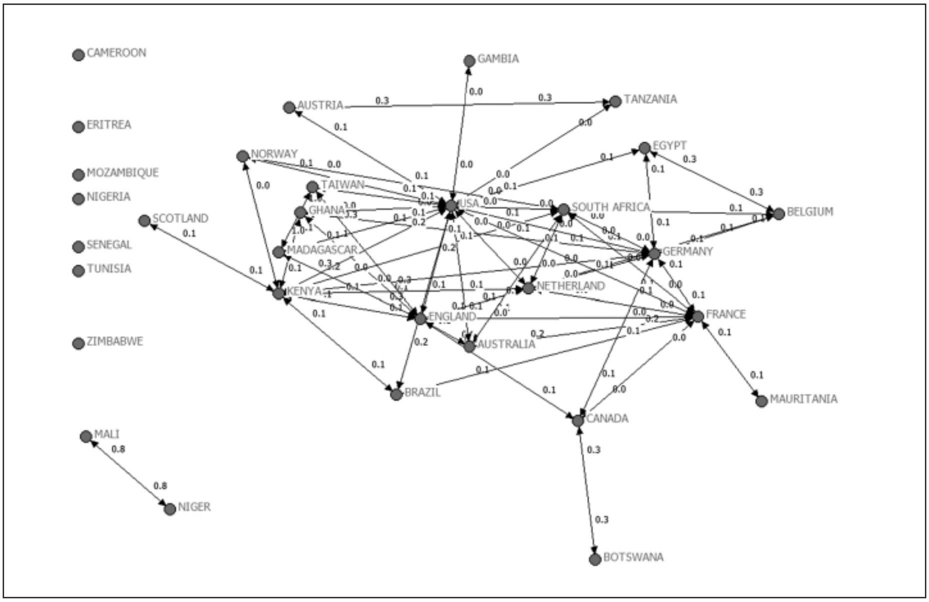


Fig 1: Ego network for African countries engaged in GWR research

Method and materials

The study used the widely accepted indicator of research collaboration, i.e. the co-authorship of papers, to measure country-wise global warming research collaboration in the selected African countries from 1990 to 2008. Data was extracted from the Thomson Scientifics' Science Citation Index (SCI®) and Social Sciences Citation Index (SSCI®) by using global warming as the preferred keyword and the names of the countries. The search was conducted within the author's address and the keyword fields. Bibliographic details of the relevant papers produced by and on the African countries were extracted for analysis. Using publications count, domestically and internationally co-authored papers and major collaborating countries between 1990 and 2008 were identified. The counting of country-wise co-authorships considered the co-occurrence of the African country with another country in the address field of each record. A country was counted only once, irrespective of how many times it appeared with the African country in the address field of the same record. Relevant data (i.e. name of author; title of publication; publication source; and author's country) was downloaded and recorded in Microsoft Excel spreadsheets. Data analysis was conducted using the TextSTAT, TI and UCINET for Windows computer softwares. Two files, namely *words.txt* (containing the names of individual countries and generated by TextSTAT) and *text.txt* (containing the names of collaborating countries) were created and sub-

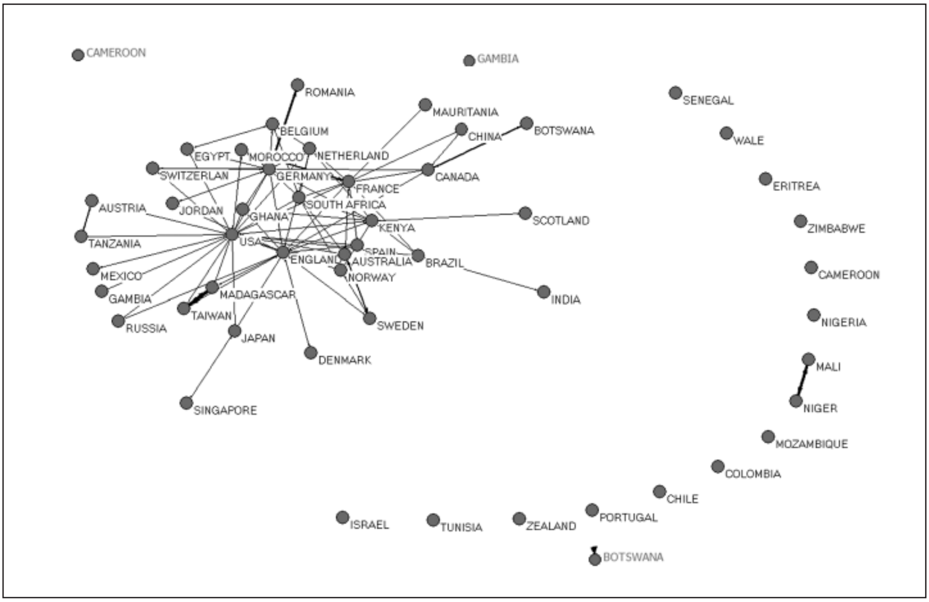


Fig 2: Social network of collaboration in GWR in Africa

jected to analysis using the TI software which produced two matrix files: COOCC. DBF and COSINE.DBF. The COOCC file consisted of raw frequency counts of the co-occurrence of two countries in the address field (thereby identifying the co-authored publications) and the COSINE.DBF comprised the normalized frequency counts of the co-authored publications. The normalized count of co-authored publications indicated the strength of association between collaborating countries. The PAJEK software was used to draw the social networks represented in Figs 1 and 2. Further analysis of the data was conducted to determine the degree of collaboration by computing the collaboration co-efficient, originally introduced by Ajiferuke in 1988 (Rao & Raghavan 2003,).

Results

Domestic vs cross-border collaboration

Domestic collaboration refers to partnership of two or more authors from two or more institutions situated within the same country while cross-border collaboration can be defined as partnership between two or more authors whose institutions are located in different countries. Out of the total 117 articles that were co-authored, 67 were internationally co-authored while 50 were co-authored by authors from the same country. The USA yielded the largest number of domestically co-authored papers (i.e. 17) followed by France (8), South Africa (6), England (4), Peoples Republic of China (3) while Argentina, Japan, Mozambique and Portugal produced 2 domestically co-authored articles each. In this category of collaboration, there was one paper each for Chile, Italy and Japan.

One-author documents were the majority (64; 23.6%) of 271 documents that provided the names of authors. Others were collaborative publication of 2(79; 29.2%), 3(43;15.9%),4(36,13.3%) and 5(24;8.9%) author papers. In total, there were 808 authors responsible for the authorship of 271 documents with most of the papers co-authored (76.4).

Degree of collaboration

The collaboration co-efficient was used as an indicator of the degree of collaboration. The number of hits recorded for each country represents the number of times a given name of a country appeared in the address field of a given record. The number of hits is therefore either greater than or equal to the total number of publications produced by a particular country. The number is greater in situations where there is more than one name of institution situated in the same country. The highest number of hits was recorded by the USA (i.e. 119) followed by France (54), South Africa (48), England (37), Germany (34), Kenya (23), Japan (17), The Netherlands (16), China (14), and Australia (10). A similar pattern was witnessed in the analysis of the publication pattern of co-authored papers. The highest number of co-authored papers came from the USA (44) while France yielded a total of 23 publications followed by England (19), Germany (15), Kenya (13), South Africa (12), The Netherlands (8) and China and Japan which produced 6 publications each.

Contrary to the aforementioned patterns of publication of multiple-authored papers, the measurement of the degree of collaboration (i.e. the collaboration co-efficient) ranks the most productive countries of multiple-author publications poorer than the less productive countries. The highest *cc* scores were recorded by China, Argentina, Brazil, Italy, Botswana, Chile, Denmark, Jordan, Madagascar, Sweden, Portugal, and Mali. Others are: Morocco, Mexico, Singapore, Taiwan, Mozambique, Norway, and Romania. These countries recorded 100% collaboration. It should however be noted that these countries produced a total number of publications ranging between 1 and 6. The most productive countries in terms of the total number of publications yielded between 3 and 75 publications. Their *cc* scores ranged between 0.33 and 0.80. In the descending order of performance, the countries include: The Netherlands (0.80), Belgium (0.75), Ghana (0.75), France (0.72), Kenya (0.68), Canada (0.67), and Austria (0.67), just to name a few. The last bunch of countries is those that recorded collaboration co-efficient of zero. They are: Eritrea, Israel, Cameroon, Colombia, Senegal, New Zealand, Tunisia, Wales, and Zimbabwe, implying that they did not participate with any other country in GWR in Africa.

Number of co-authored documents by a pair of countries

Raw frequency counts reveal that the largest number of co-authored documents (23) was jointly published by the USA and France while the partnerships between France and Morocco and Kenya and the USA yielded a total of 16 publications each. It was noted that 29/67 (i.e. 43.3%) of cross-border collaborations contained at least one name of an African country. Most publications (i.e. 38) on global warming were co-authored among foreign countries. There was therefore little collaboration endeavors among African countries, implying minimal regional collaboration.

Strength of collaboration among the countries

The strengths of collaboration between and among different countries – is indicated by the normalized frequency counts. The highest score (indicating strongest association or partnership in GWR research) was recorded by Niger and Mali (i.e. 0.8). Evidently, and as aforementioned, most collaborative research was conducted among the foreign countries. This collaboration may be originating from African scholars in the Diaspora collaborating with resident scholars or among foreign nationals.

Social networks of GWR collaborating countries

Fig 1 and 2 demonstrates the social networks of countries collaborating with African countries in GWR. Fig 1 illustrates an ego network of selected nodes (i.e. nodes representing African countries only) while Fig 2 is a social network of the entire group of countries engaged in GWR. Fig 1 identifies a total of 17 African countries involved in GWR. Seven of these countries do not have any links with any other country, thereby indicating that there was no research collaboration with each other, on the one hand, and with the rest of the African countries and the world. Fig 2 provides two clusters of countries that recorded at least one link to one other country in GWR and a number of stand-alone countries which did not have any links to/from any other country. The large cluster situated on the left hand side of the illustration consists of both domestic collaboration (i.e. collaboration among African countries) and international collaboration (i.e. collaboration between an African country and a foreign country and/or collaboration among foreign countries). Countries that did not exhibit any collaborative links include Israel, Tunisia, New Zealand, Portugal, Chile, Colombia and Mozambique. Others are: Nigeria, Cameroon, Zimbabwe, Eritrea, Wales, and Senegal.

Conclusion

Internal/domestically co-authored papers were slightly (47.7%) less than internationally oriented in general and for Africa in particular. Only South Africa (6) and Mozambique (2) produced domestically co-authored papers. The degree of collaboration and the co-authorship pattern were closely related. The measurement of the degree of collaboration/collaboration co-efficient ranks the most productive countries of multiple-author publications poorer than the less productive countries. Among the African countries, South Africa (48) and Kenya (23) produced the highest collaboration co-efficient. It was observed that at least 29(43.3%) of cross-border collaboration spotted at least one name of an African country. However, most publications (38 of 67) on GW were co-authored among foreign countries thereby implying minimal regional collaboration. Regarding social networks on GW, there were 10(of 17) countries (Botswana, Gambia, Ghana, Kenya, Madagascar, Mali, Mauritania, South Africa and Tanzania) that recorded at least one link with one another showing some collaboration. We noted that collaboration within and between countries were loose, sporadic and did not produce any logical pattern. For example, it was not possible to link the nature of collaboration with the countries colonial past despite frequent collaboration between France and an African country. Further research will extend the domain by using Scopus and Google Scholar as well as apply other methods for unraveling GW research.

References

Berger, A., Melice, J.L. and Loutre, M.F. 2005. On the origin of the 100-kyr cycles in the astronomical forcing. Journal of Paleoceanography, Vol.20,1-9. [Online]. http://www.agu.org/pubs.Accessed April 10, 2008

IPCC .2007. Intergovernmental Panel on Climate Change, Fourth Assessment Report, "Climate Change 2007: The Physical basis", Summary for policy Makers.[Online] http://www.ipcc.ch/SPM2feb.pdf. Accessed 20 April 2003

Katz, J.S. and B.R. Martin. 1997. What is research collaboration? *Research policy* 26(1): 1–18.

Mathews, J. 2007. Seven Steps to curb global warming: Energy Policy 35, 4247-4259.[Online] http://www.eslsevier.com/locate/enpol. Accessed 18 April

Ocholla, Dennis N. and Ocholla, Lyudmila (2008) Research output and scientific impact of global warming research productivity and literature from 1980 -2007. An informetric analysis .Fifteenth Jubilee Crimea Conference 2008, Crimea, Sudak, Ukrain, 7th – 15th June. [Online] http://www.gpntb.ru/win/inter-events/crimea2008/eng/cd/157.pdf

Walther, G.O., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J.C., Fromentin, J.M., Guldberg, O.V., and Bairlein, F. 2002. Ecological response to recent climate change. Review article. *Nature,* Vol 416, 389-391.

Addresses of congratulating authors:

Dennis N. Ocholla
University of Zululand
South Africa
Email: docholla[at]pan.uzulu.ac.za

Omwoyo Bosire Onyancha
University of South Africa
South Africa
Email: bonyancha[at]unisa.ac.za

Lyudmila Ocholla
University of Zululand
South Africa
Email: locholla[at]pan.uzulu.ac.za

The Janus Faced Scholar

Olle Persson

Umeå university, Umeå, Sweden

Introduction

> ”Janus was usually depicted with two heads facing in opposite directions. According to a legend, he had received the gift to see both future and past from the god Saturn in reward for the hospitality received.” (http://en.wikipedia.org/wiki/Janus)

This quotation perfectly matches my picture of Peter. Peter has always had a strong sense of what is coming in our field. Maybe the best example is webometrics, which Peter started with the papers on informetric analysis of the web, and calculation of web impact factors (Almind & Ingwersen 1997, Ingwersen 1998). Fifteen years before that, Peter was a research fellow at ESA in Italy. Here he developed the zoom command which enabled a lot of interesting online bibliometric studies (Ingwersen 1983). In fact that was one of the main reasons for my early success in bibliometrics. I showed a Swedish R&D group, that were planning a trip to Japan, how to use the zoom command to find Japanese scientists and labs in research on gallium arsenide as a semi conducting material (Persson 1984). The Japanese hosts were quite impressed by the deep and detailed knowledge their Swedish guests had about them.

Peter can also look into the past having been part of our field from its start some forty years ago.

More importantly, Peter is looking in two directions at the same time. One face is looking in the direction of information searching and seeking, and the other towards informetrics. We can all sense that. I will show that this is really the case by presenting some bibliometric maps.

Data and method

From Web of Science I downloaded all Peter's papers, and the papers citing any of his work as defined by a cited reference search. All in all 1334 was found. With the help of BibExcel (Persson, Danell & Schneider 2009), all cited first authors were extracted from the cited reference field, duplicate names within a reference list were removed, then all authors cited in at least 100 papers were selected, and finally the co-citations among the authors selected were calculated. Next, from the same set of papers 5972 direct citation links between the downloaded papers were identified. The weight of these links (WDC) was calculated by adding the number of shared references and the number co-citations within the citation graph (Persson 2010), and these weights vary between 1 (no indirect links) to 76 (sum of shared references and co-citations). To give a few examples, Table 1 lists the direct citation links with the highest weights. All of them are within webometrics.

Weighted Direct Citation (WDC)	Citing paper	Cited paper
76	Smith, 1999, V55, P577, A tale of two web spaces: Comparing sites using web impact factors	Ingwersen, 1998, V54, P236, The calculation of Web impact factors
69	Ingwersen, 1998, V54, P236, The calculation of Web impact factors	Almind, 1997, V53, P404, Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'
66	Thelwall, 2001, V52, P1157, Extracting macroscopic information from Web links	Ingwersen, 1998, V54, P236, The calculation of Web impact factors
50	Smith, 2002, V54, P363, Web impact factors for Australasian universities	Ingwersen, 1998, V54, P236, The calculation of Web impact factors
47	Bjorneborn, 2001, V50, P65, Perspectives of webometrics	Ingwersen, 1998, V54, P236, The calculation of Web impact factors
45	Thelwall, 2002, V53, P995, Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites	Thelwall, 2001, V52, P1157, Extracting macroscopic information from Web links
45	Snyder, 1999, V55, P375, Can search engines be used as tools for web-link analysis? A critical view	Ingwersen, 1998, V54, P236, The calculation of Web impact factors

Table 1. The strongest direct citation links between papers citing the works of Peter Ingwersen

Results

Figure 1 looks like two prisms connected by Peter. The left side has scholars within information seeking and searching, while the right part contains several leading names from informetrics. Except from being co-cited I guess Peter have read something of everyone and probably met all of them. On the other hand, I doubt that all on the left side know everyone on the right side and vice versa. I would then suggest that Peter is the ultimate social and cognitive key person of our field. Ask Peter, and he will guide you in any direction!

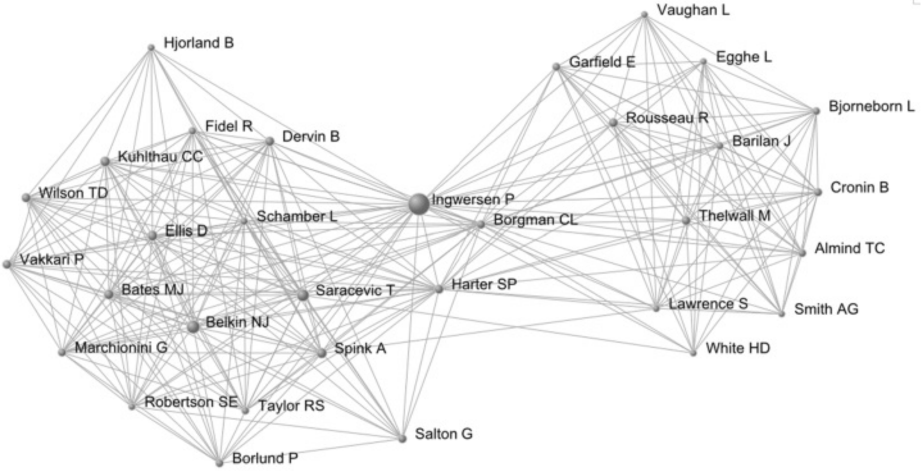


Figure 1. Author Co-citation Analysis (ACA) –map. Authors co-cited in at least 100 papers and co-citations >=25

Figure 2 and 3 shows the citation links among the papers. Figure 2 is hardly readable since it contains all citation links. But we can see two parts emerging. If we

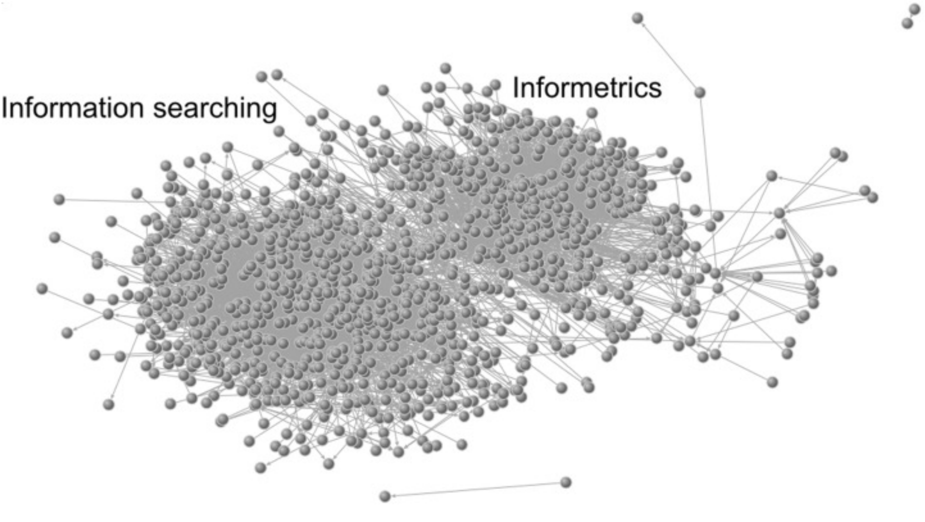


Figure 2. Weighted Direct Citation (WDC)–map on article level. All 5972 links included.

reduce the network, only allowing papers cited at least 5 times and citation links with a weight value of at least 5, then the graph becomes much easier to read. Interestingly, online bibliometrics comes out as a small part close to the dense webometrics area.

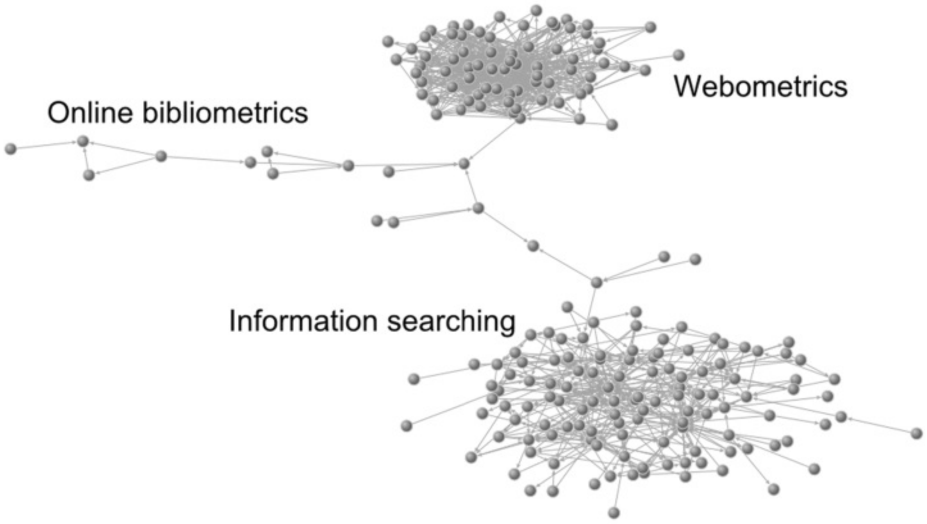


Figure 3. Weighted Direct Citation (WDC)–map on article level. Indegree >=5 and WDC-value > =5.

Conclusion

Peter is a Janus faced scholar. He has always been looking ahead for new indicators and concepts. As one of the most influential researchers in the field, his historical roots are strong and deep. Peter has played, and is still playing a social and intellectual key role in the two sub domains of our fields, informetrics and information searching.

Peter has other virtues too, however not easily discovered in the scholarly literature. Hospitality is one of them. He has opened his home for many of us and is the most active organizer of conferences and workshops all over the world. Saturn must love him.

Talking about Saturn, Galileo sent this anagram to Kepler in 1610, to patent his discovery of Saturn's rings:

s m a i s m r m i l m e p o e t a l e u m i b u n e n u g t t a u i r a s

And the solution of the anagram reads: Altissimum planetam tergeminum observavi.

"I have observed the highest (most distant) planet [Saturn] to have a triple form."

Here is my suggestion: For all scientists and innovators, why not open "The Web Site of Scientific Anagrams" to which you could post an anagram describing a new discovery. Then, if someone else would claim to have made the same discovery, the anagram will protect your priority to it.

By the way, "Peters Renewing", "Steepen Wringer", "Weeping Sterner", and "Sneering Pewter" are all anagrams of "Peter Ingwersen". So there is also a way of hiding the name of the inventor.

References

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. *Journal of Documentation, 53*(4), 404-426.

Ingwersen, P. (1983). Information in Italy. *Journal of Information Science, 6*(2-3), 91-94.

Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation, 54*(2), 236-243.

Persson, O. (1984). *Svensk elektronik : en bibliometrisk studie av Engineering index online = [Electronics in Sweden] : [a bibliometric study of Engineering index]*. Umeå.

Persson, O. (2010). Identifying research themes with weighted citation links. *Journal of Informetrics*. (Forthcoming; published online)

Persson, O., Danell, R. & Schneider, J. (2009). How to use Bibexcel for various types of bibliometric analysis *Celebrating scholarly communication studies : A Festschrift for Olle Persson at his 60th Birthday* (pp. 9-24): International Society for Scientometrics and Informetrics.

Address of congratulating author:

Olle Persson
Department of Sociology
Umeå University
SE-901 87 Umeå, Sweden
Email: olle.persson[at]soc.umu.se

Bibliographic Coupling and Co-citation as Dual Notions

Ronald Rousseau [1,2,3]

[1] KHBO (Association K.U.Leuven), Oostende, Belgium
[2] University of Antwerp, Antwerpen, Belgium
[3] K.U.Leuven, Leuven (Heverlee), Belgium

Abstract: The notions of bibliographic coupling and co-citation, originally defined for articles, can be generalized in many directions. Some examples are provided such as author bibliographic coupling, key-word co-mentioning and departmental co-citation. Care must be taken to precisely define these derived notions and the corresponding coupling and co-citation strengths.

Keywords: bibliographic coupling; co-citation analysis; duality; key-words

1 Introduction

Bibliographic coupling and co-citation analysis are two informetric techniques that originate from information retrieval in citation databases, but which gradually became more important for describing and mapping the structure of science or one of its subfields. In this article, dedicated to Peter Ingwersen, we recall some definitions related to different forms or applications of bibliographic coupling and co-citation analysis.

2 Duality between bibliographic coupling and co-citation

2.1 Bibliographic coupling: article level [1],[2]

Consider two articles *a* and *b* and their reference lists. These reference lists are considered as sets of publications (cited publications). Next one considers the intersection of these reference lists. If this intersection is non-empty one says that articles *a* and *b* are bibliographically coupled. The number of items in this intersection is called the bibliographical coupling strength. The bibliographical coupling strength divided by the number of items in the union of the two reference lists is called the relative coupling strength. This is essentially a Jaccard index [3]. It has been shown in [4] how bibliographic coupling can be performed online, using a clever combination of the RANK and TARGET commands in Dialog. Nowadays, software for this type of analysis is made freely available by Loet Leydesdorff at his website http://www.leydesdorff.net/ (among others).

2.2 Co-citation: article level [5],[6]

Consider two articles *x* and *y* and their sets of citing articles, e.g. articles that cite article *x* respectively article *y*. As usual these citing articles come from a pool of articles such as all journals covered by the WoS or by Scopus. This pool will not be mentioned further on, but will, in practice, always influence numerical outcomes. One considers the intersection of these citing articles. If this intersection is non-empty then articles *x* and *y* are co-cited. The number of elements in this intersection is called the co-citation strength. Dividing the co-citation strength by the number of articles in the union of the two citing sets yields the relative co-citation strength, which is also a kind of Jaccard index [3].

A bibliographic coupling relation is determined by cited articles, while a co-citation relation is determined by citing articles. If *x* and *y* are co-cited then there is at least one article such that articles *x* and *y* both belong to that article's reference list (cited set). If articles *a* and *b* are bibliographically coupled then there is at least one article such that articles *a* and *b* both belong to its citing set. Bibliographic coupling and co-citation are clearly dual notions.

2.3 Co-citation: author level (provisional definition)

Replacing the term 'article' by 'author' leads to the following definition of author co-citation. Consider two authors X and Y and for each of them the set of all articles that cited (at least one of) their articles (published during a given period, cited over possibly another period). One considers the intersection of these citing articles. If this intersection is non-empty authors X and Y are co-cited.

In Rousseau & Zuccala [7] a hierarchic classification of author co-citations is proposed consisting of: pure first-author co-citations, pure co-citations, general co-citations and co-author/co-citations. The definition of co-citations on the author level as given above is actually the definition of the notion of co-author/co-citation. Indeed, if X and Y have co-authored one article and if this article is cited in article T then author X as well as author Y occur in the list of authors cited in T, hence are co-cited according to the previous definition. As the co-authorship relation is usually considered to differ from a co-citation relation, we change the definition of co-citation on the author level as follows.

2.4 Co-citation: author level (based on the notion of pure co-citations)

Consider two authors X and Y. We denote by J(X,Y) the set of all publications co-authored by X and Y (and possibly other colleagues). For X and Y we consider the sets of articles that cited (at least one of) their publications that do not belong to the set J(X,Y) (as before published during a given period and cited over possibly another period). One considers the intersection of these citing articles sets. If this intersection is non-empty authors X and Y are co-cited, in the pure co-citation sense. The number of elements (different citing articles) in this intersection is called the author co-citation strength. Dividing the co-citation strength by the number of articles in the union of the two citing sets yields the relative author co-citation strength.

If X has authored one article that is co-cited with one article of Y (not co-authored by X) then authors X and Y are author co-cited and conversely, i.e. if X and Y are author co-cited then there exists at least one article written by X and one article written by Y (not co-authored by X) such that these articles are co-cited.

Besides this definition of author co-citation strength we also propose the following one in which one takes all actual co-citation occurrences into account. In this proposal the author co-citation strength is defined as the sum of all article co-citation strengths, where the sum is taken over pairs of articles (x,y) where article x is written by author X and article y is written by author Y, excluding all X-Y co-authored articles. This author co-citation strength is called the total author co-citation strength of authors X and Y. In order to distinguish between this method and the previous one the result of the first counting method will be referred to as simple co-citation strength.

Yet, other definitions of co-citation strength are feasible. We illustrate this by a simple example (see Table 1).

Author X	Author Y
Article A_1 cites article X_1	Article A_5 cites article Y_1
Article A_2 cites article X_2	Article A_1 cites article Y_1
Article A_2 cites article X_3	Article A_2 cites article Y_1
Article A_3 cites article X_3	Article A_3 cites article Y_2
Article A_3 cites article X_4	Article A_3 cites article Y_3
Article A_4 cites article X_4	
Article X_5 is not cited	

Table 1. Citations of articles written by author X (X1 to X5) and by author Y (Y1 to Y3)

Author X has written 5 articles and author Y has written three articles (none co-authored with author X). The intersection of the citing articles consists of articles

$\{A_1,A_2,A_3\}$, hence the simple co-citation strength is 3. The union of the two article sets of Table 1 consists of $\{A_1,A_2,A_3,A_4,A_5\}$, hence the relative simple co-citation strength is 3/5 = 0.6. The total co-citation strength is calculated based on the couples: (X_1,Y_1) in A_1; (X_2,Y_1) in A_2; (X_3,Y_1) in A_2; (X_3,Y_2) in A_3; (X_3,Y_3) in A_3; (X_4,Y_2) in A_3 ; (X_4,Y_3) in A_3. Hence the total co-citation strength is 7. The relative total co-citation strength has not yet been defined above. It can be defined with respect to all articles that are at least cited once (and all citing articles), or with respect to all published articles (and again all citing articles). In the first case the maximum number of possible co-cited pairs is 4x3, leading to a value of 7/(12x5) ≈ 0.12 (there are five citing articles); in the second case the maximum number is 5x3, leading to a value of 7/(15x5) ≈ 0.11.

2.5 Alternative counting methods

Alternative I. One may for each co-citing article, here the set $\{A_1,A_2,A_3\}$, count the number of articles written by X or Y, and add these results. In the example of Table 1 this is: 2+3+4= 9. This is another way of determining a co-citation strength. The relative co-citation strength is then obtained by dividing by the maximum number of articles that may be cited. If uncited articles are taken into account this is 5x(5+3) = 40, leading to a relative co-citation strength of 9/40 ≈0.225; when only cited articles are taken into account this leads to a relative co-citation strength of 9/[5x(4+3)] ≈ 0.26.

Alternative II. In this alternative each citing article is weighted according to the minimum number of cited X and Y articles. In the example A_1, A_2 have weight 1; A_3 has weight 2, while A_4 and A_5 have zero weight, leading to a co-citation strength of 4. The maximum weight for each citing article is min(5,3) = min (4,3) = 3, hence the relative co-citation strength using this alternative is 4/(3x5) = 0.27.

As we do not want to make this contribution too complicated, we will not mention these alternatives anymore, although, for instance, a version of alternative II *is* used in practice, see [8]. These alternatives moreover illustrate that there are many ways possible for determining a co-citation strength and a relative co-citation strength.

It is well known that author co-citation analysis has been introduced by White and Griffith [9]. Which definitions did they use? Although it is not completely clear to us, it seems that they used the simple co-citation strength. Moreover, they only considered first authors.

A similar problem occurs when we try to define bibliographic coupling on the author level.

2.6 Bibliographic coupling: author level (provisional definition)

Consider two authors A and B and for each of them the set of all articles they have cited (articles published during a given period, with cited articles published over possibly another period). Then one considers the intersection of these articles lists. If this intersection is non-empty one says that authors A and B are bibliographically coupled.

Again we run into the problem that if authors A and B have co-authored an article then all items in the reference list of the co-authored article belong to the intersection of all articles cited by A and by B. Hence, it is best to exclude co-authored publications, leading to the following definition.

2.7 Bibliographic coupling: author level

Consider two authors A and B and for each of them the set of all articles they have cited in other articles than in those for which A and B are co-authors are considered (published during a given period, cited articles published over possibly another period). Then one considers the intersection of these article lists. If this intersection is non-empty one says that authors A and B are author-bibliographically coupled. The number of items in the intersection is called the simple author bibliographical coupling strength. The simple author bibliographical coupling strength divided by the number of items in the union of the two cited article lists may be called the relative simple author bibliographical coupling strength. If one article written by A and one article written by B (not co-authored by A) are bibliographically coupled, then A and B are author bibliographically coupled and conversely, i.e. if A and B are author bibliographically coupled then there exists at least one article written by A and one article written by B (not co-authored by A) such that these articles are bibliographically coupled.

Also for author bibliographical coupling other methods to determine the author bibliographical coupling strength may be applied. As announced earlier we will only consider the dual notion of total author bibliographical coupling strength. This notion is defined as the sum of all article bibliographical coupling strengths, where the sum is taken over pairs of articles (a,b) where article a is written by author A and article b is written by author B, excluding all A-B co-authored articles. A relative total author bibliographical coupling strength may be defined in a similar way as for the co-citation case.

The notion of author bibliographical coupling has been introduced by Zhao and Strotmann [8]. They used what we refer to as bibliographic coupling on the author level, but, as they only consider first authors they avoid the problem related to co-authorship. In defining the author bibliographic coupling strength they consider the number of times an article occurs in an author's reference lists. If article C_1 occurs two times in author A's reference lists and 4 times in author B's reference lists then this article contributes min(2,4) = 2 to the author bibliographical coupling strength. The author bibliographic coupling strength of authors A and B is then the sum of all these minimum values. This means that Zhao and Strotmann use the author bibliographical coupling strength variant of alternative II above.

2.8 A weaker form of author co-citation and author bibliographical coupling

2.8.1 Author co-citation: weak form

Consider two authors X and Y. Again we denote by J(X,Y) the set of all publications co-authored by X and Y (and possibly other colleagues). For X and Y we consider the sets of authors that cited (at least one of) their publications that do not belong to the set J(X,Y) (as before published during a given period and cited over possibly another period). One considers the intersection of these citing author sets. If this intersection is non-empty authors X and Y are weakly author co-cited. The term weakly is used as X and Y do not need to be co-cited. It just suffices that there is an author that has cited X and also (possibly in another article) author Y. The number of elements in this intersection is called the (simple) weak author co-citation strength. Dividing the weak author co-citation strength by the number of authors in the union of the two citing author sets yields the (simple) relative weak author co-citation strength. If X and Y are co-cited then they are automatically also weakly co-cited. Similarly we define a weak form of author bibliographical coupling.

2.8.2 Author bibliographic coupling: weak form

Consider two authors A and B and for each of them the set of all authors they have cited in other articles than in those for which A and B are co-authors (published during a given period, cited authors publishing over possibly another period). Then one considers the intersection of these two author lists. If this intersection is non-empty one says that authors A and B are weak author-bibliographically coupled. The number of items in the intersection is called the simple weak author bibliographical coupling strength. Again we note that A and B do not have to be bibliographically coupled. It suffices that there is an author that has been cited by A and (in another article) also by B. The weak author bibliographical coupling strength divided by the number of authors in the union of the two cited author lists may be called the simple relative weak author bibliographical coupling strength. If A and B are bibliographically coupled then they are automatically also weakly author bibliographically coupled.

3 Other forms of co-occurrence

Bibliographical coupling and co-citation can be viewed as special cases of co-occurrences. Besides co-citation and bibliographic coupling also other co-occurrences have been studied. It seems that the first such study was performed by Karl-Erik Rosengren in 1966. In 1968 Rosengren published a book called "*Sociological* *aspects of the literary system"*, in which he introduced the co-mention approach in order to graphically display the reference frame of fiction book reviewers [10],[11]. In literary reviews of fiction books he identified the most frequently mentioned authors and their co-occurrences. From these data he was able to draw a map with authors as nodes while the distances between them were estimated using their co-occurrences [12]. Similarly Fano [13] was a precursor of Kessler in the case of bibliographical coupling.

Key-words may also be used in co-occurrence studies. These will be considered in the next section.

4 Articles and key-words

Instead of considering reference lists as in the definitions of co-citation analysis or bibliographic coupling given above, one may consider key-words. This leads to the definition of key-word coupling (on the article level):

4.1 Key-word coupling: article level

Consider two articles *a* and *b* and their key-word lists. Then one considers the intersection of these key-words lists, this is the set of key-words common to these two articles. If this intersection is non-empty one says that articles *a* and *b* are key-word coupled. The number of items in this intersection is called the key-word coupling strength. The key-word coupling strength divided by the number of items in the union of the two key-word lists is called the relative key-word coupling strength. Similarly one may study if and how often key-words are mentioned together in the same key-word list.

4.2 Key-word co-mentioning: article level

Consider two key-words *v* and *w* and, for each of them the set of articles for which they are a key-word. Then one considers the intersection of these articles. If this intersection is non-empty key-words *v* and *w* are said to be co-mentioned. The number of elements in this intersection is called the co-mentioning strength. Dividing the co-mentioning strength by the number of articles in the union of the two article sets yields the relative co-mentioning strength.

Note that classical bibliographic coupling and co-citation are properties of articles or of sets of articles (authors' œuvres), while key-word coupling and co-mentioning are only properties of articles in the coupling case but not in the co-mentioning case. Co-mentioning is indeed a property of key-words.

In parallel to our investigations of bibliographic coupling and co-citation we consider now the case of authors, i.e. oeuvres.

4.3 Key-word coupling: author level

Consider two authors A and B and for each of them the set of all key-words of their articles, except for those articles that they have co-authored. Then one considers the intersection of these key-words lists. If this intersection is non-empty one says that authors A and B are author key-word coupled. The number of items in the intersection is called the simple author key-word coupling strength. The author key-word coupling strength divided by the number of items in the union of the two key-word lists may be called the relative simple author key-word coupling strength. If one article written by A and one article written by B (not co-authored by A) are key-word coupled, then A and B are author key-word coupled and conversely, i.e. if A and B are author key-word coupled then there exists at least one article written by A and one article written by B (not co-authored by A) such that these articles are key-word coupled. Taking the sum of all these authors' articles author key-word coupling strengths (and not taking joint-articles into account) leads to the total author key-word coupling strength.

Author level key-word coupling has been introduced by Liu & Zhang [14]. However, they do not eliminate co-authored articles, so that – in our opinion - the resulting maps resemble co-authorship maps. It seems that they used the simple key-word coupling strength. The dual notion, namely co-mentioning can also be defined on the level of oeuvres.

4.4 Key-word co-mentioning: author level

Consider two key-words *v* and *w* and, for each of them the set of authors (oeuvres) for which these words are considered to be a key-word (in at least one article written by that author). Then one considers the intersection of these authors. If this intersection is non-empty key-words *v* and *w* are said to be author co-mentioned. The number of elements in this intersection is called the author co-mentioning strength. Dividing the author co-mentioning strength by the number of authors in the union of the two author sets yields the relative author co-mentioning strength of these key-words. Again one my take the actual number of occurrences into account leading to a total author co-mentioning strength. This notion might be useful to detect typical words co-used by a group of authors (scientific 'schools' ?) or even by one author.

One can even move up to a higher level of aggregation and consider larger groups of authors, such as all authors working at departments (e.g. chemistry departments) at different institutes or universities.

4.5 Key-word coupling: departmental level

Consider two departments D and E and for each of them, the set of all key-words of their articles, except those articles that they have co-authored. This means that articles where department D and department E both appear in the byline are not taken into consideration. Then one considers the intersection of these key-word lists. If this intersection is non-empty one says that departments D and E are key-word coupled on departmental level. The number of items in the intersection is called the departmental key-word coupling strength. The departmental key-word coupling strength divided by the number of items in the union of the two key-word lists may be called the relative departmental key-word coupling strength. If one article written by a member of department D and one article written by a member of department E (not co-authored) are key-word coupled, then D and E are key-word coupled on departmental level and conversely. Taking the sum of the key-word coupling strengths of all articles of these departments (and not taking joint-articles into account) leads to the total departmental key-word coupling strength.

It is clear that the larger the unit (departments or even whole universities) the easier it becomes to be coupled, and the more important the actual coupling strength is. Key-word coupling on institutional level has been introduced by Yang and Jin [15].

4.6 Key-word co-mentioning: departmental level

We finally mention the dual notion of key-word coupling, namely key-word co-mentioning on departmental level. Consider two key-words *v* and *w* and, for each of them the set of departments for which these words are considered to be a key-word (in at least one article having this department in the byline). Then one considers the intersection of these departments. If this intersection is non-empty key-words *v* and *w* are said to be co-mentioned on departmental level. The number of elements in this intersection is called the departmental co-mentioning strength. Dividing the departmental co-mentioning strength by the number of departments in the union of the two sets yields the relative departmental co-mentioning strength of these key-words. Again, as in the previous cases, one my take the actual number of occurrences into account leading to the total departmental co-mentioning strength.

5 Conclusion

The notions of bibliographic coupling and co-citation can be generalized in many directions. Care must be taken to make these definitions as precise as possible, otherwise results using these notions may become irreproducible.

Some authors only consider first authors in co-citation or bibliographic coupling studies. It seems obvious to us that not including all authors may lead to serious distortions. Luckily, more and more colleagues perform all-author studies [16], [17], [18], [19].

There are many other forms of co-occurrences not mentioned here, the most important ones being journal co-citation and co-word occurrences (co-word analysis). These other forms, see e.g. [20] and their generalizations, especially in networks, will be the topic of subsequent research.

Acknowledgment. I would like to thank all colleagues who helped during the preparation of this contribution, especially Raf Guns (UA), Yuxian Liu (Tongji Univ.), Liying Yang (CAS, Beijing) and Dangzhi Zhao (Univ. of Alberta). All remaining errors are of course the responsibility of the author.

References

1. Kessler, M.M.: An experimental study of bibliographic coupling between technical papers. M.I.T., Lincoln Laboratory (1962)
2. Kessler, M.M.: Bibliographic coupling between scientific papers. American Documentation, 14, 10-25 (1963)
3. Egghe, L. & Rousseau, R.: Introduction to Informetrics. Quantitative methods in library, documentation and information science. Amsterdam: Elsevier. ISBN: 0 444 88493 9 (1990)
4. Christensen, F.H. & Ingwersen, P.: Online citation analysis – A methodological approach. Scientometrics, 37(1), 39-62 (1996)
5. Marshakova, I.V.: System of document connections based on references (in Russian). Nauchno-Tekhnicheskaya Informatsiya, ser.2, 6, 3-8 (1973)
6. Small, H.: Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265-269 (1973)
7. Rousseau, R. & Zuccala, A.: A classification of author co-citations: definitions and search strategies. Journal of the American Society for Information Science and Technology, 55(6), 513-529 (2004)
8. Zhao, DZ. & Strotmann, A.: Evolution of research activities and intellectual influences in information science 1996-2005: introducing author bibliographic coupling analysis. Journal of the American Society for Information Science and Technology, 59(13), 2070-2086 (2008)
9. White, H.D. & Griffith, B.C.: Author cocitation: a literature measure of intellectual structure. Journal of the American Society for Information Science, 32, 163-171 (1981)

10. Rosengren, K.E.: The literary system. Unpublished licentiate thesis in sociology, University of Lund, Sweden (1966)
11. Rosengren, K.E.: Sociological aspects of the literary system. Stockholm: Natur och Kultur (1968)
12. Persson, O. : The Literature Climate of Umeå - Mapping Public Library Loans. Bibliometric Notes 4(5) (2000)
http://www.umu.se/inforsk/BibliometricNotes/BN5-2000/BN5-2000.htm.
13. Fano, R.M.: Information theory and the retrieval of recorded information. In: Documentation in Action, J. H. Shera, A. Kent, J.W. Perry (Eds.), 238–244, New York: Reinhold Publ. Co. (1956)
14. Liu, ZH. & Zhang ZQ.: Author keyword coupling analysis: an empirical research. Journal of the China Society for Scientific and Technical Information, 29(2), 268-275 (In Chinese) (2010)
15. Yang, LY. & Jin, BH. : A co-occurrence study of international universities and institutes leading to a new instrument for detecting partners for research collaboration. ISSI Newsletter, 2(3), 7-9 (2006)
16. Eom, S.: All author cocitation analysis and first author cocitation analysis: a comparative empirical analysis. Journal of Informetrics, 2(1), 53-64 (2008)
17. Persson, O.: All author citations versus first author citations. Scientometrics, 50(2), 339-344 (2001)
18. Schneider, J., Larsen, B. & Ingwersen, P. : A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. Scientometrics, 80(1), 103-130 (2009)
19. Zhao, DZ. & Strotmann, A.: Comparing all-author and first-author co-citation analysis of information science. Journal of Informetrics, 2(3), 229-239 (2008)
20. Leydesdorff, L.: The position of Tibor Braun's œuvre: bibliographic journal coupling. In: The multidimensional world of Tibor Braun, W. Glänzel & A. Schubert (Eds.), 37-43, Leuven: ISSI (2007)

Addresses of congratulating author:

Ronald Rousseau

KHBO (Association K.U.Leuven), Industrial Sciences and Technology,
Zeedijk 101, B-8400 Oostende, Belgium

University of Antwerp, IBW
Venusstraat 35, B-2000 Antwerpen, Belgium

K.U.Leuven, Dept. Mathematics,
Celestijnenlaan 200B, B-3000 Leuven (Heverlee), Belgium

Email: ronald.rousseau[at]khbo.be

On Measuring the Publication Productivity and Citation Impact of a Scholar: A Case Study

Tefko Saracevic[1] **& Eugene Garfield**[2]

[1] Rutgers University, New Brunswick, USA
[2] ThomsonReuters Scientific (formerly ISI), Philadelphia, USA

Abstract. The purpose is to provide quantitative evidence of scholarly productivity and impact of Peter Ingwersen, a preeminent information science scholar, and at the same time illustrate and discuss problems and disparities in measuring scholarly contribution in general. Data is derived from searching Dialog, Web of Science, Scopus, and Google Scholar (using Publish or Perish software). In addition, a HistCite profile for Peter Ingwersen publications and citations was generated.

Keywords: Scholarly productivity; citation impact; quantitative measures.

Introduction

The paper is honoring the scholarly contribution of Peter Ingwersen, a scholar extraordinaire in information science. With his ideas, publications, presentations, and collaborations Professor Ingwersen attained a global reach and impact. The purpose here is to provide some numerical evidence of his productivity and impact with a further objective of using this data as a case study to illustrate and discuss the problems, difficulties and disparities in measuring scholarly contributions in general.

The essence of scholarship is proposition of ideas or explanation of phenomena in concert, at some time or another, with their verification. Since antiquity to the present day these were represented in publications – books, treatises, journal articles, proceedings papers etc. – in a variety of forms. Traditionally, their quality was assessed by peer review and recognition, critical examination, and verification of claims. The impact was the breadth and depth of these assessments and even more so their effects on scholarship that followed. Scholarly productivity and impact was a qualitative assessment.

In contrast, close to a century ago quantitative metrics associated with scholarly publications started to appear. Counting various aspects provided a further picture of productivity and impact. At first they were numbers such as publications per author, numbers of references and citations, and other indicators. Bibliometrics emerged in the mid of last century as an area of study of quantitative features and laws of recorded information discourse. Finally, a decade or so thereafter scientometrics focused on the scientific measurement of the work of scientists, especially by way of analyzing their publications and the citations within them – it is application of mathematical and statistical methods to study of scientific literature. Scholarly productivity and impact was also quantified.

Contemporary advances in information and communication technologies enabled innovative creation of large databases incorporating publication and citation data from which, among others, a variety of metrics are derived. Scholarly productivity and impact is being derived quantitatively from massive databases. Results are often used for a variety of evaluative purposes.

Thus, a distinction is made between relational bibliometrics/scientometrics, measuring (among others) productivity and evaluative bibliometrics/scientometrics measuring impact. In this paper we deal with both,

2 Problems, issues

A number of databases now provide capabilities to obtain comprehensive metrics related to publications of individual scholars, disciplines, journals, institutions and even countries. As to statistics related to publications, i.e. relational bibliometrics, they provide straight forward relational data. But as to impact, i.e. evaluative bibliometrics, they also compute a variety of citation-related measures or metrics. In other words, citations are at the base of evaluative bibliometrics. Three issues follow.

The first issue is about the very use of citations for impact studies. Numerous caveats are expressed questioning such use and warning of possible misuse. Leydesdorff [1] is but one of numerous articles addressing the problem. While fully recognizing the caveats and this problem we will not deal with them. Let it be said that such caveats should be applied to data presented here as well.

The second issue is operational and relates to the quality of citations from which evaluative data is derived. Citations are not necessarily "clean" data; ambiguities, mistakes, inaccuracies, inabilities to differentiate, and the like are present at times. Citation hygiene differs. White [2] is but one of numerous articles that discusses possible ambiguities in presentation and use of citation data. Again, while recognizing this issue and problem we will not deal with it here.

The third issue, the one that we will deal with here, is also operational, but relates to coverage and treatment of sources from which publication and impact metrics are derived. *Science Citation Index* appeared in 1963, compiled by the Institute for Scientific Information (ISI), followed a few years later by *Social Science Citation Index* and then by *Arts & Humanities Citation Index*. Using and enlarging on these indexes, in 1997 ISI, (now part of Thomson Reuters) released the *Web of*

Science (WoS) [3]. For four decades, - from 1960s till 2004 – these indexes, including WoS, were the sole source for citation studies and impact data. Thus, for a long while life for deriving and using such data was simple and unambiguous.

In 1972 the Lockheed Missiles and Space Company launched Dialog as a commercial search services, incorporating a number of indexing and abstracting databases for standardized access and searching. [4]. (After several owners, Dialog is now a part of ProQuest). Dialog became by far the largest and most diversified "supermarket" of databases available for searching. Among others, Dialog offered and is still offering ISA citation indexes for citation searches and analyses.

In 2004 Elsevier launched Scopus, a large indexing and abstracting database. At first Scopus covered science, engineering, medicine, and social sciences and later included humanities as well. But from the start, Scopus incorporated citation analyses of various kinds, including impact data. WoS and Scopus provide similar kind of citation analytic capabilities [5]. Suddenly, life was not simple any more. Two different sources for citation analyses became available.

In 2005 Google launched Google Scholar, with the goal to cover scholarly literature. The coverage is broad. As to citations, a "cited by" link is provided but citation analysis can not be done directly. Independently, enters Anne-Wil Harzing, a professor at the University of Melbourne, Australia, and in 2006 releases Publish or Perish (PoP), a free tool or app for deriving various citation analyses, including impact data, from Google Scholar [6]. With three large databases available for citation analyses and impact metrics life got really complicated.

Soon after appearance of Scopus and then Google Scholar a number of papers compared features of these two with WoS (e.g. [7]). But the more interesting question was not comparison of features, but of results. The issue is: How do citation results from these three giant databases compare? For instance, do publication data or impact metrics differ? If so, why and by how much? E.g. If we search for citation and impact data for an author – in this case Peter Ingwersen – are results from the three databases close? Or not?

Not surprisingly, a number of studies were launched trying to answer these questions, i.e. comparing results of citation searches from the three databases. A cottage industry developed addressing the issues and problems. This paper is one of them. Here is but a sample of more recent studies from various fields comparing citation results from WoS, Scopus, and Google Scholar (GS).

Meho and Yang compared ranking of 25 top scholars in library and information science and found that "Scopus significantly alters the relative ranking of those scholars that appear in the middle of the rankings and that GS stands out in its coverage of conference proceedings as well as international, non-English language journals...[and that] WoS, helps reveal a more accurate and comprehensive picture of the scholarly impact of authors."[8].

Kulkarni, et al. compared the citation count profiles of articles published in general medical journals and found that “Web of Science, Scopus, and Google Scholar produced quantitatively and qualitatively different citation counts for articles published in 3 general medical journals.” [9].

Bar-Ilan compared citations to the book “Introduction to Informetrics” from the three databases and found that ”Scopus citations are comparable to Web of Science citations ... each database covered about 90% of the citations located by the other. Google Scholar missed about 30% of the citations covered by Scopus and Web of Science (90 citations), but another 108 citations located by Google Scholar were not covered either by Scopus or by Web of Science.” [10].

Taking it all together: there were differences in results from the three databases, but the magnitude differs from study to study and field to field.

3 Method

Four databases, - Dialog, Web of Science (WoS), Scopus, and Google Scholar (GS) (using Publish or Perish (PoP) software) - were searched for author “Ingwersen P” or “Ingwersen Peter” to identify:

- number of publications,
- number of citations including self-citations,
- number of citations excluding self-citations,
- the h-index,
- papers with highest citation rate, and.
- number of collaborators.

In addition, analysis of Ingwersen publications and citations was done using HistCite, described below.

In **Dialog** the following four files were searched: Social SciSearch (file 7), SciSearch 1990 - (file 34), SciSearch 1974-1989 (file 434), and Arts and Humanities Search (file 439). These files are incorporated in WoS, but their organization and searching in Dialog is very, very different.

WoS was searched using the version available through Rutgers University Libraries – subscription in this version is restricted to WoS data from 1984 to present. Thus, this is a **partial WoS**, but it does contain most Ingwersen publications and citations that appeared in WoS covered journals, since Ingwersen started publishing in 1980.

Scopus was searched in its entirety. Scopus covers journals and other sources that substantially overlap with those in WoS, but also covers some additional ones.

PoP was used to extract data from Google Scholar. GS covers many types and sources of publications but it is not transparent what the coverage is as to sources or time period [7].

HistCite, developed by Eugene Garfield, is a software package that provides a variety of bibliometric analyses and mappings from data in WoS [11]. Input is generated form whole WoS but it also allows input of publications not in WoS (e.g. books, proceeding papers) to search for their citations. Here, the input (collection) for HistCite included: (a) papers by "P Ingwersen" downloaded from whole WoS; (b) papers that contained the cited author "P Ingwersen" also downloaded from WoS; **plus** (c) selected papers not in WoS from an Ingwersen bibliography of 126 publications supplied by Birger Larsen, Royal School of Library and Information Science, Denmark. In other words, papers from that bibliography not in WoS were added to HistCite collection.

All searches were done in the second week of May 2010.

4 Results

This section provides results from searches and analyses in a tabular form. The next section, Discussion, provides interpretation of these results linked to each table. In other words, results are presented all together in one section and discussion again all together in another one. In this way, a reader can look at the results alone and draw own interpretations, and then follow our discussion.

4.1 Publications, citations, h-index

Basic results related to Peter Ingwersen's publications, citations and h-index are presented in Table 1.

Database	No. of publications by P. Ingwersen	Total citations with self-citations	Total citations without self-citations	h-index
Dialog	53	902	859	NA
Scopus	55	1208	1123	14
Web of Science 1984-present	52	1101	663	16
Google Scholar	279	4639	NA	27
HistCite	85	1850	1696	20

Table 1. No. of publications, citations, and h-index for Peter Ingwersen from Dialog, WoS (1984-date), Scopus, Google Scholar (using PoP) and HistCite.

4.2 Time span of publications and citations

Table 2 shows the number of publications per year by Ingwersen from 1984 to 2009. Table 3 shows the number of citations received by Ingwersen's papers per year from years 1984 to 2009. Both are derived by WoS (1984-present).

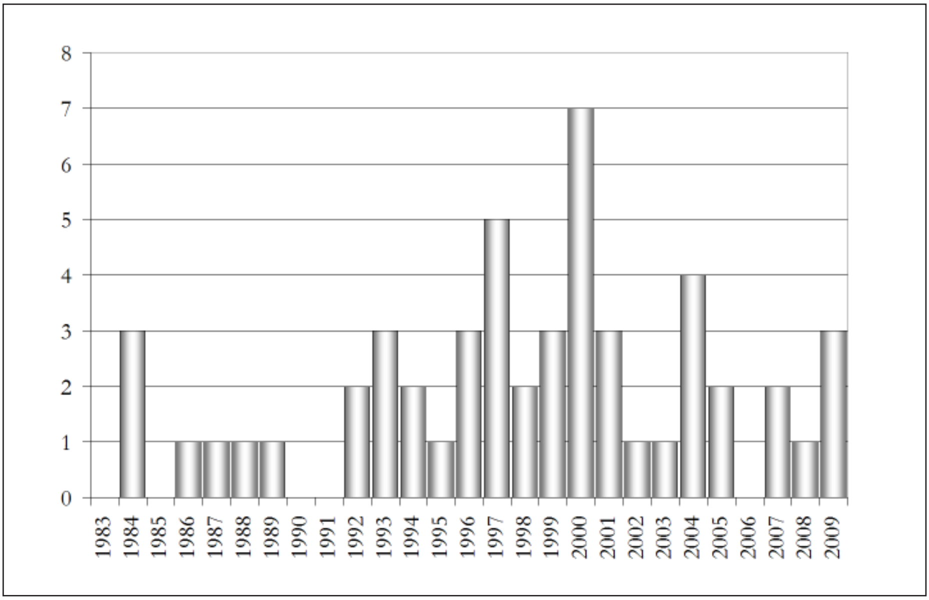


Table 2. WoS (1984-present): No. of publications by Peter Ingwersen published over the years.

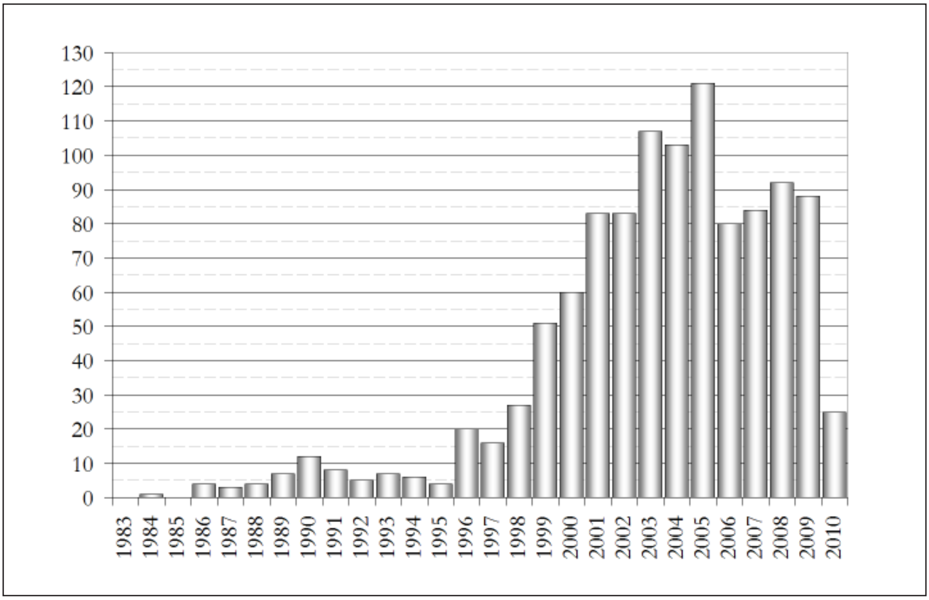
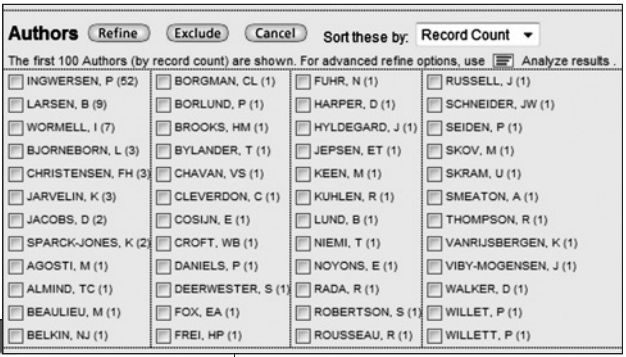


Table 3. WoS (1984-present): No. of citations to Peter Ingwersen's papers over the years.

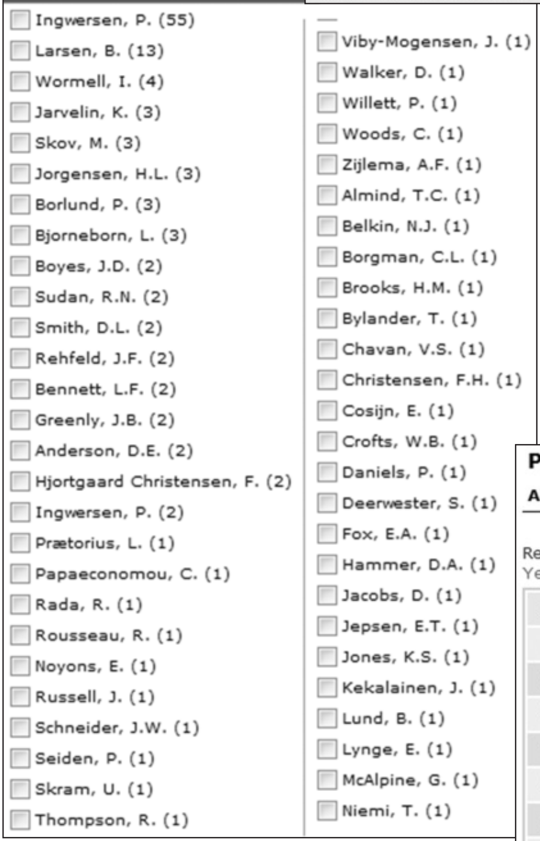
Table 4. WoS (1984-present): List of Ingwersen's co-authors. In this collection Ingwersen has 52 papers with 47 different co-authors (although Willett and Willet are the same author); as example, he co-authored 9 papers with Larsen.



Authors Refine Exclude Cancel Sort these by: Record Count

The first 100 Authors (by record count) are shown. For advanced refine options, use Analyze results .

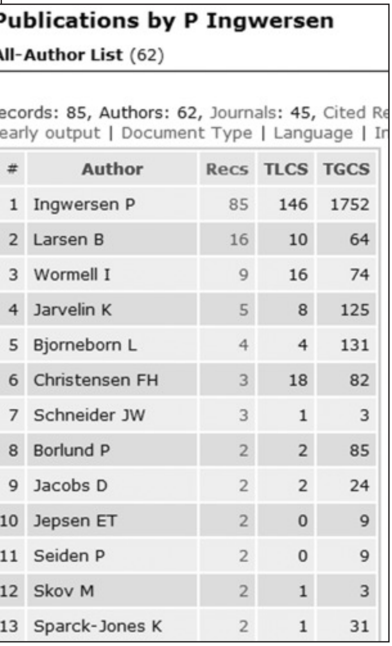
□ INGWERSEN, P (52)
□ LARSEN, B (9)
□ WORMELL, I (7)
□ BJORNEBORN, L (3)
□ CHRISTENSEN, FH (3
□ JARVELIN, K (3)
□ JACOBS, D (2)
□ SPARCK-JONES, K (2
□ AGOSTI, M (1)
□ ALMIND, TC (1)
□ BEAULIEU, M (1)
□ BELKIN, NJ (1)
□ BORGMAN, CL (1)
□ BORLUND, P (1)
□ BROOKS, HM (1)
□ BYLANDER, T (1)
□ CHAVAN, VS (1)
□ CLEVERDON, C (1)
□ COSIJN, E (1)
□ CROFT, WB (1)
□ DANIELS, P (1)
□ DEERWESTER, S (1
□ FOX, EA (1)
□ FREI, HP (1)
□ FUHR, N (1)
□ HARPER, D (1)
□ HYLDEGARD, J (1)
□ JEPSEN, ET (1)
□ KEEN, M (1)
□ KUHLEN, R (1)
□ LUND, B (1)
□ NIEMI, T (1)
□ NOYONS, E (1)
□ RADA, R (1)
□ ROBERTSON, S (1
□ ROUSSEAU, R (1)
□ RUSSELL, J (1)
□ SCHNEIDER, JW (1)
□ SEIDEN, P (1)
□ SKOV, M (1)
□ SKRAM, U (1)
□ SMEATON, A (1)
□ THOMPSON, R (1)
□ VANRIJSBERGEN, K (1)
□ VIBY-MOGENSEN, J (1)
□ WALKER, D (1)
□ WILLET, P (1)
□ WILLETT, P (1)



Author Name

□ Ingwersen, P. (55)
□ Larsen, B. (13)
□ Wormell, I. (4)
□ Jarvelin, K. (3)
□ Skov, M. (3)
□ Jorgensen, H.L. (3)
□ Borlund, P. (3)
□ Bjorneborn, L. (3)
□ Boyes, J.D. (2)
□ Sudan, R.N. (2)
□ Smith, D.L. (2)
□ Rehfeld, J.F. (2)
□ Bennett, L.F. (2)
□ Greenly, J.B. (2)
□ Anderson, D.E. (2)
□ Hjortgaard Christensen, F. (2)
□ Ingwersen, P. (2)
□ Prætorius, L. (1)
□ Papaeconomou, C. (1)
□ Rada, R. (1)
□ Rousseau, R. (1)
□ Noyons, E. (1)
□ Russell, J. (1)
□ Schneider, J.W. (1)
□ Seiden, P. (1)
□ Skram, U. (1)
□ Thompson, R. (1)
□ Viby-Mogensen, J. (1)
□ Walker, D. (1)
□ Willett, P. (1)
□ Woods, C. (1)
□ Zijlema, A.F. (1)
□ Almind, T.C. (1)
□ Belkin, N.J. (1)
□ Borgman, C.L. (1)
□ Brooks, H.M. (1)
□ Bylander, T. (1)
□ Chavan, V.S. (1)
□ Christensen, F.H. (1)
□ Cosijn, E. (1)
□ Crofts, W.B. (1)
□ Daniels, P. (1)
□ Deerwester, S. (1)
□ Fox, E.A. (1)
□ Hammer, D.A. (1)
□ Jacobs, D. (1)
□ Jepsen, E.T. (1)
□ Jones, K.S. (1)
□ Kekalainen, J. (1)
□ Lund, B. (1)
□ Lynge, E. (1)
□ McAlpine, G. (1)
□ Niemi, T. (1)

Table 5. Scopus: List of Ingwersen's co-authors In this collection Ingwersen has 55 papers with 52 different co-authors; he co-authored 13 papers with Larsen.



Publications by P Ingwersen

All-Author List (62)

Records: 85, Authors: 62, Journals: 45, Cited Re
Yearly output | Document Type | Language | Ir

#	Author	Recs	TLCS	TGCS
1	Ingwersen P	85	146	1752
2	Larsen B	16	10	64
3	Wormell I	9	16	74
4	Jarvelin K	5	8	125
5	Bjorneborn L	4	4	131
6	Christensen FH	3	18	82
7	Schneider JW	3	1	3
8	Borlund P	2	2	85
9	Jacobs D	2	2	24
10	Jepsen ET	2	0	9
11	Seiden P	2	0	9
12	Skov M	2	1	3
13	Sparck-Jones K	2	1	31

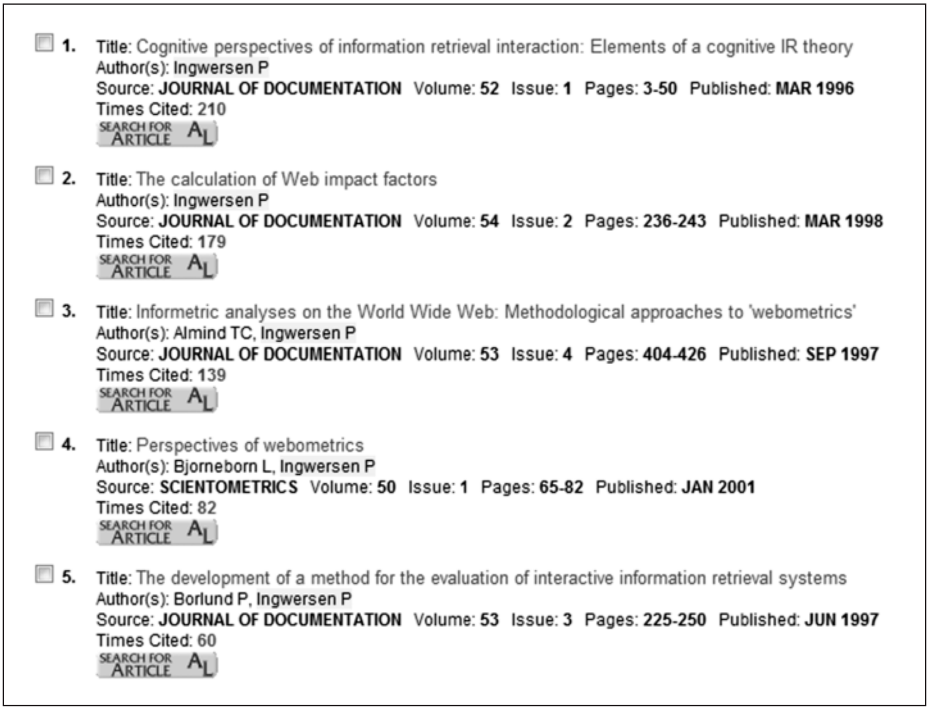
*Table 6. **HistCite:** List of Ingwersen's co-authors up to co-authorship of two papers. In this collection Ingwersen has 85 papers with 62 different co-authors; he co-authored 16 papers with Larsen. Recs = number of records; TLCS = Total Local Citation Score, shows the count of cited papers within the collection; TGCS = Total Global Citation Score, shows the Citation Frequency based on the total count in the Web of Science.*

4.3 Co-authors

In doing research and publishing papers Ingwersen collaborated with a number of scholars. List of Ingwersen's co-authors as listed in WoS (1984-present) are shown in Table 4, as listed in Scopus in Table 5, and as listed in HistCite in Table 6; this table shows co-authors who published 2 or more papers with Ingwersen; single co-authorship list is not shown, because it is too long.

4.4 Highest cited papers

Five highest cited papers by Ingwersen as listed in WoS are shown in Table 7, in Scopus in Table 8, and in HistCite in Table 9.



□ **1.** **Title:** Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory
Author(s): Ingwersen P
Source: **JOURNAL OF DOCUMENTATION** Volume: **52** Issue: **1** Pages: **3-50** Published: **MAR 1996**
Times Cited: 210

□ **2.** **Title:** The calculation of Web impact factors
Author(s): Ingwersen P
Source: **JOURNAL OF DOCUMENTATION** Volume: **54** Issue: **2** Pages: **236-243** Published: **MAR 1998**
Times Cited: 179

□ **3.** **Title:** Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'
Author(s): Almind TC, Ingwersen P
Source: **JOURNAL OF DOCUMENTATION** Volume: **53** Issue: **4** Pages: **404-426** Published: **SEP 1997**
Times Cited: 139

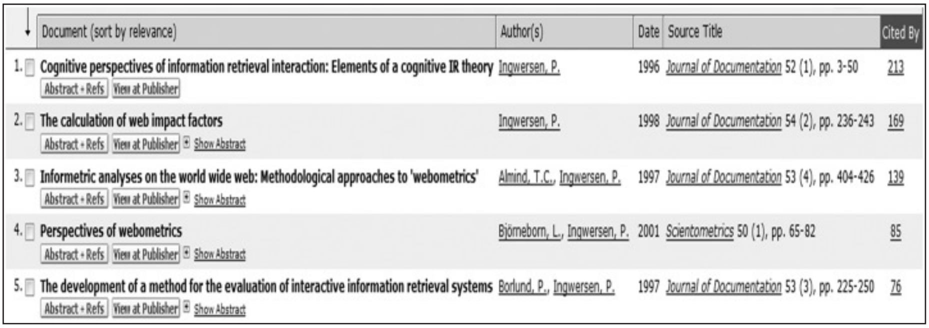
□ **4.** **Title:** Perspectives of webometrics
Author(s): Bjorneborn L, Ingwersen P
Source: **SCIENTOMETRICS** Volume: **50** Issue: **1** Pages: **65-82** Published: **JAN 2001**
Times Cited: 82

□ **5.** **Title:** The development of a method for the evaluation of interactive information retrieval systems
Author(s): Borlund P, Ingwersen P
Source: **JOURNAL OF DOCUMENTATION** Volume: **53** Issue: **3** Pages: **225-250** Published: **JUN 1997**
Times Cited: 60

Table 7. WoS (1984-present): Five highest cited papers by Ingwersen with number of citations for each.

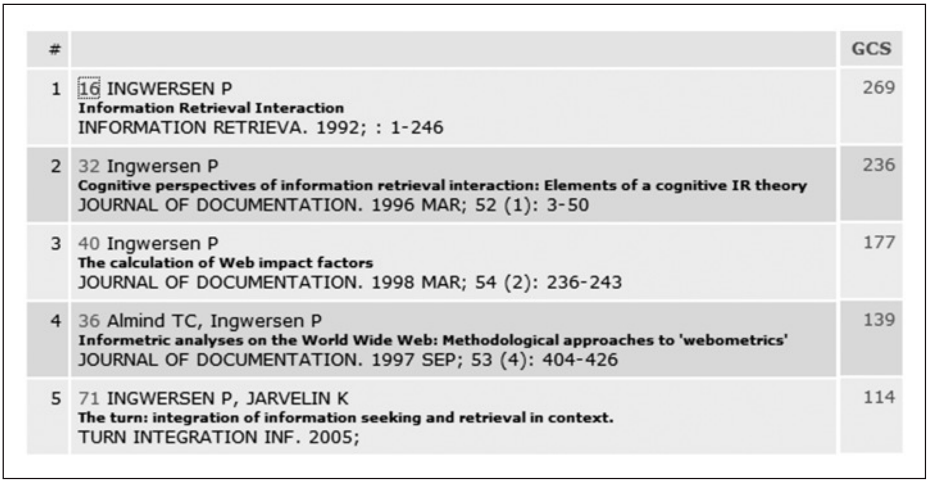
4.5 HistCite

As mentioned, HistCite produces a variety of analyses and mappings using WoS data, but allows input of publications that are not necessarily in WoS, as was the



	Document (sort by relevance)	Author(s)	Date	Source Title	Cited By
1. □	Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory Abstract + Refs View at Publisher	Ingwersen, P.	1996	*Journal of Documentation* 52 (1), pp. 3-50	213
2. □	The calculation of web impact factors Abstract + Refs View at Publisher ⊞ Show Abstract	Ingwersen, P.	1998	*Journal of Documentation* 54 (2), pp. 236-243	169
3. □	Informetric analyses on the world wide web: Methodological approaches to 'webometrics' Abstract + Refs View at Publisher ⊞ Show Abstract	Almind, T.C., Ingwersen, P.	1997	*Journal of Documentation* 53 (4), pp. 404-426	139
4. □	Perspectives of webometrics Abstract + Refs View at Publisher ⊞ Show Abstract	Björneborn, L., Ingwersen, P.	2001	*Scientometrics* 50 (1), pp. 65-82	85
5. □	The development of a method for the evaluation of interactive information retrieval systems Abstract + Refs View at Publisher ⊞ Show Abstract	Borlund, P., Ingwersen, P.	1997	*Journal of Documentation* 53 (3), pp. 225-250	76

Table 8. Scopus: Five highest cited papers by Ingwersen with number of citations for each.



#		GCS
1	16 INGWERSEN P **Information Retrieval Interaction** INFORMATION RETRIEVA. 1992; : 1-246	269
2	32 Ingwersen P **Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory** JOURNAL OF DOCUMENTATION. 1996 MAR; 52 (1): 3-50	236
3	40 Ingwersen P **The calculation of Web impact factors** JOURNAL OF DOCUMENTATION. 1998 MAR; 54 (2): 236-243	177
4	36 Almind TC, Ingwersen P **Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'** JOURNAL OF DOCUMENTATION. 1997 SEP; 53 (4): 404-426	139
5	71 INGWERSEN P, JARVELIN K **The turn: integration of information seeking and retrieval in context.** TURN INTEGRATION INF. 2005;	114

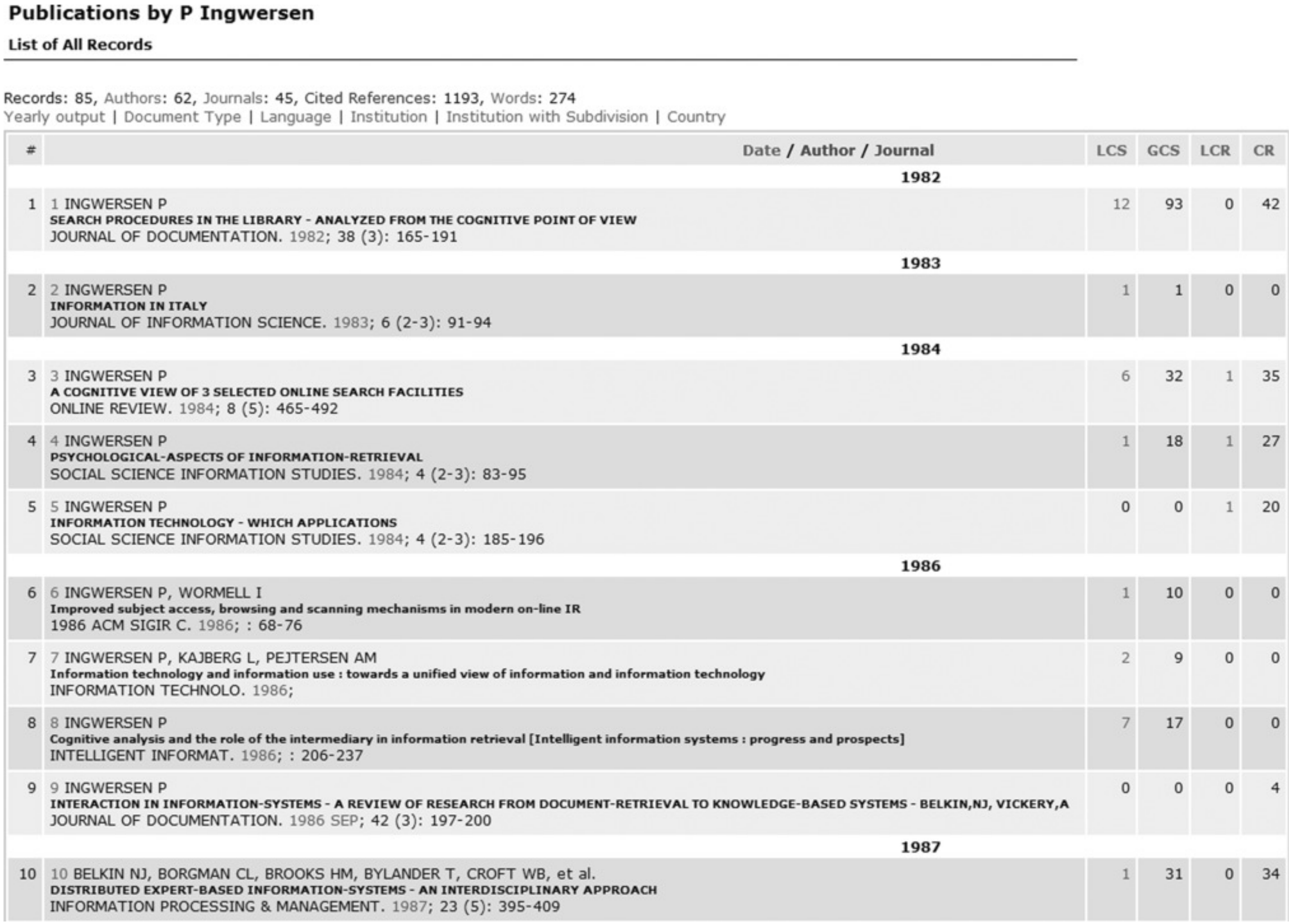
Table 9. HistCite: Five highest cited papers with number of citations for each. GSC= Global Citation Score, shows the Citation Frequency based on the total count in the Web of Science.

case here where selected papers from Larsen's bibliography for Ingwersen that were not in WoS were added.[1] Only a sample of HistCite data is presented here; full array of data can be accessed as follows:

Publications by Peter Ingwersen are available at
http://garfield.library.upenn.edu/histcomp/ingwersen-p_auth/index-tl.html

Papers citing Peter Ingwersen are available at
http://garfield.library.upenn.edu/histcomp/ingwersen-p_citing/index-tl.html

[1] HistCite data presented here is derived from data available online at mentioned sites and is on par with a static report. If one uses the actual HistCite software (available for a free trial at http://www.histcite.com/), the experience is different as more information is available and there are numerous ways to edit and define the collection to ascertain a variety of different statistics. Coupled with the ability to export to Excel, there are many different ways to use data through HistCite software.



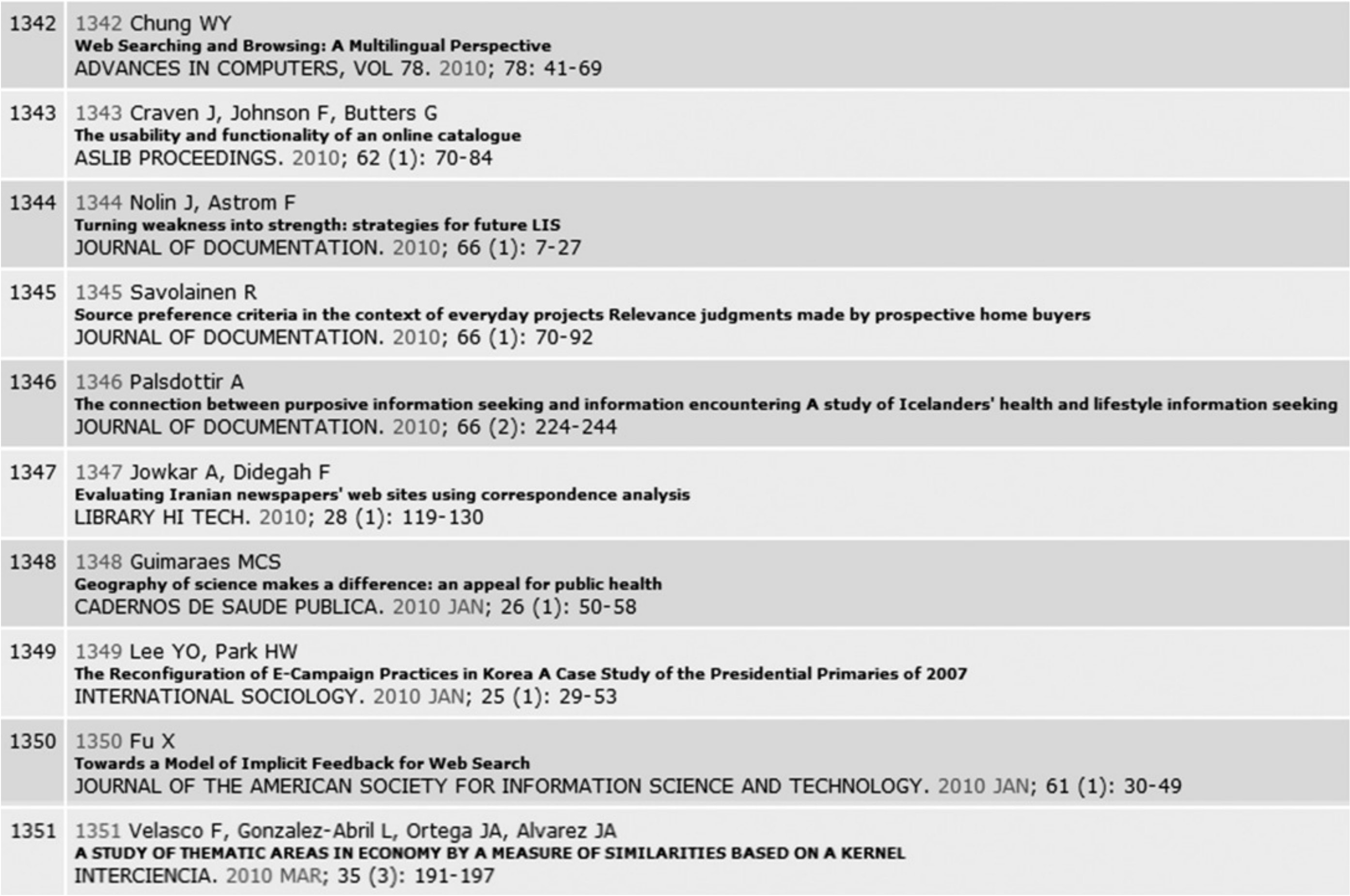
Publications by P Ingwersen

List of All Records

Records: 85, Authors: 62, Journals: 45, Cited References: 1193, Words: 274
Yearly output | Document Type | Language | Institution | Institution with Subdivision | Country

#	Date / Author / Journal	LCS	GCS	LCR	CR
	1982				
1	1 INGWERSEN P **SEARCH PROCEDURES IN THE LIBRARY - ANALYZED FROM THE COGNITIVE POINT OF VIEW** JOURNAL OF DOCUMENTATION. 1982; 38 (3): 165-191	12	93	0	42
	1983				
2	2 INGWERSEN P **INFORMATION IN ITALY** JOURNAL OF INFORMATION SCIENCE. 1983; 6 (2-3): 91-94	1	1	0	0
	1984				
3	3 INGWERSEN P **A COGNITIVE VIEW OF 3 SELECTED ONLINE SEARCH FACILITIES** ONLINE REVIEW. 1984; 8 (5): 465-492	6	32	1	35
4	4 INGWERSEN P **PSYCHOLOGICAL-ASPECTS OF INFORMATION-RETRIEVAL** SOCIAL SCIENCE INFORMATION STUDIES. 1984; 4 (2-3): 83-95	1	18	1	27
5	5 INGWERSEN P **INFORMATION TECHNOLOGY - WHICH APPLICATIONS** SOCIAL SCIENCE INFORMATION STUDIES. 1984; 4 (2-3): 185-196	0	0	1	20
	1986				
6	6 INGWERSEN P, WORMELL I **Improved subject access, browsing and scanning mechanisms in modern on-line IR** 1986 ACM SIGIR C. 1986; : 68-76	1	10	0	0
7	7 INGWERSEN P, KAJBERG L, PEJTERSEN AM **Information technology and information use : towards a unified view of information and information technology** INFORMATION TECHNOLO. 1986;	2	9	0	0
8	8 INGWERSEN P **Cognitive analysis and the role of the intermediary in information retrieval [Intelligent information systems : progress and prospects]** INTELLIGENT INFORMAT. 1986; : 206-237	7	17	0	0
9	9 INGWERSEN P **INTERACTION IN INFORMATION-SYSTEMS - A REVIEW OF RESEARCH FROM DOCUMENT-RETRIEVAL TO KNOWLEDGE-BASED SYSTEMS - BELKIN,NJ, VICKERY,A** JOURNAL OF DOCUMENTATION. 1986 SEP; 42 (3): 197-200	0	0	0	4
	1987				
10	10 BELKIN NJ, BORGMAN CL, BROOKS HM, BYLANDER T, CROFT WB, et al. **DISTRIBUTED EXPERT-BASED INFORMATION-SYSTEMS - AN INTERDISCIPLINARY APPROACH** INFORMATION PROCESSING & MANAGEMENT. 1987; 23 (5): 395-409	1	31	0	34

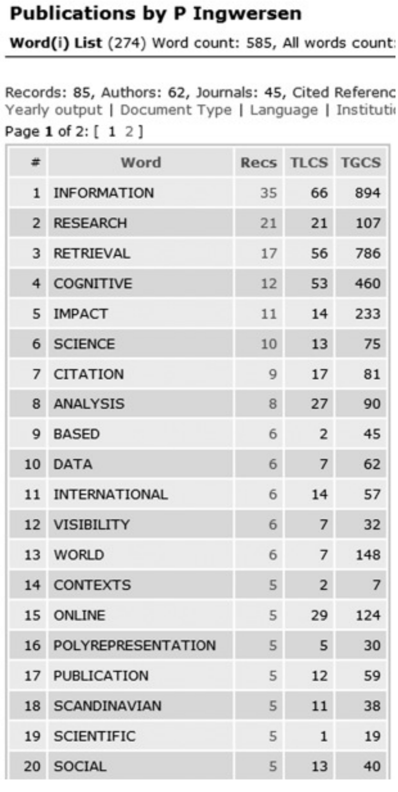
Table 10. HistCite: Sample from 85 publications by Ingwersen; listed are 11 publications from 1982 to 1987. LCS= Local Citation Score, shows the count of cited papers within the collection; GSC= Global Citation Score, shows the Citation Frequency based on the total count in WoS; LCR= Local Citation Score, shows the Citation Frequency within the collection; CR= Cited References, shows the number of all cited references as given in the paper's bibliography.



1342	1342 Chung WY **Web Searching and Browsing: A Multilingual Perspective** ADVANCES IN COMPUTERS, VOL 78. 2010; 78: 41-69
1343	1343 Craven J, Johnson F, Butters G **The usability and functionality of an online catalogue** ASLIB PROCEEDINGS. 2010; 62 (1): 70-84
1344	1344 Nolin J, Astrom F **Turning weakness into strength: strategies for future LIS** JOURNAL OF DOCUMENTATION. 2010; 66 (1): 7-27
1345	1345 Savolainen R **Source preference criteria in the context of everyday projects Relevance judgments made by prospective home buyers** JOURNAL OF DOCUMENTATION. 2010; 66 (1): 70-92
1346	1346 Palsdottir A **The connection between purposive information seeking and information encountering A study of Icelanders' health and lifestyle information seeking** JOURNAL OF DOCUMENTATION. 2010; 66 (2): 224-244
1347	1347 Jowkar A, Didegah F **Evaluating Iranian newspapers' web sites using correspondence analysis** LIBRARY HI TECH. 2010; 28 (1): 119-130
1348	1348 Guimaraes MCS **Geography of science makes a difference: an appeal for public health** CADERNOS DE SAUDE PUBLICA. 2010 JAN; 26 (1): 50-58
1349	1349 Lee YO, Park HW **The Reconfiguration of E-Campaign Practices in Korea A Case Study of the Presidential Primaries of 2007** INTERNATIONAL SOCIOLOGY. 2010 JAN; 25 (1): 29-53
1350	1350 Fu X **Towards a Model of Implicit Feedback for Web Search** JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY. 2010 JAN; 61 (1): 30-49
1351	1351 Velasco F, Gonzalez-Abril L, Ortega JA, Alvarez JA **A STUDY OF THEMATIC AREAS IN ECONOMY BY A MEASURE OF SIMILARITIES BASED ON A KERNEL** INTERCIENCIA. 2010 MAR; 35 (3): 191-197

Table 11. HistCite: Sample of publications citing Ingwersen; listed are 10 (out of 23) publications that were published in 2010.

Here are excerpts from main results. Table 10 shows publications by Ingwersen from 1982 to 1987 – listed are 11 out of a total of 85 publications in HistCite. In addition to date as shown here, data can be sorted by various parameters indicated in blue. Table 11 shows a sample of 10 papers published in 2010 that cite Ingwersen. Table 12 shows 20 most significant words in tiles of papers by Ingwersen. Table 13 shows an example of a historiograph – a map – generated by HistCite; in this case it shows connections of the listed paper in the center of the map. On the above site, connecting papers can be identified by scrolling over them.



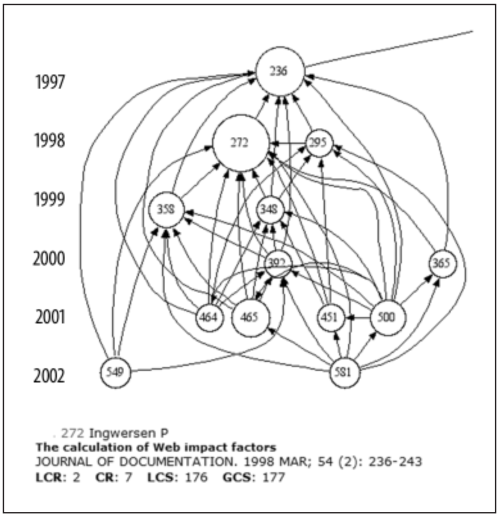
Publications by P Ingwersen

Word(i) List (274) Word count: 585, All words count:

Records: 85, Authors: 62, Journals: 45, Cited Referenc
Yearly output | Document Type | Language | Instituti
Page **1** of 2: [1 2]

#	Word	Recs	TLCS	TGCS
1	INFORMATION	35	66	894
2	RESEARCH	21	21	107
3	RETRIEVAL	17	56	786
4	COGNITIVE	12	53	460
5	IMPACT	11	14	233
6	SCIENCE	10	13	75
7	CITATION	9	17	81
8	ANALYSIS	8	27	90
9	BASED	6	2	45
10	DATA	6	7	62
11	INTERNATIONAL	6	14	57
12	VISIBILITY	6	7	32
13	WORLD	6	7	148
14	CONTEXTS	5	2	7
15	ONLINE	5	29	124
16	POLYREPRESENTATION	5	5	30
17	PUBLICATION	5	12	59
18	SCANDINAVIAN	5	11	38
19	SCIENTIFIC	5	1	19
20	SOCIAL	5	13	40

Table 12. HistCite: Top 20 significant words (out of 274) used in titles of papers by Ingwersen. Recs = shows the number of records where the word appears; TLCS= Total Local Citation Score, shows the count of cited papers within the collection; TGSC= Global Citation Score, shows the Citation Frequency based on the total count in WoS.



. 272 Ingwersen P
The calculation of Web impact factors
JOURNAL OF DOCUMENTATION. 1998 MAR; 54 (2): 236-243
LCR: 2 **CR:** 7 **LCS:** 176 **GCS:** 177

Table 13. HistCite: Map of connections (historiograph) for Ingwersen paper 272 The calculation of Web impact factors to demonstrate mapping. This is a part of a larger map. In the original hisoriograph connecting papers are identified by scrolling over.

5. Discussion

Publications, citations, and h-index (Table 1): Dialog, Scopus, and WoS contained a similar number of papers **by** Peter Ingwersen but produced differing number of citations **to** Ingwersen. (Note that the version of WoS used here is from 1982-present and not the whole WoS). It is not clear how Dialog, supposedly containing the same databases as WoS, produced a lesser number of citations. On the other hand, WoS produced the smallest number of citations without self-citations. Possibly, computing algorithms and rules may differ. The h-indexes were almost identical.

Google Scholar produced by far the largest number of publications by and citations to Ingwersen. This is partially a reflection of a much broader coverage than other databases, particularly in proceedings and non-English publications, and partially because of a lack of quality control e.g. counted are multiple versions of the same paper, ghost citations and the like as enumerated by Jacsó [12]. In general, data from Google Scholar are inflated.

As mentioned, HistCite included papers by Ingwersen from whole WoS plus those not in WoS added from Larsen's bibliography of Ingwersen. Thus, the base collection for searching for citations was larger. This produced probably the most realistic numbers of citations and h-index – simply because more of Ingwersen's publications were used. He has written quite a bit more than what is covered by WoS or Scopus but not as much as indicated by Google Scholar.

Time span of publications and citations (tables 2 and 3). Data here are generated from WoS only, even though a similar display can be obtained from Scopus as well. His highest productivity in publishing papers was the time span 1997-2000. His highest number of citations was in publications that were published in the time span of 2001 to 2009. His impact, as measured by the number of citations, is continuing to this day. In other words, the impact of his publications goes on unabated.

Co-authors (tables 4, 5, and 6). Here we can see significant differences among databases. WoS includes 52 Ingwersen papers with 47 different co-authors. Scopus has 55 papers with 52 co-authors. HistCite has 85 papers with 62 co-authors. Larsen is the highest placed co-author in all three databases, but WoS shows that Larsen co-authored 9 papers with Ingwersen, Scopus 13, and HisCite 16. This may be due to evident difference in coverage, but it could be also that method of processing – policies and/or algorithms may differ.

Highest cited papers (tables 7, 8, and 9). Again, here we can see significant differences. The order of five highest cited papers for WoS and Scopus are the same, but not for HistCite; however, the number of citations that these papers receive differs from one database to the other. And again this may be due to differences in coverage, algorithms, and policies, but with citations this may also be due to type and intensity of quality control. As mentioned, citation data are not “clean” at all, thus the question is: how effectively are theyt cleaned?

HistCite (tables 10, 11, 12, and 13). Full and rich data from HistCite are provided at listed URLs, thus all the tables presented here offer only a set of static and very limited examples of what is available there. In dynamic HistCite, data can be sorter in numerous ways – date, authors, journals, papers, citations, references, yearly output, and more. This provides for a dynamic interaction with data and even discovery. One such example is analysis of significant words used in titles of papers by Ingwersen. The top 20 words almost describe his oeuvre of interest and research. The maps show connections that can be further explored.

6. Conclusions

The purpose of the paper was to provide quantitative evidence of scholarly productivity and impact of Peter Ingwersen; at the same time another objective was to illustrate and discuss problems and disparities in measuring scholarly contribution in general.

As to the main purpose: Data confirm a long and sustained effort by Ingwersen over the span of some three decades. Moreover, they also confirm several other aspects: the large impact he had on other work and research, the amazing breadth of his collaborations, and the international connections he nurtured.

But as to the second objective of the paper, data also confirm considerable difference in results from one database to another. This was observed in a number of other studies, thus, here is another confirmation. Reasons for such disparity were not investigated here – they could be due primarily to differences in coverage, but also in policies, algorithms, and quality control in handling of data. A number of larger questions related to use or even misuse of such data can also be raised [e.g. 13], but are not discussed here.

Still, with all the caveats and problems we are better off with such data then without.

Acknowledgment. Assistance of Soren Paris from HistCite LLC in creating of HistCite collections for Peter Ingwersen is gratefully recognized.

References

Leydesdorff, L. Caveats For The Use Of Citation Indicators In Research And Journal Evaluations. J. Am. Soc. for Inf. Sc. and Techn., 59 (2), 278–287. (2008)

White, H. D. Citation Analysis. In: M. J. Bates and M. N. Maack (Eds.) Encycl. of Libr. and Inf. Sci. New York: Taylor & Francis. Third Edition, 1: 1, 1012–1026. (2010)

McVeigh, M. E. Citation Indexes and the Web of Science. In: M. J. Bates and M. N. Maack (Eds.) Encycl. of Libr. and Inf. Sci. New York: Taylor & Francis. Third Edition, 1: 1, 1027 — 1037 (2010)

Bellardo-Hahn, T. Pioneers Of The Online Age. Inf. Proc. & Mngmt. 32 (1), 33-48. (1996)

Dess, H.M. Data Base Reviews and Reports: Scopus. Issues in Science and Technology Librarianship. issue 45. (2006) http://www.istl.org/06-winter/databases4.html

Harzing, A.W. Publish or Perish. http://www.harzing.com/pop.htm

Jacsó, P. As we may search – Comparison of major features of the *Web of Science*, *Scopus*, and *Google Scholar* citation-based and citation-enhanced databases. Current Science, 89 (9), 1537-1547. (2005) http://www.ias.ac.in/currsci/nov102005/1537.pdf

Meho, L.I., Yang, K. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. J. Am. Soc. for Inf. Sc. and Techn., 58(13), 2105-2125. (2007)

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals. JAMA – J. Am. Med. Assoc. *302*(10), 1092-1096. (2009)

Bar-Ilan, J. Citations to the "Introduction to Informetrics" Indexed by WOS, Scopus and Google Scholar. Scientometrics, , 1-12. (2010)

HistCite. Bibliometric Analysis and Visualization Software. http://www.histcite.com/index.htm

Jacsó, P. Google Scholar Ghost Authors, Lost Authors and Other Problems. Libr. J. 134(18), 26-27. (2009) http://www.libraryjournal.com/article/CA6698580.html

Garfield, E. The History and Meaning of the Journal Impact Factor. JAMA – J. Am. Med. Assoc. (293): 90-93, (2006) http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

Addresses of congratulating authors:

Tefko Saracevic
School of Communication and Information, Rutgers University
4 Huntigton Street, New Brunswick, NJ 08901, USA
Email: tefkos[at]rutgers.edu

Eugene Garfield
Chairman Emeritus
ThomsonReuters Scientific (formerly ISI)
1500 Spring Garden Street, Philadelphia, PA 19130, USA
Email: Garfield[at]codex.cis.upenn.edu

Cognitive Perspectives of Peter Ingwersen

Henry Small

Thomson Reuters, Philadelphia, USA

Over the years I have been impressed by Peter's ability to move comfortably and effortlessly between the worlds of information retrieval and bibliometrics, giving a tutorial at SIGIR and then presenting a webometrics paper at an ISSI conference (This is not to mention his formidable organizational skills.) However, on rereading some of his primordial and highly cited works [1], [2], I wonder now whether this seeming split personality was not the product of a larger, overarching world view.

In his well known papers on the cognitive perspective of information retrieval, he discusses the complexity of retrieval as an interaction of multiple domains and representations. He speaks of the user's diverse "models of the world" and "cognitive spaces", the separate and differing worlds of retrieval systems, indexers, document representations, search intermediaries. The interaction of these multiple systems results in substantial uncertainty and unpredictability in information retrieval outcomes, as well as possibilities for enhancing outputs.

The transition from retrieval to bibliometrics moves the focus from consumer of information to the producer, but, in principle, the scholarly or scientific author's cognitive space is similar if not identical to that of the information seeker. Peter has described the information user's cognitive state in terms of his/her work task, goal, problem, information behavior, degree of specificity or vagueness, social context, and even emotional condition. Similarly scientists may on occasion search for new perspectives on a problem that will enable them to make advances in their fields, and arrive at new insights and discoveries, and gain credit and priority. The information seeking process for them can be exploratory and uncertain, because, by definition, they are not sure what they are looking for. Emotions also enter the picture due to disagreements with colleagues and competition for recognition.

The parallel between bibliometrics and information retrieval is clearly seen in the case of the citing author viewed as an information user. A bibliographic reference represents a concrete outcome of a retrieval process, and a crystallized relevance judgment. The citing passage, as Peter has discussed, gives meaning and context to the cited text. An example from my recent research may serve as a roundabout confirmation of his finding of uncertainty in the retrieval process.

We can think of the citing scientist as living in two worlds: the world of his/her own research specialty and the wider world of the scientific community as a whole. Most of the relevant information that the scientist needs resides in the local community where the players and terminology are familiar, and retrieval outcomes are somewhat predictable, consistent with Peter's "internal" classification of information needs. But when the scientist ventures outside this narrow world to find new ideas, inspiration, and perhaps forging new interdisciplinary links, the relevance of information is much less certain (Peter's external or "berry-picking" mode).

We can study this information behavior by comparing co-citation contexts drawn from within specialties to those that cross disciplinary boundaries. For both samples, the citing passages are collected and each defines a word corpus which can be analyzed by the methods of corpus linguistics (Wordsmith Tools). Surprisingly the key terms for the cross disciplinary sample compared to the within specialty sample, as identified by the log likelihood statistic, reflect a state of contingency and uncertainty on the part of the citing author toward the cited references. The prominence of subjunctives such as "may" and "could" and adjectives such as "possible", "potential" and "promising" suggest that when authors cite the more distant literatures as opposed to their own fields, the projected outcomes of research are hopeful but uncertain. An earlier study also found a high incidence of thinking by analogy for such interdisciplinary connections [3]. By contrast the most prominent key terms from the within specialty contexts reflect the utility of the cited references, signaled by terms such as "with", "employ", "applied", and "utilize". At the least I think we can safely conclude that interdisciplinarity is one important source of uncertainty in both retrieval outcomes and referencing.

Finally, the fields of scientometrics and bibliometrics would benefit from following Peter's cognitive perspective more thoroughly. There is a tendency in these fields to pursue evaluative aims without regard to the content and context of the object being evaluated. A better approach would be to detail the cognitive states of both the evaluated and the evaluator, much as Peter has provided for the seeker of information and the information system. This cognitive approach to bibliometrics should also include addressing the uncertainties in the evaluation process.

References

1. Ingwersen, P.: Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. Journal of Documentation, 52(1), 3-50 (1996).
2. Ingwersen, P.: Information Retrieval Interaction. London: Taylor Graham (1992).
3. Small, H.: Maps of science as interdisciplinary discourse: co-citation contexts and the role of analogy. Scientometrics, 83(3), 835-849 (2010).

Address of congratulating author:

Henry Small
Thomson Reuters
1500 Spring Garden, Philadelphia, PA., 19130, USA
Email: henry.small[at]thomsonreuters.com

Blog Issue Analysis: An Exploratory Study of Issue-Related Blogging

Mike Thelwall & David Wilkinson

University of Wolverhampton, Wolverhampton, United Kingdom

Abstract: The blogosphere contains a huge amount of discussion and public opinion about a wide variety of topics. This article introduces an information science approach to analyse topic-specific discussions on a large scale within the blogosphere from the perspective of bloggers. First, relevant blogs are identified through search engine queries. Motivated by the research of Peter Ingwersen and his PhD student Lennart Björneborn, the blogs are then analysed in terms of the distribution of site sizes, inlinks, outlinks and co-inlinks with respect to a reference data set of random blogs. In addition, content analyses are conducted on a sample of apparently issue-relevant blogs. The approach is illustrated through six case studies of religious issues. The results show that although the more numerical techniques seem to have little value blog issues can have significantly have different web signatures and that some of the methods described here are capable of yielding interesting information.

1. Introduction

Blogs, web sites containing a series of entries posted in reverse chronological order, have been widely and rapidly adopted since free blogging systems like OpenDiary emerged in 1998. The most prominent blogs seem to be those of political, news and technology commentators but there are many different kinds, with personal blogs dominating numerically (Herring, Scheidt, Bonus, & Wright, 2004). In addition to blogs in free blog sites like blogger.com and blogsky.com, some social network sites (boyd & Ellison, 2007) like MySpace, Cyworld and Live Spaces contain blogs as part of a suite of facilities offered to members. Particularly because of social network sites it is difficult to estimate the number of active bloggers in the world but it seems likely to be well over 100 million (e.g., Technorati claimed to track 112.8 million blogs by February 22, 2008: http://technorati.com/about/ and BlogPulse claimed to have found 126.9 million in May 8, 2010 http://www.blogpulse.com/). In the US, blogging is widespread. In 2009 it had declined amongst youth by about 50% over three years, however, to 14-15% of teens and young adults (Lenhart, Purcell, Smith, & Zickuhr, 2010). Nevertheless, 72-73% of online US teens and young adults having a social network site, possibly including a type of blog or microblog, and blogging is increasing for older US adults, to about 11% in 2009 (Lenhart et al., 2010). Moreover, a much larger segment of the population reads blogs than writes them. Given the large number of bloggers and much discussion about their informational (Bar-Ilan, 2005; Bond & Abtahi, 2005; Thelwall, 2007) and democratic (Coleman, 2005; Elmer et al., 2007) potential, it seems important to both exploit blogs to study issues of interest and to analyse blogspace to evaluate its influence over topics for which it is considered important.

In the commercial sector, blogs and similar "consumer-generated media" are already being exploited to give information on public reactions to brands and advertising campaigns (Pikas, 2005) by companies including Microsoft (Gamon, Aue, Corston-Oliver, & Ringger, 2005), IBM (Gruhl, Chavet, Gibson, Meyer, & Pattanayak, 2004) and Nielsen (buzzmetrics.com). One approach, for example, is to use automatic sentiment analysis techniques (Kanayama & Nasukawa, 2006) to estimate the daily number of positive and negative comments about a brand. Blogs are an attractive data source for these applications not only because of the large number of bloggers but also because blogs posts are time-stamped so that time series analyses can be conducted, even retrospectively (Thelwall, 2007). In terms of techniques to systematically quantitatively analyse blogs to generate results of interest to social scientists, there are a few (e.g., Schmidt, 2007), but none are general purpose for analysing issues. For example, one detects sentiment in blogs to identify the mood of the nation (Dodds & Danforth, in press) and another seeks to estimate the proportion of texts matching a certain category of interest (Hopkins & King, 2010).

In this article the focus is on conducting social science descriptive analyses of bloggers alluding to a specific issue. These analyses are based upon the bloggers mentioning the issue and the hyperlinks to and from their blogs. The reason for using hyperlinks as the raw data, an approach popularised by Ingwersen's (1998) highly cited and pioneering Web Impact Factors paper, is that links can reveal important blogs and related sites (see below). The most similar previous work is an analysis of topic-related connections between different online services based upon an analysis of hyperlinks in blogs (Bhagat, Cormode, Muthukrishnan, Rozenbaum, & Xue, 2007). The analysis, although not focussing on a given issue, incorporated a variety of issue-relevant statistics and useful techniques, many of which are similar to those presented below. A range of blog hyperlink analyses have also been previously published, often with a political focus, (Ackland, 2005; Adamic & Glance, 2005; Park & Thelwall, 2008), but these have tended to use approaches specific to the topic studied – such as starting with politicians' blogs - rather than a generic approach.

This article introduces a new generic approach, *blog issue analysis (BIA)*, for analysing bloggers alluding to an issue within blogspace. It comprises five components:

- Identifying issue-relevant blogs
- Size distribution analysis of the blogs
- Hyperlink analyses of links to and from the blogs
- Content analysis of the text in the blogs
- Benchmarking against a different collection of blogs to identify anomalies

Like a similar previous technique, *web issue analysis* (Thelwall, Vann, & Fairclough, 2006), it includes investigations into the international spread of an issue but, in contrast, it is focussed on bloggers, avoids the use of natural language processing to analyse the contents of pages, and includes a wider range of link analyses. In this article blog issue analysis is described and investigated through a series of cases. These cases, each relating to a well-known religion, were chosen to cover issues with varying degrees of centrality to the lives of the bloggers likely to mention them, and hence should give significantly different results and highlight the value of the methods in differentiating between issue structures.

2. Blog Issue Analysis 1: Identifying issue-relevant bloggers

The first BIA stage is to gather a large collection of blogs mentioning the chosen issue. This can be indirectly achieved by taking advantage of the domain naming feature of many blog hosting sites, which allocate subdomains to members' blogs. For example blogs in Live Spaces (spaces.live.com) have domain names ending in spaces.live.com, such as searchtextmining.spaces.live.com and qianjiayi1122.spaces.live.com. This structural feature means that searching the domain spaces.live.com through the `site:` keyword, is an effective way to identify bloggers, as illustrated below. The BIA method is to build an issue-specific query and then submit it to search engine site-specific searches for a range of blog hosting sites that include blogs with unique domain names for members. For example, for the issue Integrated Water Resource Management, the search

```
"Integrated Water Resources Management" OR IWRM OR "Integrated
Water Resource Management" site:X
```

could be submitted with X replaced by the domain name of a blog hosting site such as spaces.live.com or blogsky.com. Each search would yield up to 1,000 URLs of pages in blogs matching the query (more if using query splitting, Thelwall, 2008). Extracting the unique domain names of these results gives a list of potentially issue-relevant blogs. The process of submitting the queries and recording the matching URLs can be automated using search engine APIs (Mayr & Tosques, 2005) and the domain extraction and duplicate elimination can be automated using the free software *LexiURL Searcher* (lexiurl.wlv.ac.uk, Thelwall, 2009). The critical part of this initial step, however, is the construction of an effective query. In some cases terms may need to be explicitly excluded from the results in order to avoid false matches. For example, if searching for the animal *jaguar* then any query may need to subtract *car* and *auto* in order to eliminate as many jaguar car references as possible (e.g., `jaguar -car -auto site:blogger.com`).

Note that the method potentially captures any *blog* that has ever mentioned the issue, rather than just blogs that focus on the issue or *blog entries* that discuss the issue. This is a result of an intentional focus on blogs rather than discussion. Nevertheless, it is an important fact to remember when analysing the results.

The heart of BIA is gathering various types of data on the blogs identified, or on a random sample of identified blogs if too many match the searches. The following types of information are proposed as a core set. These are core in the sense that they are all relatively easy to calculate using free online software and provide generic descriptive information about an issue. An analysis may want to add additional issue- or goal-specific methods, however.

3. Blog Issue Analysis 2: Blog size distribution

Estimates of the number of pages in any blog can be obtained from Google, Yahoo! or Bing through a simple query `site:D` where *D* is the domain name of the blog. Estimates of the sizes of blogs obtained in this way can be inaccurate because they are based upon the pages found by the search engines, which may have missed some. In addition, blog web sites are typically database-driven, with automatically generated pages that archive old posts and so the total number of pages in a blog web site is only a crude indicator of the total number of posts in the site. Moreover search engine estimates in some cases can be inconsistent – for example large estimates may reflect the number of pages found whereas small estimates may reflect the number of pages found after filtering for duplicates and near-duplicates (Thelwall, 2008).

The size distribution of blogs, plotted on a log-log scale, can be expected to obey a power law linear shape, as much web data does (Barabási, 2002; Rousseau, 1997). However, the useful information that may be identified is the identity of the largest blogs and any deviation from a pure power law. To illustrate the latter, a larger number of small blogs than expected might indicate a high level of interest from casual bloggers. The identity of the largest blogs is potentially relevant as they may be particularly important for the issue.

4. Blog Issue Analysis 3: Link analysis

Hyperlinks are an important and widely studied phenomenon in webometrics because a link to a web site may indicate either a connection between the owners of the source and target web site or an endorsement of the target web site by the owner of the source web site (Björneborn & Ingwersen, 2001, 2004; Brin & Page, 1998; Kleinberg, 1999). Hyperlinks can be analysed in various different ways in order to get useful insights into a topic (Björneborn, 2006).

Blog inlink count distribution. Estimates of the number of inlinks to each blog and lists of the URLs of these inlinks can be obtained from Yahoo! via site inlink searches (Ingwersen, 1998). If D is the domain name of a blog then the query `linkdomain:D -site:D` matches pages outside of the blog that contain a link to any page within the blog. It is important to exclude self-links, which are pages within the blog that link within the same blog, because these do not reflect external interest in the blog and are hence not relevant to a blogosphere analysis. The value of inlink counts is that each inlink reflects some kind of interest in the target blog, so that better linked-to blogs are likely to be the most important. Although the distribution of inlink counts is likely to be a power law, as in the case of the site size distribution, any deviation from a power law may indicate useful aspects of the distribution of authority. For example, there may be relatively few key blogs or one blog that is overwhelmingly important. Note, however, that the link count estimates of Yahoo! are subject to the same search engine reliability provisos discussed above.

Blog inlink URL analysis. In addition to the inlink count distribution and the names of the most linked-to blogs, information can be extracted from the URLs of the pages linking to the blogs. First, the names of the web sites and/or URLs most frequently linking to the blogs can be extracted by combining the lists of URLs matching all the site inlink searches submitted to Yahoo! This data reaches out from the blogosphere to the wider web by identifying pages and sites that frequently link to relevant blogs. In addition, some indication of the international distribution of the issue may be gained by counting the top-level domains (TLDs) of the inlinking web sites. This may show, for example if the issue seems particularly relevant to the U.K., the U.S. or another country. This TLD analysis is rather weak, however, because many web sites are published in generic TLDs which sometimes cannot be attributed to countries, such as .com, .net and .org. Moreover, if the issue blog search used in the first stage was language-specific then the TLDs are likely to reflect countries speaking that language. Finally, when counting domains and TLDs, multiple URLs from the same web site linking to the same blog should be excluded, this is because links are sometimes automatically replicated to different pages of a site, as happens with blogroll links. The elimination of multiple links from the same site can be achieved by converting the lists of URLs linking to each blog (from Yahoo!) into lists of domain names linking to each blog (by truncating the URLs) and eliminating duplicates. This processing can be done automatically using LexiURL Searcher (by generating a standard LexiURL Searcher report from the search engine results and using the domain or site column in the summary table produced).

Blog outlinks. Blog outlinks, i.e. the URLs of pages outside the blog that are linked to by the blog, can be obtained from Bing using its linkfromdomain: command. An alternative is to use a web crawler to download all of the pages in each blogs and then to extract the links. This is feasible but consumes a lot of resources (bandwidth and computer time) for the value of the information extracted. A practical alternative, however, is to download only the home pages of each blog and to extract its links. This can be achieved by feeding a list of all blog URLs into *SocSciBot* (socscibot.wlv.ac.uk). The restriction to the home pages alone means that the links extracted may not be relevant to the topic, however, and are likely to be dominated by the blogroll links, which may express the main interests of the blogger. The analyses that can be performed on these outlinks parallel those for inlinks: the most commonly targeted domains and TLDs. The distribution of the number of outlinks per blog home page does not seem relevant, however, but the processes would be similar to those for blog inlink analysis.

5. Blog Issue Analysis 4: Content analysis

When conducting a blog issue analysis it is important to conduct content analyses of sites and links in order to understand the context in which they exist. To give an extreme example, most links to blogs might indicate support for the ideas expressed; conversely they might predominantly indicate disagreement or could even indicate nothing of value if they are paid advertising.

The content analysis should be based upon a random sample of links (i.e., pages linking to the blogs, as found by the Yahoo! blog inlink searches), and a set of 100 should be sufficient to give an approximate indication of what the links represent. Larger samples should be used if the purpose is to compare different blog issues because the larger numbers will allow more fine-grained comparisons. The links should be coded into categories that are relevant to the analysis objective. As part of this, the categories for links should probably include one for irrelevant links, including spam and paid-for links. Once the content analyses are complete, the quantitative results can be interpreted more clearly in terms of the issue.

A content analysis should also be conducted of at least 100 of the blog pages mentioning the issue (i.e., the pages matching the issue-specific searches). This will allow the bloggers to be connected to the issue in some way, by identifying the range of reasons why the bloggers alluded to the issue. This also serves as a check that the bloggers have genuinely mentioned the issue.

6. Blog Issue Analysis 5: Benchmarking

Most of the statistics above can be better interpreted through comparisons with other, similar statistics. For example TLD distributions are dependant upon the spread of TLD usage, and so it is difficult to tell whether any particular TLD count for an issue indicates unusually high or low interest from the associated country. For instance, if 10% of links come from the Spanish .es domain, is does this reflect an unusually high or low interest in the topic from Spain? Similarly, some web sites, like Google, are widely linked to and so it would not be significant to see them in any list of top linked-to sites. Comparisons can help to remove such popular sites from consideration when analysing the key sites for an issue.

The problem of a need for comparison may be avoided if two blog issue analyses are conducted in parallel and their results compared. If this is not relevant to a particular research objective then an alternative is to use a neutral analysis to serve as a comparator. This neutral analysis can be conducted on a random list of blogs, for example obtained through keyword searches, as described for the first issue analysis step, but using neutral e.g. content-free, but language-specific, keywords like 'the', 'because' or 'when'.

7. Case Studies

To illustrate BIA, several cases are presented here. Although it is not normally relevant to present several issues together, this is done here for compactness and to illustrate differences between BIA results. A set of religious terms was chosen as likely to exhibit differing degrees of concentration within blogs – the degree to which individual bloggers are likely to focus solely on the issue associated with the term, if they mention it.

- *Wicca* – the name for a nature-based religion often thought of as witchcraft or paganism. This term seems likely to be mainly used by practitioners and, if used in a blog, seems likely to indicate a blogger for which Wicca is important. This probably represents a strongly concentrated blogging issue.
- *ISKCON* – the acronym of the International Society for Krishna Consciousness. This is assumed to be little-known and mainly used by devotees. If used in a blog, it seems likely to be used by an ISKON sympathiser and for ISKCON to be an important part of the blog. This probably represents a strongly concentrated blogging issue.
- "*Hari Krishna*" – the colloquial name for the ISKCON movement. This is assumed to be known much more widely than ISKCON and for bloggers mentioning Hari Krishna to be much less likely to focus on it throughout their blog. This probably represents a moderately concentrated blogging issue.

- *Scientology* – associated with a U.S.-based Church of Scientology religious movement. This term is widely known and so this probably represents a moderately concentrated blogging issue.
- *Islam* – a major world religion. This seems likely to be a disparate blogging topic, with separate foci: religious and political (both for and against). It is also a well-known and so it probably represents weakly concentrated blogging issue with islands of stronger concentration.
- *Christ* – a figure associated with Christianity, another major world religion. The term is also used for cursing. As a result of the common use of the term, this probably represents a weakly concentrated blogging issue.
- *When, because* – common non-nouns. These both probably represent non-concentrated blogging "issues" and are included as a benchmark or control group.

In the remainder of this paper the case studies are described in parallel, stage by stage, drawing lessons where appropriate.

7.1 BIA 1: Identifying issue-relevant bloggers

A range of predominantly English blog hosting web sites was chosen as the core set to be investigated. These were identified by an ad-hoc collection of web searching methods, giving the following list: blogwebsites.com, spaces.live.com, xanga.com, typepad.com, inknoise.com, weblogs.us, blog-city.com, blogsome.com, pitas.com, blogspirit.com, squarespace.com, weblogger.com, journalspace.com, blog.com, deadjournal.com, livejournal.com, motime.com, blogharbor.com, blogdrive.com, fotopages.com, blogstream.com, ebloggy.com, tabulas.com, blogspot.com. The search `A site:X` was submitted to both Google and Bing using query splitting technique to extended lists of results, and replacing *A* with each of the terms listed above, and *X* with each of the blog hosting site names above. Yahoo! could also have been used but was avoided to avoid sending too many queries to this search engine, which was required for the link searches. After combining the Google and Bing results for each term, a list of all unique domain names (i.e., blogs) returned by searches for each term was extracted by LexiURL Searcher. These lists formed the set of "issue-relevant" blogs.

7.2 BIA 2: Blog Size Distribution

For each blog B identified in the above stage, the query `site:B` was submitted to Bing in order to get an estimate of it size (number of pages). Although all three search engines could have been used, only Bing was used in order to cut down the total number of searches sent to the search engines, because 107,585 queries were needed, one per blog. For each case, a graph was plotted of the number of

Search	#Blogs	Median pages	Largest blog matching search, excluding official blog site blogs
Wicca	9,379	6	timjblair.spaces.live.com "Activist blog U.S. army Vietnam veteran, longtime civic leader and broadcaster the independent voice of Seattle"
ISKCON	1,396	6	linan1978.spaces.live.com – personal blog
Hari Krishna	1,021	16	eternalsunshine-on-me.spaces.live.com – personal blog
Scientology	11,598	13	trixie.spaces.live.com – personal blog
Islam	20,095	14	d3vmax.spaces.live.com – hack MSN technical blog
Christ	29,499	14	aceybongos.spaces.live.com – Xbox promotional blog
Because	34,597	16	blogquest.spaces.live.com – personal blog

Table 1. Median sizes of blogs matching each search.

blogs of each size. It is interesting that blogs for all of the issues, except for Hari Krishna, are smaller than the benchmark figure of 16 for 'because'. This suggests that the issues are all mentioned predominantly in smaller blogs, which are presumably mainly amateur, informal blogs.

The shape of the graphs was similar in each case, with Figure 1 showing the shape most clearly. It is a classic "hooked power law" (Pennock, Flake, Lawrence, Glover, & Giles, 2002). The basic power law suggests a rich-get-richer/Matthew Effect phenomenon. Blogs with many pages are extremely rare in comparison to blogs with only a few pages. Nevertheless, it is common for a blogger to create blogs with up to 50 pages, giving the relatively flat hooked part of the graph on the

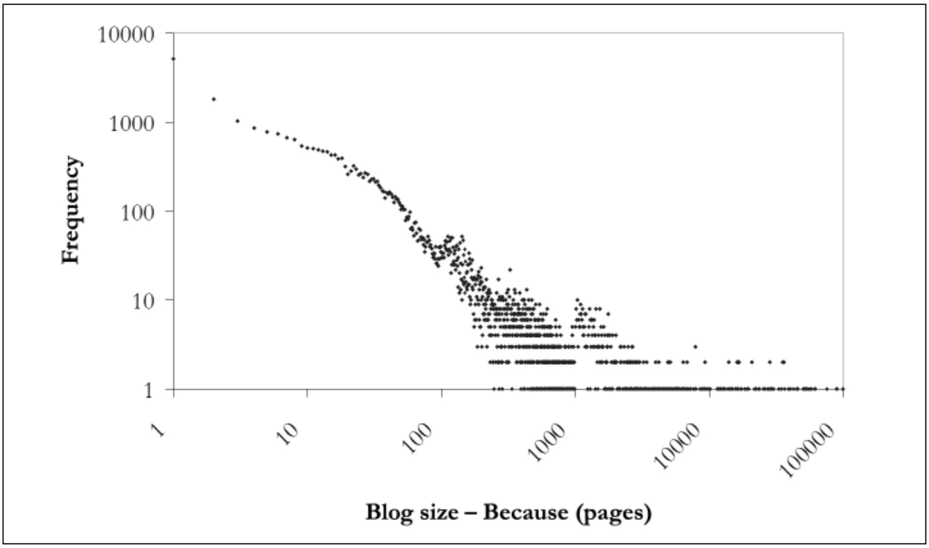


Figure 1. Distribution of the number of pages in blogs including the word "because" (note the log-log scale).

left. This is probably due to the blogging software automatically creating archived pages. There is also a peak at just over 100 pages which is probably also an artefact of the blogging software. The extra peak at 1000 is an artefact of search engine reporting, as discussed in the next section.

Table 1 also lists the largest blog for each search, after excluding blogs associated with the owner of the blogging service (e.g., technical support pages, general photo pages). None of the results are very interesting, however, since most are personal blogs in spaces live that have many pages partly as a result of the way in which Spaces Live organises its blog sites.

7.3 BIA 3a: Blog inlink count distribution

All the inlink graphs except Hari Krishna and ISKCON have a distinctive shape, as illustrated below for Christ blogs (Figure 2). The additional two peaks at 1000 and 10000 are artefacts caused by the search engine rounding its results for values over 1000 (Thelwall, 2008; Uyar, 2009a, 2009b). Ignoring this, the graph is a pure power law. The other two graphs have too few points to have a distinctive shape but could be interpreted as partial versions of the same graph. The graph shape similarities suggest that no new information can be found by comparing them. The slopes of the neck of the graph (left hand side) varied from -0.77 to -1.08, with the larger graphs having slopes very close to -1. The slopes were too similar to read any significance into the differences. It was not possible to identify outliers as the blogs with most inlinks were in all cases very general blogs. In summary, *the inlink count distribution did not yield useful information.* Table 2 lists the most popular (most linked to) blogs for each topic. This is not really

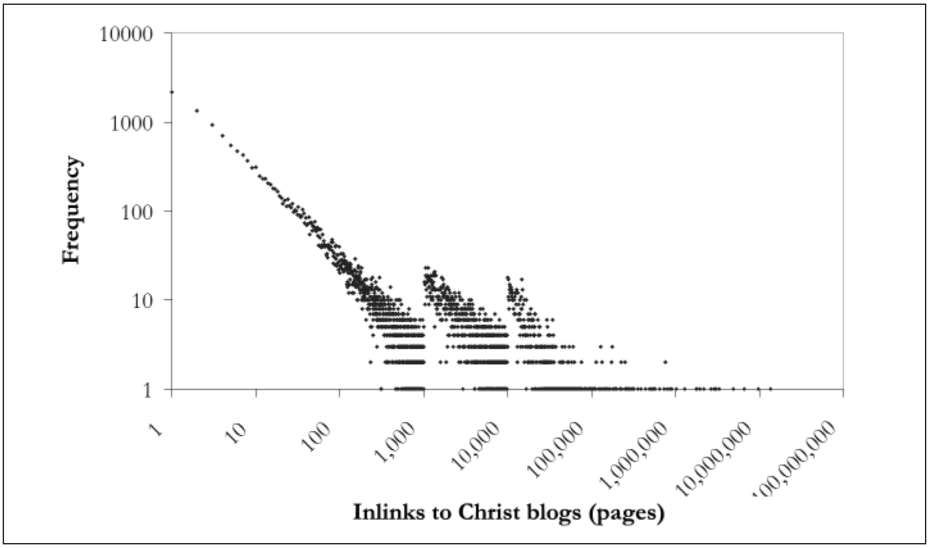


Figure 2. Distribution of the number of links to blogs mentioning Christ (note the log-log scale).

Search	Name and description
Wicca	atrios.blogspot.com Eschaton blog – a personal but popular blog with an author associated with Media Matters for America, an organisation "dedicated to comprehensively monitoring, analyzing, and correcting conservative misinformation in the U.S. media" http://mediamatters.org/p/about_us/
ISKCON	dilbertblog.typepad.com The general and political blog of Scott Adams, the Dilbert cartoonist.
Hari Krishna	timesonline.typepad.com The Times (UK) newspaper.
Scientology	timesonline.typepad.com The Times (UK) newspaper.
Islam	awis.blogspot.com Alexa search engine blog
Christ	atrios.blogspot.com See Wicca.
When	esnippers.typepad.com – the personal blog of Lonnie Hendrickson

Table 2. Blogs matching each search with most inlinks.

useful, however, as they tend to be very general blogs rather than blogs that are mainly about the topic or important to the topic. It seems that a more complex technique would be needed to identify blogs that are important to the topic (Kleinberg, 1999).

7.4 BIA 3b: Blog inlink URL distribution

For each base query, (Christ, Wicca etc.) the lists of inlinking pages (i.e., pages linking to blogs matching a base query) were merged into one combined list and the number of unique URLs, domains, sites and TLDs were counted. Figure 3

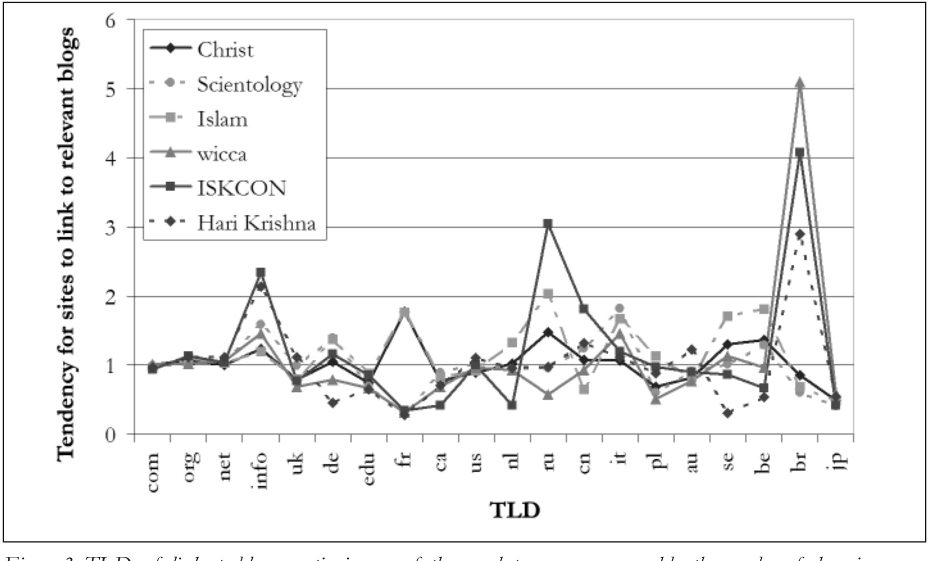


Figure 3. TLDs of links to blogs mentioning one of the search terms, as measured by the number of domains matching the TLD relative to the number of domains matching the TLD for the term "when".

illustrates the TLDs of web sites linking to the blogs, relative to the number for the query "when". Whilst the proportion of queries matching the main gTLDs is approximately the same for all religions terms, they vary by country. ISKCON seems particularly successful in Russia and Brazil attracts many links for three of the terms: Wicca, ISCKON and Hari Krishna. All of these seem genuine country associations except for Wicca, which is prominent in Brazil through posting the lyrics of a band of the same name. Although the band is relevant (e.g., one song is: "An Elizabethan Devil Worshipper's Prayer Book") its popularity probably signifies little about the popularity of Wicca itself.

7.5 BIA 4: Content analysis

Table 3 reports an estimate of the total number of correct and appropriate matches for each search, based upon a content analysis of a random sample of blogs. The final column estimates the total number of blogs that correctly and appropriately

Search	#Blogs	Classification sample	Correct mention	Appropriate mention	Estimated # appropriate
Wicca	9,379	118	102 (86%)	94 (80%)	7,471
ISKCON	1,396	258	236 (91%)	224 (87%)	1,212
Hari Krishna	1,021	184	122 (66%)	98 (53%)	544
Scientology	11,598	128	105 (82%)	94 (73%)	8,517
Islam	20,095	162	119 (73%)	100 (62%)	12,404
Christ	29,499	156	122 (78%)	100 (64%)	18,910

Table 3. Accuracy of religious blog identification: Correct mentions indicate the term found in the blog page and appropriate mentions indicate that the religion is mentioned or alluded to.

Search	Blogger	Common topics
Wicca	Follower, studying Wicca, sympathiser	Personal statement or list of interests, descriptions of Wicca activities
ISKCON	Sympathiser, follower	Religious thoughts, temples
Hari Krishna	Sympathiser, opponent, neutral, followers	Joke, public processions, temples, George Harrison, meeting followers
Scientology	Opponent, sympathiser/ follower	Warnings, conspiracy theories
Islam	Student of Islam, follower, opponent	Book reviews, terrorism, Islamic 'intolerance', Islamophobia
Christ	Follower	Religious thoughts, personal blog, discussion of the film: "The passion of the Christ"

Table 4. Common discussion topics and blogger orientations.

mention the religion. Incorrect mentions were classes as those where the term was not found in the blog page and inappropriate mentions were use of the term outside of the context of the religion, such as personal or place names and typos.

Table 4 reports the common topics of blog posts and the common types of blog orientation to the religion. It is clear in the case of Hari Krishna and its near-synonym ISKCON that the choice of term makes a big difference in the types of blog and topic returned. It is also clear that the average context in which the religions are discussed varies significantly. A more in-depth content analysis could make these findings more precise and verify them with additional coders in order to establish inter-coder consistency.

8. Discussion and conclusions

The blog issue analysis methods introduced and demonstrated in this paper reveal some interesting but expected findings about how blogs can be studied on the web. Note, however, that the case study is exploratory rather than thorough and that, as a consequence, no conclusions should be drawn about the religious topic from the data presented here.

Although some of the results did not seem to give useful information, particularly the more abstract numerical graphs, other sections were more revealing about the issues studied. In Table 1, the number and median size of the blogs matching each issue search seems to be useful background information to understand the context of issue-related blogging. Similarly, the TLD distribution graph Figure 3 seems useful to compare the international spread of influence for issue blogging, for example showing a high Russian affiliation for ISKCON. The data on the percentage of blogs matching the search that directly alluded to the issue in Table 3 seems valuable to understand the extent to which an issue is tightly defined (e.g., ISKCON, Wikka) in the sense of being predominantly mentioned in a relevant context (c.f., Hari Krishna). Perhaps most valuable is the data in Table 4, however, the results of the content analysis. Although only reported briefly, the differences between the issues in the way in which they are typically discussed are considerable. For example, Scientometrics is typically discussed in a very negative fashion, Islam discussions are more mixed and ISKCON is quite religious.

In terms of future developments of BIA, the core techniques can be applied to particular issues for extended case studies and a more robust assessment of the value of the approach. Extensions are also possible, perhaps including appropriate visualisations, such as network diagrams. One of the unsuccessful aims of BIA was to identify important or large blogs relevant to the issue. This did not work because the blogs found were either general or did not seem particularly relevant.

A new method is therefore needed to identify the key blogs for any given issue. This may well exploit Ingwersen's link analysis and Kleinberg's algorithm to separate out the important relevant blogs.

References

Ackland, R. (2005). *Mapping the U.S. political blogosphere: Are conservative bloggers more prominent?* Retrieved February 3, 2006, from http://acsr.anu.edu.au/staff/ackland/papers/polblogs.pdf

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. *WWW2005 blog workshop*, Retrieved May 5, 2006 from: http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf.

Bar-Ilan, J. (2005). Information hub blogs. *Journal of Information Science, 31*(4), 297-307.

Barabási, A. L. (2002). *Linked: The new science of networks.*Cambridge, Massachusetts: Perseus Publishing.

Bhagat, S., Cormode, G., Muthukrishnan, S., Rozenbaum, I., & Xue, H. (2007). No Blog is an island - Analyzing connections across information networks. *International Conference on Weblogs and Social Media 2007*, Retrieved March26, 2007 from: http://www.icwsm.org/papers/2002--Bhagat-Cormode-Muthukrishnan-Rozenbaum-Xue.pdf.

Björneborn, L. (2006). 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space. *Scientometrics, 68*(3), 395-414.

Björneborn, L., & Ingwersen, P. (2001). Perspectives of Webometrics. *Scientometrics, 50*(1), 65-82.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227.

Bond, M., & Abtahi, M. A. (2005). The blogger of Tehran. *New Scientist, 188*(2521), 48-49.

boyd, d., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), Retrieved May 7, 2009 from: http://jcmc.indiana.edu/vol2013/issue2001/boyd.ellison.html.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Coleman, S. (2005). Blogs and the new politics of listening. *Political Quarterly, 76*(2), 273-280.

Dodds, P. S., & Danforth, C. M. (in press). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies.*

Elmer, G., Ryan, P. M., Devereaux, Z., Langlois, G., Redden, J., & McKelvey, F. (2007). Election bloggers: Methods for determining political influence. *First*

Monday, 12(4), Retrieved June 5, 2007 from: http://firstmonday.org/issues/issue2012_2004/elmer/index.html.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text (IDA 2005). *Lecture Notes in Computer Science, 3646*, 121-132.

Gruhl, D., Chavet, L., Gibson, D., Meyer, J., & Pattanayak, P. (2004). How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Systems Journal, 43*(1), 64-77.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. In *Proceedings of the Thirty-seventh Hawaii International Conference on System Sciences (HICSS-37).*Los Alamitos: IEEE Press.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229-247.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation, 54*(2), 236-243.

Kanayama, H., & Nasukawa, T. (2006). Fully automatic lexicon expanding for domain-oriented sentiment analysis. In *EMNLP: Empirical Methods in Natural Language Processing* (pp. 355-363). Stroudsburg, PA: ACL.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604-632.

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social media & mobile internet use among teens and young adults. *Pew Internet & American Life Project*, Retrieved February 5, 2010 from: http://pewinternet.org/Reports/2010/Social-Media-and-Young-Adults.aspx.

Mayr, P., & Tosques, F. (2005). Google Web APIs: An instrument for webometric analyses? Retrieved January 20, 2006 from: http://www.ib.hu-berlin.de/%2007Emayr/arbeiten/ISSI2005_Mayr_Toques.pdf.

Park, H. W., & Thelwall, M. (2008). Web linkage pattern and social structure using politicians' websites in South Korea. *Quality & Quantity, 42*(6), 687-697.

Pennock, D., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences, 99*(8), 5207-5211.

Pikas, C. K. (2005). Blog searching for competitive intelligence, brand image, and reputation management. *Online, 29*(4), 16-21.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics, 1*(1), Retrieved July 25, 2006 from: http://www.cindoc.csic.es/cybermetrics/articles/v2001i2001p2001.html.

Schmidt, J. (2007). Blogging practices: An analytical framework. *Journal of Computer-Mediated Communication, 12*(4), Retrieved March 3, 2008 from: http://jcmc.indiana.edu/vol2012/issue2004/schmidt.html.

Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review, 31*(3), 277-289.

Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology, 59*(1), 38-50.

Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology, 59*(11), 1702-1710.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences.*New York: Morgan & Claypool.

Thelwall, M., Vann, K., & Fairclough, R. (2006). Web issue analysis: An Integrated Water Resource Management case study. *Journal of the American Society for Information Science & Technology, 57*(10), 1303-1314.

Uyar, A. (2009a). Google stemming mechanisms. *Journal of Information Science, 35*(5), 499-514.

Uyar, A. (2009b). Investigation of the accuracy of search engine hit counts. *Journal of Information Science, 35*(4), 469-480.

Addresses of congratulating authors:

Mike Thelwall
Statistical Cybermetrics Research Group
School of Computing and Information Technology
University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, United Kingdom
Email: m.thelwall[at]wlv.ac.uk

David Wilkinson
Statistical Cybermetrics Research Group
School of Computing and Information Technology
University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, United Kingdom
Email: d.wilkinson[at]wlv.ac.uk

Ingwersen's Identity and Image Compared

Howard D. White

Drexel University, Philadelphia, USA

Abstract. A citation identity comprises data on who an author cites, ranked by how frequently. A citation image comprises data on who the same author is cited with, ranked by how frequently. The top ranks of the two constructs are displayed and interpreted here for Peter Ingwersen and myself. Both analyses support the notion that least effort plays a part in authors' behavior as they seek relevant citations. Specific examples of least-effort behavior are included.

1 Introduction

About a decade ago, I introduced citation identities and citation images as useful ways of studying an individual author—a topic also developed in later papers [1-3]. Since the present volume honors Peter Ingwersen, I will immediately use his name as my example in discussing these constructs, whose relationship I take up in a new way here. The Ingwersen citation *identity* consists of all the authors he has cited over time. When these authors are ranked high to low by how frequently he has cited them, they form a distinctive self-portrait of his intellectual world, with his top-ranked citees being presumably of greatest interest. The Ingwersen citation *image*, in contrast, comprises all the authors with whom he has been co-cited. It is the field's-eye view of his work, created over time by citers in general. It is also the starting point for the map displayed on the next page. My main goal, however, is to compare the top ranks of his identity and image, and to discuss the causes of their similarities and differences.

Peter and I won the Price Medal together in 2005, and, although he is younger, our publication and citation records are comparable. I have therefore analyzed my own identity and image as a check on the claims I make about his. As it turns out, both his data and mine lead to much the same conclusions.

1.1 Image-Based Maps

The Pathfinder Networks (PFNets) in Figure 1 set the stage. The AuthorMap system that produced them, designed by Xia Lin, Jan Buzydlowski, and me in the late 1990s, has been described at greater length elsewhere [e.g., in 4]. AuthorMap in this instance was given Ingwersen's name and mine as seeds for extracting the 24 other

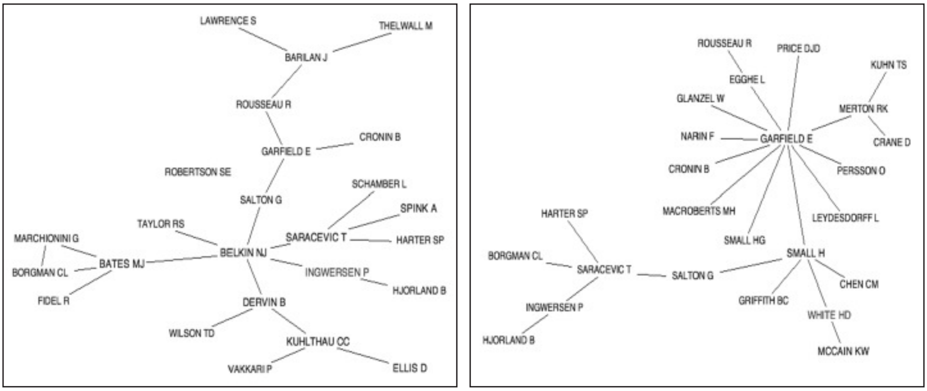


Fig. 1. Pathfinder networks (PFNets) for the top 24 authors co-cited with Peter Ingwersen (left) and Howard D. White (right). The displays were made with AuthorMap, a system created at Drexel University for on-the-fly mapping of bibliographic terms. The data were gathered from Social Scisearch on Dialog in June 2010. Names appear in the surname-initials format used in the Thomson Reuters citation databases.

authors with whom we are most frequently co-cited in the Dialog version of Social Scisearch. This covers citations made in the journal literature of the social sciences from 1980 to the present; citations from books are unfortunately excluded. Lin's algorithm then obtains the co-citation counts for every other pair of authors in the retrieved set, creating a square symmetric matrix of counts for (25*24)/2 = 300 pairs. These are passed to the PFNet algorithm, which draws links between authors on the basis of the highest (or tied highest) count for each pair; all other possible links are eliminated. A spring embedder positions the author nodes for display, as in Figure 1.

Note that Ingwersen has co-citation counts above 70, and I above 50, with every other name in our respective maps. But his highest count is with Nicholas Belkin, and mine with Henry Small, so those are the authors to whom we are attached. In my map, Small appears twice (and Anthony van Raan was left out) because AuthorWeb cannot reconcile counts for multiple forms of a name. In Ingwersen's map, Gary Marchionini would ordinarily be connected only to Christine Borgman, but he and Borgman have tied counts with Marcia Bates, which accounts for the links seen there.

A last detail: I have removed from the images the citations made by Ingwersen and myself, which by default would have been included. My reason will be apparent later. These maps thus represent our intellectual connections purely from the standpoint of other citers.

1.2 The Ingwersen Map

Ingwersen's map shows the perceived—and admirably extensive—usefulness of his work in three major subdisciplines of information science: user behavior studies, information retrieval, and bibliometrics (including webometrics). His preponderant co-citees place him squarely in the user-oriented wing of the discipline, in which Belkin is the central influence. These are the authors who, starting in the late 1970s, sought to elaborate and alter, or even to overturn, the tradition of formalistic design and evaluation of document retrieval systems they associated with, e.g., the paradigmatic research projects at the Cranfield College of Aeronautics in England and the work of Gerard Salton. Broadly speaking, they found this tradition, which dated from the 1950s, to be entirely too "systems-oriented," too little concerned with human psychology, human-computer interaction, and the behavior of real information-seekers in real settings (as opposed to judges in artificial retrieval experiments). These lacks they wished to remedy with new empirical research and new theorizing of their own—an effort that continues to the present day.

Looking back, it is possible to see the change as something of a paradigm shift or "cognitive revolution." Ingwersen's important book of 2005, *The Turn: Integration of Information Seeking and Retrieval in Context* [5], deliberately fuses the older and the newer traditions, emphasizing a cognitive framework and interactivity as foundations for viable future research. In so doing, it assimilates numerous writings by information scientists in the lower part of the Ingwersen map: the theorists on relevance linked to Tefko Saracevic, the modelers of user behavior linked to Brenda Dervin, and the investigators of interactive retrieval linked to Marcia Bates. It may be recalled that Stephen Robertson, linked in the map to Salton and still identified with the "systems-oriented" wing, was Belkin's co-author in an early paper on the cognitive framework. In fact, far from overturning the retrievalist tradition, Ingwersen and many of his co-citees have deepened it, making its assumptions more realistic.

What the upper part of Ingwersen's map implies more clearly than *The Turn* is that his interests go beyond users of retrieval systems. He has contributed, for example, to evaluative bibliometrics, as represented by Eugene Garfield and Blaise Cronin, and to webometrics, as represented by Ronald Rousseau and the other three authors at top.

1.3 The White Map

Citers see me mainly as a citation analyst specializing in co-citation studies. Small is one of the founding co-citationists; through him I am linked to Garfield, father of the Institute for Scientific Information and much of modern bibliometrics. The authors around Garfield, from Michael MacRoberts through Derek Price, are bibliometricians who involve themselves in citation analysis to varying degrees. Continuing clockwise, the sociologists Robert Merton and Diana Crane and the historian Thomas Kuhn are all famous for their social studies of science, to which they see citation analysis as at least a somewhat pertinent adjunct. Olle Persson and Loet Leydesdorff are, like me, both citationists and visualizers of co-citation data. The same holds true for other authors connected to Small or me—Chaomei Chen, Belver Griffith, and Katherine McCain, all of whom are or were my fellow professors at Drexel University. (There is a bit of a Philadelphia mafia in bibliometrics, counting Garfield, Small, Crane, and Narin as present or former members.) Interestingly, my map includes 12 recipients of the Price Medal, as well as the multi-talented Derek Price for whom it is named.

Although the authors extending leftward from Small have explicit connections to bibliometrics, they are not primarily identified with it. They represent my ties with Ingwersen's primary world of "sociocognitive" research; he himself appears, and everyone else is in his map. This is the world I am trying to move toward, both here and in other recent papers (whether my co-citation record ever captures it or not). I want, that is, to interpret citation data more psychologically, bringing them closer to the central notion of information science, which is relevance. As many readers will know, important work on relevance has been done by authors who appear in both maps of Figure 1, such as Stephen Harter, Birger Hjörland, Salton, Saracevic, and Ingwersen himself.

2 Identities and Images

For me, and I hope for anyone immersed in information science, the image-based maps in Figure 1 have a remarkable characteristic: high intelligibility. In other words, it is easy to relate the authors in them to the seeds and to say why various pairs are connected (even if the explanations are mistaken). There are felicities not only in topical connections but in social ones as well. I have mentioned the Drexel group, but look at other linked pairs who work for the same employer: Ingwersen-Hjörland, Belkin-Saracevic, Bates-Borgman, Garfield-Small. Look at co-author pairs linked on the basis of co-citation: Egghe-Rousseau, Belkin-Ingwersen, Saracevic-Spink, Small-Griffith, White-McCain. One can only conclude that the names in the maps are perceptibly *relevant* to each other in many ways. They and their positionings seem to make sense, to be not in error. One can, moreover, imagine maps that would look wrong or at least dubious, either in the co-citees named or in the linkages drawn. For example, it would be unsettling to see Ingwersen directly linked to a mathematical bibliometrician like Egghe, given what we know of their work. While not impossible, such an outcome goes against our sense of ready interpretability, which is another way of saying our immediate perceptions of relevance.

Suppose you were asked, "Which author is more relevant to Peter Ingwersen—Nick Belkin or Leo Egghe?" I think almost any information scientist would quickly say Belkin, because Ingwersen's work is easier to associate with Belkin's than with Egghe's. That makes *the effort of processing the stimuli* a key part of the perception of relevance; choosing Belkin over Egghe is a least-effort response to the question. It requires specialized domain knowledge, to be sure, but analogous questions could

Identity		Image	
Ingwersen P (seed)	41	*Ingwersen P (seed)*	783
Wormell I	17	*Belkin NJ*	273
Belkin NJ	16	*Saracevic T*	243
Garfield E	11	Spink A	186
Saracevic T	11	Kuhlthau CC	182
Bates MJ	10	*Bates MJ*	170
Christensen F	10	Dervin B	161
Cronin B	10	*Ellis D*	160
Jones KS	9	Wilson TD	148
Rousseau R	9	Borgman CL	125
Brookes BC	8	Vakkari P	123
Croft WB	8	Harter SP	121
DeMey M	8	*Rousseau R*	118
Ellis D	8	Marchionini G	113
Pejtersen AM	8	Thelwall M	111
Almind TC	7	Salton G	109
Egghe L	7	Fidel R	107
Lancaster FW	7	*Taylor RS*	107
MacKay DM	7	*Cronin B*	106
Pao ML	7	*Garfield E*	94
Schamber L	7	Hjörland B	84
Searle JR	7	*Schamber L*	82
Taylor RS	7	Bar-Ilan J	79
Glänzel W	6	Robertson SE	76
Kochen M	6	Lawrence S	73

Table 1. Top 25 names and counts in Peter Ingwersen's citation identity and citation image.

Identity		Image	
White HD (seed)	40	*White HD (seed)*	747
McCain KW	17	*Small H*	265
Griffith BC	15	*McCain KW*	212
Small H	12	*Garfield E*	207
Garfield E	11	Price DJD	151
Mullins NC	9	*Cronin B*	141
Lancaster FW	8	Leydesdorff L	105
Sandstrom PE	8	*Borgman CL*	98
Chen CM	7	*Chen CM*	87
Harter SP	7	Salton G	84
Wilson P	7	Merton RK	77
Baldi S	6	Egghe L	71
Borgman CL	6	Rousseau R	67
Drott MC	6	Crane D	66
Jones KS	6	*Hjörland B*	65
Saracevic T	6	*Griffith BC*	64
Schvaneveldt RW	6	*Ingwersen P*	63
Bradford SC	5	Kuhn TS	62
Cole S	5	Glänzel W	61
Cozzens SE	5	Narin F	60
Cronin B	5	Persson O	60
Edge D	5	MacRoberts MH	59
Hjörland B	5	*Harter SP*	58
Ingwersen P	5	*Saracevic T*	58
Kruskal JB	5	Van Raan AFJ	58

Table 2. Top 25 names and counts in Howard White's citation identity and citation image.

be asked in any scientific or scholarly domain, and the least-effort responses would, I believe, be made most frequently. In Ingwersen's case Belkin is near-fetched whereas Egghe is relatively far-fetched.

Even Egghe, however, would probably be easier for most of us to associate with Ingwersen than, say, Ludwig Wittgenstein. Although Ingwersen has not cited Wittgenstein himself, he and the famous philosopher have been co-cited in 14 articles, including one by Salton. If the articles refer to this pair in the same passages, I cannot guess why with any confidence. Would Ingwersen himself experience the same difficulty? Presumably he would have little trouble explaining why he has been so frequently co-cited with Belkin or Saracevic—why they are so relevant to him. He could use the same explanations that underpin his own heavy use of them as orienting figures in his identity.

This leads me to my main concern: to what extent do authors—Ingwersen, for example—repeatedly cite the same people with whom they are repeatedly co-cited? Do they obviously exhibit least-effort behavior as they decide which authors to cite in their own work? What varieties of such behavior can we infer from the citees they choose? Answers lie in the content of citation identities and images, and the degree of match between them. The content of authors' identities grows slowly as a by-product of their efforts to document claims in multiple publications, and it is rarely something they are much aware of. But for that very reason, it may reveal something of their thought processes as they create their intellectual worlds.

Using Dialog software on the Thomson Reuters citation databases, one can obtain an identity and an image for any given seed author. Since both constructs can be very long, Tables 1 and 2 give only the top 25 names in each for Ingwersen and myself. The data are again from Social Scisearch as of early June 2010. Counts are of citing articles, not of cited references. My focus will be on least-effort behaviors in the identities, where individual authors are clearly agents who seek to make their citations relevant to their main texts.

In the identity, a seed author cites certain people many times. In the image, the seed and certain people are co-cited many times. Now, if the two sets of top names and their frequencies are significantly correlated, it means that both the seed author and other citers find it comparably easy to associate the same names with the seed's work. If no such correlation exists, one could still argue that the seed exhibits least-effort citing behavior, but not that other citers confirm it.

Images aggregate the identities of many citers at a time, including that of the seed author unless it is removed. As in the maps, I have removed Ingwersen's and my identities from our images. (A simple Boolean command does it.) There is thus no overlap in the two sets of counts, and they can be validly correlated.

2.1 Ingwersen's Identity and Image

Table 1 shows that Ingwersen's image and identity share the 10 names in italics—one of the lower intersection counts I have seen. However, seven of the matching names are in the top 10 of his identity, where a seed's easiest choices manifest themselves though heavily repeated citation.

For the italicized names, both image and identity counts are on display. But all the names in Table 1 have image or identity counts in lower ranks not shown. Before correlating them, I removed Ingwersen's own two counts from the data because they unduly raise the correlation. When image counts are filled in for all other names in his identity, the Pearson's r is 0.48 (p=.009, 1-tailed, n=24). When identity counts are filled in for all other names in his image, the Pearson's r is 0.54 (p=.003, 1-tailed, n=24). While these are middling correlations, they are still significant below the .01 level for a relatively small number of cases. This indicates a commonality of structure in the names that Ingwersen and other authors find easiest to cite, given his name as context. The correlations might well be higher if cited references rather than citing articles were counted and data from books were added to the counts. For example, Amanda Spink, Carol Kuhlthau, and Brenda Dervin are names high in his image. He has cited multiple works of theirs in *The Turn* and in some of his articles, yet they are not among the top names in his identity.

With that identity before us, I can better justify phrases like "easiest to cite" and "least-effort behavior." The first consideration is what tops the identity—self-citation. Just as no editorial law forbids self-citation (at least within measure), so none requires it. Authors are thus free to refrain from citing themselves, perhaps even routinely, as a matter of principle. That so few do, and that so many cite themselves more frequently than anyone else, suggests that a cognitive principle is at work. In considering the documents relevant to a writing being composed, published scholars and scientists tend to lessen their effort by choosing items they already know best—those they have authored themselves. Cognitively, they build from the inside out. This is not egotism; it is simply the way the mind operates, and Ingwersen in this regard is no exception.

The second consideration is social ties. Such ties are generally less important than the intellectual relevance of writings, but intellectually relevant writings gain in salience if the citer also knows the people who wrote them. In Ingwersen's identity, the second most cited author is his wife. She and others among his citees in Table 1 —Finn Hjorthgaard Christensen, Annelise Mark Pejtersen, Tomas Crone Almind—are or were colleagues associated with his school in Denmark. When he cites them, he is sometimes citing documents he co-authored. I believe he is also personally acquainted with many of the remaining authors in the table and, again, has been a co-author with some of them.

Third are explicit terminological links between Ingwersen's writings and those of his citees—links that make their mutual relevance plain to see. A large-scale content analysis would be needed to produce the evidence, from titles, abstracts, subject indexing, and bibliographically coupled citations, if not full text. I would simply note here that authors such as Belkin and Saracevic are far likelier to produce complete or partial matches with Ingwersen's terms than, say, the philosopher John Searle, who is lower in the ranks of his identity.

The last effort-saving item in this non-exhaustive list is the use of what I will call "personal concept symbols." Small [6] found that when many citations accumulate for a scientific paper, scientists are often using that paper as symbolic shorthand for one particular concept. The same thing can happen at the individual level. That appears to be the case, for example, with Ingwersen's citations to the neuroscientist Donald MacKay, all of which invoke his radio talk "What Makes a Question?" published in *The Listener* in 1960. Mind you, there is nothing wrong this; it is extremely common for authors (me included) to re-use a valued concept symbol again and again rather than reading and citing ever-new things. But it does show cognitive economy at work.

2.2 White's Identity and Image

It will be seen in Table 2 that my identity and image have 12 names in common, and, as with Ingwersen, seven of those are in the top 10 of the identity. When image counts and identity counts are correlated for all names in my identity except mine, Pearson's r is 0.63 (p=.001, 1-tailed, n=24). Doing the same thing for the names in my image, the r value is, coincidentally, also 0.63 (p=.001, 1-tailed, n=24). (Counts have been emended in the image for the variant name-forms of Small, Price, and Chen.) Again, a considerable degree of non-random structure appears in the data. Again, citers in general have made many of the same intellectual connections that I did, with roughly comparable ease.

I can add with assurance that my citation identity was not intentionally produced; it simply accreted over the years as I followed paths of least effort (including high self-citation and the citation of acquaintances, especially local ones, that I noted in Ingwersen). Furthermore, I certainly had no control over those who produced my citation image. I would be very surprised if the same is not true of Ingwersen and every other academic author.

What causes the differences in names that move to the top ranks of an author's identity and image? Ingwersen's identity and mine provide some strong clues. The identity names that get demoted in the image are those whose connection to the seed is more personal, more idiosyncratic, and therefore harder for information scientists in general to see. For instance, Ingwersen's Scandinavian authors disappear from the top 25 ranks of his image, as do his non-information scientists—Donald MacKay, Marc DeMey, and John Searle. In my case, non-information scientists whose work I use as "personal concept symbols"—Nicholas Mullins, Roger Schvaneveldt, David Edge, Stéphane Baldi, and J. B. Kruskal—likewise disappear. So do colleagues whose relevance to my publications is more apparent to me than others, such as Carl Drott (another Drexel professor), Patrick Wilson (my dissertation advisor), and Pamela Sandstrom (on whose dissertation I consulted). In Ingwersen's image and mine, such connections are not part of the mainstream consensus on "who most goes with whom." Images, in truth, bear some comparison to typecasting. They are full of associations that are relatively easy for our publics to make. Identities are closer to how we see ourselves. It is to Ingwersen's credit, damn him, that he seems a bit less typecast than I do.

3 A Final Note

As some may have noticed, my arguments about "ease of processing" and "least effort" as components of relevance are taken directly from Sperber and Wilson's relevance theory, a major specialty at the interface of linguistic pragmatics and cognitive science [7]. Rather coyly, I have not introduced RT in the present context because the exposition would require more space than a short paper like this affords. But I have an article in progress that will do some of the necessary bridge-building, with evidence taken from citation identities and images as well as other constructs from bibliometrics. The goal is to use bibliometric data to make points about relevance-seeking in human cognition. That, of course, is a longstanding interest of Peter's and a good reason for continuing to read his work.

References

1. White, H. D. Toward Ego-Centered Citation Analysis. In: Cronin, B., Atkins, H. B. (eds.) The Web of Knowledge: A Festschrift in Honor of Eugene Garfield. pp. 475–496. Information Today, Medford, NJ (2000).
2. White, H. D. Authors as Citers over Time. JASIST 52, 87–108. (2001)
3. White, H. D. Author-Centered Bibliometrics through CAMEOs: Characterizations Automatically Made and Edited Online. Scientometrics 51, 607–637. (2001)
4. White, H. D., Lin, X, Buzydlowski, J., Chen, C. User-Controlled Mapping of Significant Literatures. PNAS 101 (suppl. 1), 5297–5302. (April 6, 2004)
5. Ingwersen, P., Järvelin, K. The Turn: Integration of Information Seeking and Retrieval in Context. Springer, Dordrecht, The Netherlands. (2005)
6. Small, H. G. Cited Documents as Concept Symbols. Social Studies of Science 8, 327–340. (1978)
7. Sperber, D., Wilson, D. Relevance: Communication and Cognition. 2nd Ed. Blackwell, Oxford, UK, Cambridge, MA. (1995)

Address of congratulating author:

Howard D. White
College of Information Science and Technology
Drexel University, Philadelphia, PA 19104 USA
Email: whitehd[at]drexel.edu

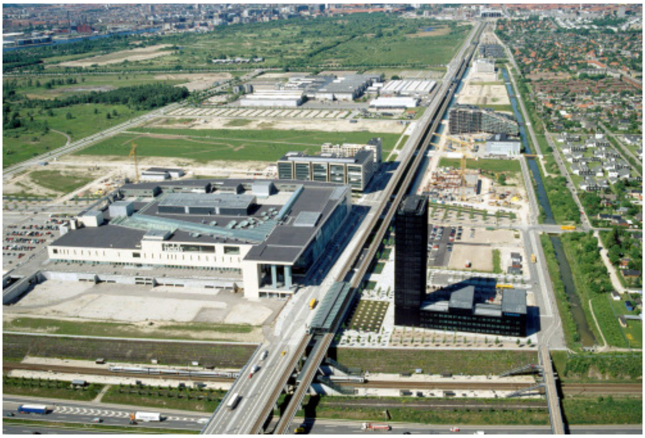
Information Science

Peter som pendler: Undervejs i Ørestadens arkitektoniske univers

Nan Dahlkild

Det Informationsvidenskabelige Akademi, København, Danmark

Den, der som Peter har pendlet med tog under og over Øresund og med metro langs Amager Fælled, har nøje kunnet følge udviklingen af en helt ny by, der er vokset frem inden for det nye årtusinds første årti. Det er gået stærkt med både planlægning, projektering og byggeri. Fælleden er forandret for altid. Amagers hulter-til-bulter af gamle villabyer, kolonihaver og sociale boligbyggerier er blevet udfordret af en ny avantgardearkitektur. Ja, selv uddannelsesinstitutioner med etableret beliggenhed i området som den daværende Danmarks Biblioteksskole har fået konkurrence fra det nye KUA, fra IT-Universitetet og fra en medieinstitution som Danmarks Radio. Nysgerrige må spørge sig selv: Hvorfor og hvorledes?



»Ørestaden set mod nord fra metrostation Ørestad men metroen og Amager Boulevard som trafikkorridor. Forrest Field's og Ferrings højhus, i mellemgrunden Bella Center og Amager Fælled og øverst Ørestad Nord. Øverst til højre ligger Det Informationsvidenskabelige Akademi. Foto: Ørestadsselskabet.«

Forhistorien

Ørestaden er et vidt begreb, som oprindelig blev brugt bredt om hele Øresundsregionen, men nu oftere anvendes mere afgrænset om den nye københavnske bydel langs metroanlægget på Vestamager eller endnu mere præcist om metrostationen Ørestad. Da byplanlæggerne begyndte at arbejde med begrebet i halvtredserne, omfattede det hele Øresundsregionen fra Helsingør og Helsingborg i nord til Køge og Trelleborg i syd. I dagbladet på Politiken blev der i 1959 skitseret et stort samlet byområde med forslag om to broer, lufthavn på Saltholm og atomkraftværk på Hven. Overskriften lød »Øresund som indsø i Nordeuropas største bylandskab«. Visionen forudså de kommende årtiers udvikling og vækst i området, men blev ikke gennemarbejdet som egentlig plan og derfor heller ikke realiseret.

Planer for området fortsatte dog med at dukke op. Københavns Kommune ønskede at inddrage Vestamager i byudviklingen, så der kunne bygges flere boliger i den efterhånden fuldt udbyggede kommune. Det første initiativ blev Urbanplanen, opkaldt efter kommunens daværende socialdemokratiske overborgmester Urban Hansen, der i 1962 var blevet valgt på bl.a. et løfte om at bygge 25.000 flere boliger i kommunen. Urban-planen kom til at bestå af Remisevænget i vest, Hørgården i nord og Dyvekevænget i øst. Remisevænget og Hørgården blev opført som karakteristisk ensartet montagebyggeri med hvide betonelementer og sort træværk i perioden 1965-1971. Boligområdet kom til at omfatte 2.500 boliger med i alt cirka 5.400 indbyggere. Der blev plads til grønne områder og legepladser, men i uhyre standardiseret udgave, der hurtigt kom til at virke nedslidt. Byggeriet blev i flere betydninger forbundet med »urbaniseringen« af Amager: Kolonihaver måtte vige for betonbyggeri. Det fik langt fra samme spektakulære udformning som efterfølgende byggerier, men er i de seneste år blevet renoveret i samme farveskala som Ørestadens nye bebyggelser.

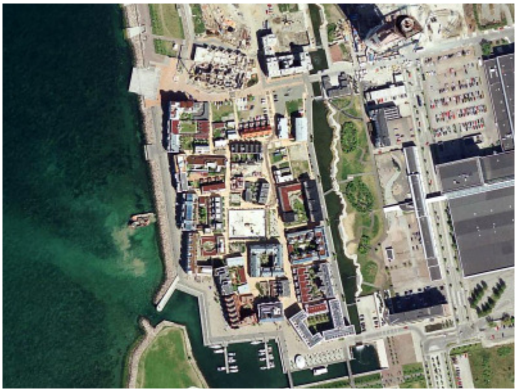
Interessen for bebyggelsesmæssig udnyttelse af Vestamager resulterede endvidere i udskrivelsen en idékonkurrence i 1964-65. Vinderforslaget foreslog et system at båndbyer over hele Amager Fælled.

Det store skift i udviklingen kom med vedtagelsen af Loven om en fast forbindelse over Øresund og den sammenhørende Lov om Ørestaden i 1991. Her fastlagdes de principper, som blev styrende for hele det efterfølgende anlæg. Det kan illustreres med et par citater fra lovens par. 12: »Ørestaden skal have en bymæssig udformning, der er af høj arkitektonisk kvalitet som en pendant til City og Middelalderbyen og samtidig med hensyntagen til områdets naturværdier. Stk. 2 Den nordlige del af Ørestaden anvendes primært til anvendelse af højere læreanstalter, forskningsinstitutioner, forskningsbaserede virksomheder, boliger og kulturelle anlæg. Den sydlige del af Ørestaden anvendes primært til cityorienteret erhvervsbebyggelse med mulighed for etablering af kulturelle anlæg og boliger.« Mellem linjerne kan man her læse om forestillingen om et internationalt kraftcenter med mulighed for tiltrækning af kapital, forskning og den eftertragtede kreative klasse.

Det blev samtidig vedtaget, at der skulle afholdes en offentlig arkitektkonkurrence, og at Ørestadsselskabet skulle forstå en udvikling i etaper, svarende til metroanlæggets stationsudbygning. Selskabet stiftedes i 1993. Finansieringen af den kommende metro blev dermed nært forbundet af salget af grunde, der ejedes af staten og kommunen i fællesskab, og som steg i værdi med netop anlægget af metroen og attraktionsværdien ved tilstedeværelsen af store offentlige institutioner som Københavns Universitet og Danmarks Radio som dele af komplekset.

Den offentlige arkitektkonkurrence blev udskrevet i 1994 med 4 vinderprojekter, hvoraf især den finske tegnestue APRT kom til at præge udviklingen med forslag om en blå-grøn bydel med både en slynget »landskabelig kanal« og en lige »universitetskanal« og dermed det blå vand som bydelens »genius loci« – elementer, der genfindes i den endelige udbygning.

Heller ikke i denne fase af Ørestadens udvikling skortede det på kritik: Fra den overordnede vækstfilosofi af Ørestaden som internationalt økonomisk centrum til de konkrete placeringer af de to letbane/metrolinjer i udkanten af og ikke i midten af Amagers tættest beboede områder. Øresundsbroen blev i en bogtitel kaldt *Bro til drømmeland* og i debatbogen *Manhattan på Amager* forudså et kapitel profetisk de store omkostninger i forbindelse med flytningen af Danmarks Radio. Med Ørestadsselskabet som lokomotiv var udviklingen af den nye bydel dog sat på skinner.



»Luftfoto af Malmøs havnefront. Bo01 er blevet til bydelen Västra Hamnen. Til venstre ses "muren" langs havnepromenaden. Bag den ligger de lavere huse omkring små pladser og knækkede stræder. Mod øst et parkområde med lavt vand. Øverst mod nord Santiago Calatravas højhus Turning Torso. Foto: Skyscraperpage Forum.«

Bo01 i Malmø

Også i Malmø gav anlægget af Øresundsbroen anledning til nye initiativer. Broen blev indviet i år 2000, og i 2001 gennemførte Malmø den store boligudstilling Bo01, som senere er blevet til en selvstændig bydel Västra Hamnen ved Malmøs havnefront. Malmø Stad havde allerede i 1997 opkøbt det tidligere havne- og industriområde af Kockum Skibsværft og Saab for at omdanne det til et nyt postindustrielt bolig- og erhvervsområde. Boligmessen var et ambitiøst forsøg på at sammentænke nye boligformer, æstetik, økologi og teknologi med mangfoldighed i arkitektur og byrum. Bl.a. deltog arkitekter fra hele Europa. I forhold til udviklingen af den københavnske havnefront på samme tid, som var domineret af økonomiske interesser, var der således tale om et langt mere visionært og helhedsorienteret projekt.

Området blev bygget op omkring en helhedsplan, hvor de høje bygninger langs havnepromenaden som »mur« ud mod det barske klima fra havet beskyttede den lave og mere intime bebyggelse indenfor, som med sine smalle og uregelmæssige stræder, gyder, pladser og små grønne områder og haver kunne minde om de gamle købstæder langs sundet. Her var arkitekturen bevidst varieret og eksperimenterende med materialer og farver. Parkområder med små søer modsat havnepromenaden skabte rolige og børnevenlige fællesarealer. Udstillingen viste også, hvordan man kunne bo i husbåd.

Ud mod sundet var »muren« præget af store transparente glasfacader. Fra de øverste penthouse-lejligheder var der panoramaudsigt til Øresundsbroen og Københavns tårne. Fra de nederste lejligheder kunne man følge med i livet på havnepromenaden, og fra havnepromenaden fik man et fint indblik i, hvad der foregik i lejlighederne. Man kunne både se og blive set.

Efter udstillingen opførtes Santiago Calatravas snoede tårn Turning Torso, som kan ses over store dele af Øresundsregionen. Tilsvarende er der fra tårnet udsigt til Falsterbo och Trelleborg i syd, Helsingborg i nord og København i vest. Med sin højde på 190 meter, fordelt på 54 etager, er tårnet Skandinaviens højeste bygning. I dag rummer bydelen Västra Hamnen ca. 600 boliger, kontorer og butikker med havnepromenaden som attraktion.

Fire Ørestæder

På den københavnske side var broen sammen med anlægget af metroen en vigtig forudsætning for udbygningen af Ørestaden i forskellige etaper. Udviklingen fulgte de retningslinjer, der var nedfældet i loven om Ørestaden og yderligere konkretiseret i arkitektkonkurrencen, der blev videreudviklet til en helhedsplan i samarbejde mellem den finske tegnestue APRT og de danske KHR-arkitekter, som også kom til at stå for selve metroens design. Ørestadsselskabet, der blev dannet i 1993, varetog udviklingen frem til 2007, hvor det for bl.a. at undgå unødig indbyrdes konkurrence fusionerede med Københavns Havn i et nyt samlet selskab By & Havn, der varetager udviklingen af både Ørestaden og Havnefronten. Trafikalt blev området bundet sammen af metroen, Øresundsbroen med tilhørende anlæg af jernbane og motorvej og Kastrup Lufthavn med udvidelser.

Det smalle område langs metroen, der er 5 km langt og kun er 600 m bredt på det bredeste sted, er typisk blevet sammenlignet med et slips, der ligesom den tidligere Fingerplan i Københavns vestegn fordeler bebyggelsen langs trafikkorridorer og stationer. Ørestaden består således af 4 kvarterer: Ørestad Nord, Amager Fælled, Ørestad City og Ørestad Syd. Områdets samlede areal er 310 hektar, hvor der efter planerne i år 2020 vil bo ca. 20.000 mennesker. Der er planlagt blandede bebyggelser med 60 % erhvervs- og kontorbyggeri, 20 % boliger og 20 % øvrigt byggeri til anvendelse for kultur, service, handel og institutioner. Tilsvarende forventes der i fremtiden at være mellem 60.000 og 80.000 arbejdspladser og ca. 20.000 studerende. I år 2010 bor der ca. 5000 mennesker i Ørestaden.

Står man på metrostation DR-Byen, tidligere Universitet, har man som pendler et godt overblik over Ørestaden. Her kører metroen som højbane, og stationen er hævet over jorden med panoramaudsigt over Fælleden. Man kan se over til Havnefronten, Ørestad Nord med Frelserkirkens spir som point de vue for den lige kanal og Ørestad City med højhusene som bebyggelsesmassiv mod syd. Her kan man også blive opmærksom på en af de tre store skulpturer, som binder metroforbindelsen og Ørestaden kunstnerisk sammen, nemlig Per Kirkebys skulptur i røde mursten, Bjørn Nørgaards »Kærlighedens Ø« og længst mod syd Hein Heinsens spiralskulptur »Den store udveksler«.

Per Kirkebys murskulptur markerer overgangen mellem bebyggelse og fælled, mellem kultur og natur. I lighed med hans øvrige skulpturer bl.a. ved Gentofte Bibliotek og Humlebæk station er der tydelige arkitektoniske elementer, her døre og vindueshuller. Gennem de øverste vinduer kan man fra stationen se både ud over Fælleden og over mod byens tårne. Han har selv udtalt om skulpturen: »Lige dér står den underlige – ja, det er jo ikke nogen rigtig bygning. Den har ikke noget tag, og man kan ikke gå ind i den. Men den står lige der, hvor naturen kommer strømmende ind. Hvis man kommer fra Fælleden, møder man skulpturen som et varsel om byen. Skulpturen er så mærkelig og karakteristisk, at hvis man sender børnene med Metroen, kan man sige til dem, at de skal stå af ved den mærkelige murstensting.«

Hans fascination af mursten stammer fra hans opvækst i Københavns Nordvestkvarter omkring Grundtvigskirken med murstensbyggeri af høj håndværksmæssig kvalitet.

Længere mod syd ligger Bjørn Nørgaards »Kærlighedens Ø« som en øskulptur med en stiliseret pavillon omkring en søjle med kvindefigurer. Længst mod syd befinder sig Hein Heinsens spiralskulptur »Den store udveksler« med en dobbeltspiral med mange fortolkningsmuligheder. Også her er temaet udvekslingen mellem by og land. Samtidig afsluttes aksen med metro og boulevard fra Frelserkirkens snoede spir. Alle tre skulpturer kan ses fra metroen. De er bekostet af Ørestadsselskabet, Statens Kunstfond og År 2000 Fondet.

Hvis man i stedet for at tage metroen, bevæger sig ned i Ørestad Nord, venter der en række spændende arkitektoniske oplevelser. Bydelen er præget af store offentlige medie- og uddannelsesinstitutioner som Danmarks Radio, IT-Universitetet og KUA, Københavns Universitets Humanistiske Fakultet på Amager, blandet op med boliger og kollegier. Indkøbsmæssigt betjenes kvarteret af en enkelt Døgnnetto. Det gennemskæres af to kanaler, den lige Universitetskanal eller Emil Holms Kanal i institutionsområdet og den snoede »landskabelige« kanal i boligområdet. Mange kender sikkert området som kulisse for Danmarks Radios nyhedsinterviews: Jean Nouvels blå koncertsal, kanalen med piletræer og den høje etagebebyggelse med glasaltaner.

Det første interessante byggeri er multimediehuset DR-Byen i 4 sektioner med Jean Nouvels koncerthus som det mest iøjnefaldende. Byggeriet kan som de øvrige bebyggelser ikke beskrives udtømmende her, men det kan anbefales at læse videre i fagtidsskrifterne. Koncerthuset er i det ydre omgivet af en koboltblå skærm, der i mørke kan bruges til billedprojektioner. Om dagen kan med den rigtige belysning se den indre meteor, der udgør koncerthusets kerne. Den er inspireret af meteoren i Peter Høegs roman *Frøken Smillas fornemmelse for sne*. Der er i alt 4 koncertsale. Den største er opført som terrasser, der skyder sig frem fra alle sider omkring orkesteret. Ligesom koncertsalen i det gamle Radiohus er den beklædt med træ i gyldne nuancer.

Længere mod nord ligger de to store uddannelsesinstitutioner IT-Universitetet og Københavns Universitets Humanistiske Fakultet KUA. IT-Universitetet flyttede i 2004 til den nuværende bygning i Ørestaden, tegnet af Henning Larsens Tegnestue. Den består af to fløje i 5 etager, der omgiver et overdækket atrium. Lokaler til møder og gruppearbejde skyder sig ud i atriet som skuffer, der er trukket ud i forskellig længde. På skuffernes gavle vises digitale mønstre. I det yderste hjørne af Ørestad Nord fik KUA en ny afdeling 1997-2002, tegnet af KHR-Arkitekter.

Den nye afdeling adskilte sig markant fra det tidligere byggeri i grå beton med sin beklædning af lys travertin.

På den modsatte side af kanalen ligger Det Humanistiske Fakultetsbibliotek. Nærmer man sig den elegante glaskube med fritstående kampanile af glas, er der dog ikke meget, der umiddelbart signalerer bibliotek. Det er der heller ikke inde i bygningen, der er præget af gennemført hvid minimalisme med store glasflader og ovenlys gennem tre etager. Minimalismen understreges af geometriske skumgummimøbler i stærke lysegrønne og lyseblå farver, der er spredt ud over biblioteket som runde byggeklodser, der kan trilles i forskellige retninger. Bygningen er tegnet af arkitekterne Dissing+Weitling. Kampanilen – eller stelen – af glas med geometriske sorte figurer med blå streger er skabt som kunstnerisk udsmykning af Vibeke Mencke Nielsen.

Kvarteret rummer også to store kollegier, Tietgenkollegiet og Bikubenkollegiet, som også begge er eksempler på eksperimenterende arkitektur. Tietgenkollegiet, som er tegnet af Boje Lundgaard og Lene Tranberg og bygget 2003-6, er opført som en cirkulær form med altaner og værelser, der skyder sig ud og ind. Den cirkulære form, som har mange forbilleder i arkitekturhistorien, er her bl.a. blevet inspireret af den sydkinesiske tulou, som kombiner fællesskab og individuelle boliger. Den runde form skaber rammen om en fin grønnegård med bevoksning af piletræer. Kollegiet har plads til ca. 400 studerende. Det lidt mindre Bikubenkollegium er også fra 2006, tegnet af tegnestuen A.A.R.T. med beklædning af grå og orange fliser. Kunstneren Viera Collaro har udført en lysinstallation til kollegiet med begreber som Respect og Tolerance bøjet i neon.

Karen Blixen Parken og Boligslangen er almindelige boligbebyggelser. Karen Blixen Parken er tegnet af Vilhelm Lauritzens Tegnestue og opført 2004-6. Den ligger med sine 7 etager med altaner lige ud til Universitetskanalen og indeholder 180 lejligheder. Boligslangen zig-zagger hen over den landskabelige kanal med sine 320 boliger samt en daginstitution. Den ene halvdel Fælledhaven er tegnet af Domus Arkitekter, den anden halvdel Universitetshaven af Arkitema. Bebyggelsen er opført i samme periode. Mens Karen Blixen Parken er tæt højt byggeri, åbner Boligslangen sig ud mod fælleden.

Kanalerne, især den snoede landskabelige kanal, gør sig godt på tegninger, særlig set fra oven. Nogle vil måske spørge, om kanalerne kan bruges til noget: Som bassiner, til sejlads eller andet – men her er svaret nej. Kanalerne er en visuel effekt, og den manglende brug er utvivlsomt medvirkende til, at området generelt fremtræder stiliseret og kulisseagtigt.

Pendler man sydpå med metroen, har man en fin udsigt over Amager Fælled, som er bevaret som et stort fredet rekreativt område med afvekslende vådområder, lysninger, rørsump, krat, skovbevoksning og mange sjældne dyr og planter. På den østlige side af metroen ligger en smal bebyggelse med det poetiske navn »Solstriben« med punkthuse og rækkehuse i gule sten med store vinduer og altaner. Husene fra 2004-6 er tegnet af arkitekterne Boldsen & Holm, der også har stået for byggeriet af Bjørn Nørgaards huse på Bispebjerg i en noget anden stil. I december måned kan både metroens passagerer og bilisterne på Amager Boulevard følge med i mange former for juleudsmykning. Stationen kunne have heddet Amager Fælled, men hedder Sundby.

I Ørestad City ligger Field's og det store kompakte bebyggelsesmassiv nord for Arne Jacobsens Allé. Det er også til dette område, at den kendte polsk-amerikanske arkitekt Daniel Libeskind i 2006 fremlagde sin plan for hotel-, oplevelses- og forretningskvarteret Ørestad Down Town mellem Field's og Center Boulevard.

På den anden side af metroen ligger Ørestadsgymnasiet med åbne rum som ramme om eksperimenterende undervisningsformer med mange tests. Gymnasiet er tegnet af den århusianske tegnestue 3XN. Længere mod nord ligger VM-Husene med spidse fotogene altaner og VM-Bjerget, som kombinerer boliger med et camoufleret parkeringshus, tegnet af Bjarke Ingels Group. For at optimere pladsen i det smalle byområde er der få åbne parkeringspladser, men i stedet parkeringskældre og parkeringshuse. VM-Husene må i øvrigt siges at være den virkeliggjorte utopi for voyeurer. Der er store transparente vinduer i fuld rumhøjde, men sjældent gardiner. Her konkurrerer en åben livsform med TV-programmer som Big Brother eller Paradise Hotel.

Området præges dog af samme mangel på byliv og pladser som Ørestad Nord. Offentlige pladser og byrum er tilsyneladende rykket ind i Field's, som dog ikke kan leve op til at være en egentlig »Experience City«.

Ørestad Syd er området omkring metrostation Vestamager syd for Øresundsmotorvejen. Også her er der planlagt store bebyggelser til både erhverv og boliger, som kan ses på visualiseringer på Internettet og af den arkitektkonkurrence, der blev afholdt i 2006. Området er overvejende planlagt som boligbyggeri, bestående af fire store nord-sydgående karrébebyggelser, adskilt af let buede hovedgader. Udbygningen af Ørestad Syd ventes at være færdiggjort 2012-25. P.t. kradser finanskrisen.

Stadig undervejs

Selv om Ørestaden har mange klare intentioner om et internationalt miljø af høj arkitektonisk kvalitet, således som de formuleres i loven fra 1991, kan udviklingen ikke sættes på en enkel formel, og den er langt fra afsluttet. I sit forsøg på at undgå forstædernes almindelige tristesse giver Ørestaden sine steder associationer til futuristiske visioner som Antonio Sant'Elias *Citta Nuova* eller science fiction-film som Fritz Langs *Metropolis*. Den rummer mange spændende arkitektoniske eksperimenter, men bliver også kritiseret for at repræsentere mennesketom postmoderne slum, selv om den prøver at forbinde vidensinstitutioner og boligområder. Det er også for tidligt at gøre status over forestillingen om et vækstcenter, ikke bare for regionen, men for hele Nordeuropa.

Man kan med rimelighed kalde Peter Ørestadsborger i ordets brede betydning. Han bor i Malmø, arbejder i København og pendler tværs over Sundet. Han spiller stadig tennis på KB´s baner på Frederiksberg og har stadig konto i Den danske bank på Islands Brygge, som dog har forbindelse med Skånes Provinsbank.

Som transnational pendler og symbolanalytiker kan han også rubriceres som en del af den kreative klasse. Han er dog atypisk ved ikke at bo i nye byggerier som VM-husene, VM-Bjerget eller Bo01, men først i en ældre herskabslejlighed med en serie af værelser en suite ved Søerne i København og derefter i en herskabslejlighed i det indre af Malmø ved Triangeln.

Som Yuppie – forkortelsen for *young urban professional* – må man fortolke »young« i bred betydning for Peters vedkommende. Christopher Lasch har skrevet bogen om *Eliternes oprør*, hvor han kritiserer klassen af internationale kosmopolitter. Her må det tilføjes, at det er nogle år siden Peter optrådte som ungdomsoprører med mørkerød fløjlsvest, der samlede ind til Folkebevægelsen mod EF. På den anden side er han er atypisk som rygsækpendler. Her lever han ikke helt op til den minimalistisk klædte HRM-manager med korrekt sort mappe. Nuvel, ingen er fuldkommen.

Ørestaden har stadig store ubebyggede områder, især i Ørestad Syd. Her er de store projekter gået i stå. Til de fortsat ubebyggede områder hører også »ørkenen« syd for Danmarks Radio, hvor mange pendlere har deres gang fra og til metroen i stedet for at følge fortovet. Der var oprindelig planer om at placere et nyt Rigsarkiv, men områdets fremtidige status er uafklaret. Også Peter skyder genvej gennem dette »terrain vague«, som kan udvikle sig i mange retninger. Tiden vil vise hans fremtidige rute.

Ørestaden er fortsat undervejs – og det er Peter også!

Litteratur

Bo01 Malmø Parkprojekter in: Arkitektur DK, 2002:1, p. 12-19.
Bjerget i Ørestad/The Mountain in Ørestad in: Arkitektur DK, 2008:7, p. 50-67.
Gaardmand, Arne: Bro til drømmeland. København, Forlaget Urania, 1991.
Gaardmand, Arne: Dansk byplanlægning. København, Arkitektens Forlag, 1993.
Konkurrence: byliv og byrum i Ørestad Syd in: Landskab, 2007:4, p. 72-77.
Lasch, Christopher: Eliternes oprør, Hovedland, 1995. (The Revolt of the Elites, New York, W.W. Norton & Co., 1993)
Manhattan på Amager. Red. af Sven Skovmand, København, Aschehoug, 2000.
Tema: DR Byen i: Arkitektur DK, 2009:2.
Tietgenkollegiet i: Arkitektur DK, 2007:7.
VM-boliger i Ørestad/Housing in Ørestad in: Arkitektur DK, 2006:1, p. 2-19.
Ørestad. Københavns kommune/City of Copenhagen, 2003.

Lykønskende forfatters adresse:

Nan Dahlkild
Det Informationsvidenskabelige Akademi
Birketinget 6, 2300 Copenhagen, Denmark
Email: nd[at]iva.dk

When Information Science Reached the Royal School of Librarianship

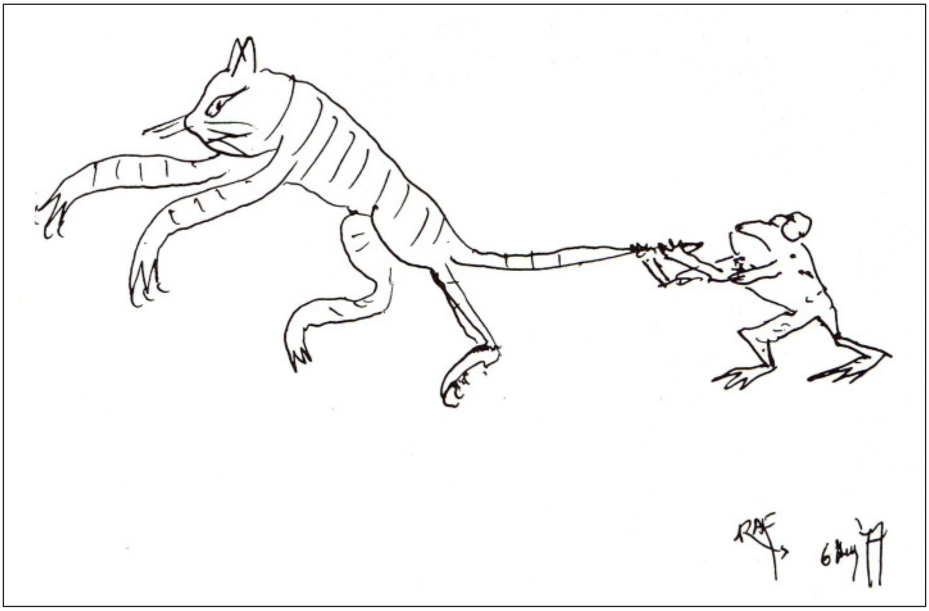
Ole Harbo

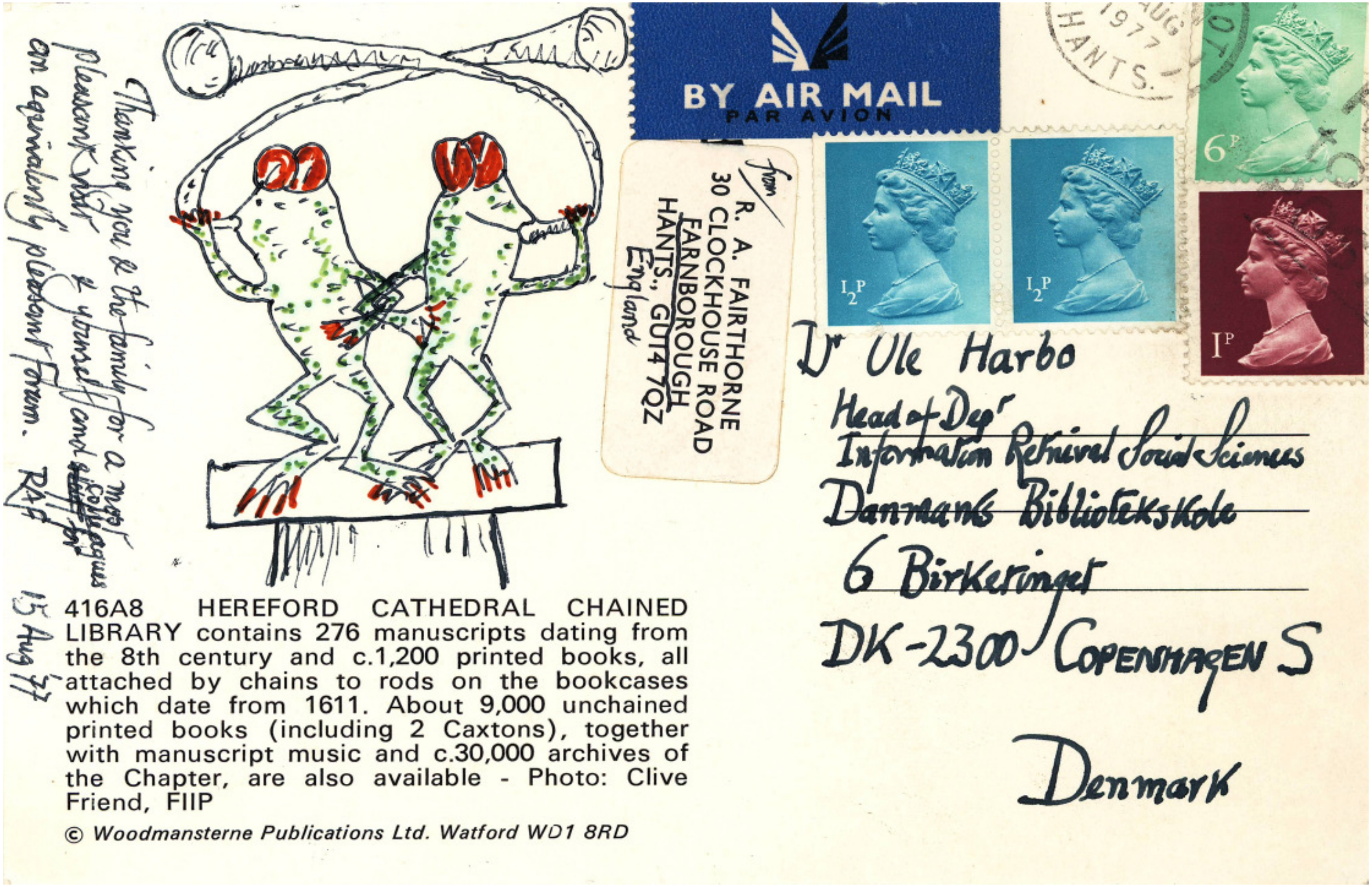
Former Head, Royal School of Library and Information Science, Copenhagen, Denmark

Abstract. The paper is about the early history of information science at The Royal School of Librarianship in the 1970`ies, how it entered into teaching, research and publishing through the work of around 10 staff members, among which were, the then young, Peter Ingwersen. The paper is mainly based on my archives from the SIRE conference.

Early research at the Royal School of Librarianship

Research is not an obvious activity at a professional school, but ever since the establishing of the independent institution, the Royal School of Librarianship in 1956, efforts have been made to have research incorporated as a natural part of the schools activities, so that the school could develop into a university for the library community.





Thanking you & the family for a most pleasant visit & yourself and ~~staff~~ colleagues for an equivalently pleasant Forum. RAF

15 Aug '77

416A8 HEREFORD CATHEDRAL CHAINED LIBRARY contains 276 manuscripts dating from the 8th century and c.1,200 printed books, all attached by chains to rods on the bookcases which date from 1611. About 9,000 unchained printed books (including 2 Caxtons), together with manuscript music and c.30,000 archives of the Chapter, are also available - Photo: Clive Friend, FIIP

© Woodmansterne Publications Ltd. Watford WD1 8RD

BY AIR MAIL
PAR AVION

from
R. A. FAIRTHORNE
30 CLOCKHOUSE ROAD
FARNBOROUGH
HANTS., GU14 7QZ
England

AUG 1977
HANTS.

Dr Ole Harbo
Head of Dept
Information Retrieval Social Sciences
Danmarks Biblioteksskole
6 Birketinget
DK-2300 COPENHAGEN S
Denmark

During the starting years efforts were concentrated on recruiting staff, development of curricula and courses and finding enough space for the fast growing institution. There were research activities at a modest level related to the traditional fields of literature-, book-, and library history and a few empirical studies about the reading habits of library users were undertaken.

A couple of the schools teachers with research potential left the school again for the university, where research conditions were much better.

After the school had moved to the address on Birketinget and after employment of many new staff members the first beginnings to information research were laid around 1970 and it was inspired and supported by two visiting American Fulbright professors, F.W. Lancaster and Allan D. Pratt.

At the organizational level research was supported by an official report on the schools future activities from 1973 in which the following topics were mentioned as relevant for research at the school: The library as an institution, information carrying materials, subject classification, documentation and library sociology. The report did not lead to a revision of the legal conditions for research at the school. The duty was still "to further research", not to do it.

Bibliometrics

Finn Hjortgaard Christensen (1937-2000) was a mathematician by education and head of the department for the natural sciences. He presented Bradford´s Law of Scattering in his early teaching material, whereas his research publications came much later in the 90´ies, some in cooperation with Peter Ingwersen.

Ole Harbo (b. 1943) economist by education and head of the social science department published in 1973 a presentation of the newly issued Social Science Citation Index and the year after the first Danish citation analysis. To this analysis the school for the first time gave a grant to a research project, 1000 DKK, to pay a student to do the citation counts.

Niels Ole Pors (b.1948) historian by education and then lecturer at the schools branch in Aalborg published about citation analyses from 1977, while Erland Munch-Petersen (1930-1997) then head of the bibliography department in 1981 was the first Dane to publish about bibliometrics in English. That same year Birger Hjørland (b.1947) lecturer at the reference department too published about bibliometrics.

Other topics

Thomas Johansen (1924-2003) head of the cataloguing department, Peter Ingwersen (b.1947) lecturer at the same department and Povl Timmermann (1931-1982) head of the technical science department started the so called JIT project, that analyzed the dialogue between users and librarians in order to improve recall and/or precision, and at the same time Annelise Mark Pejtersen (b.1944) lecturer at the fiction department started her project about classification of fiction on the basis of the criteria: subject matter, frame, authors intention and accessibility. A project which at the beginning was met with resistance from colleagues at the school and from librarians in the public libraries, but as it developed it was accepted, used and even praised in all circles.

To support the development of research at the school a research secretariat under Leif Kajberg (b.1946) was established in 1974. One of its duties was to maintain and further the schools international relations among which to deliver input to the international research database, ISORID. Harbo´s small bibliometric project was registered as the first Danish contribution and was spotted in Britain, where a large scale social science citation project was conducted at the University of Bath. Therefore an invitation to participate the first forum on information research was issued.

The IRFIS conference in London 1975

At this conference a large number of the well known American and British information scientists were present: Robert Fairthorne, B.C. Brookes, Brian Vickery and Karen Sparck Jones from Britain and Manfred Kochen, William Goffmann and Belver Griffith from the U.S.A., but also the next generation af information scientists were active participants: J.M. Brittain, Nicholas Belkin, S.E. Robertson, M.H. Heine and Marc de Mey.

At the London conference B.C. Brookes asked me to organize the next conference in Copenhagen, because he and others wanted to establish a European forum for information science, which was heavily dominated by the Americans.

With reference to the topics of the IRFIS conference I took the initiative to a study group for the Royal School´s staff members who were interested in information science themes and who would present their project for discussion in the group. About 10 persons participated and my own contribution was a 29 pages paper, Introduction to bibliometrics.

In 1976 the JIT project was presented at the EURIM 2 conference in Amsterdam and in the beginning of 1977 Ole Harbo, Peter Ingwersen and Povl Timmermann presented a joint paper at the Communication and Cognition conference in Ghent, Cognitive processes in information storage and retrieval.

Second International Research Forum on Information Science

We managed to organize the next conference in Copenhagen during the days 3.-6.- August 1977 at the Royal School under the acronym SIRE!

The organizational committee headed by Ole Harbo had as members from the schools staff Leif Kajberg, Peter Ingwersen, Povl Timmermann and Axel Andersen (b.1929) head of the reference department and the author of an overview about informatics, a term used in the Soviet Union and Eastern Europe for information science, (1976). Professor Henning Spang-Hanssen from the University of Copenhagen, Institute of Mathematical and Applied Linguistics was the external member, who had good relations to the documentation environment in Denmark at the National Technological Library and the Technical Literature Society.

The number of participants was about 100. Among the foreign participants there was a considerable overlap to the London conference, where about 50 persons participated. Fairthorne, Brookes, Vickery, Brittain, Robertson, Belkin, Heine, Goffmann, Kochen and de Mey, but in addition Kathleen Bivins, José-Marie Griffith, Tom Wilson, Ulrich Neveling, Gernot Werzig and Jean Tague (later Tague-Sutcliffe) were present.

The conference meant a major breakthrough for the information science research at the Royal School.

The school received a grant from the ministry of culture at 40.000 DKK. So the ministry was now aware of the school as a research agent.

The conference also proved that the school, technically and professionally was capable of managing such a great arrangement.

The documentation environment in Denmark, which in many ways regarded itself a competitor and a supplement to the school realized that the school was active in information science as well as in library science.

Finally the schools two own major projects, the JIT project and the classification project for fiction, were presented at an international forum with high professional competence, and both of them were well received and researchers networks were created to last for decades.

The foundation for further conferences was laid and it must be added, that the conference was mentioned in the Danish national broadcast.

The only major problem with the SIRE conference was the publishing of the proceedings publication because the publisher, with whom the original deal was made, went broke halfway through the process.

So eventually in 1980 the proceedings were issued, Theory and Application of Information research, at Mansell in London, edited by Ole Harbo and Leif Kajberg. In the foreword, which is Preben Kirkegaards opening address to the conference, it reads:" During recent years information science has received growing

Calendar

attention at our school. Members of our faculty have started teaching of this new discipline, and it has been considered in connection with our curriculum building. Besides information science related research approaches have gradually emerged."

In this paper I have tried to specify what lies behind these sentences.

Conclusion

The growth in the information science research at the school continued in the 80´ies and 90´ies and so did the research in library history, library administration and cultural studies.

The research of the 1970´ies in many ways laid the basis for the development of the Royal School leading to the legislative changes in 1985 and 1998, when the Royal School finally got status as a university institution with research, education of researchers and awarding of PhD and doctoral degrees.

Congratulating author:

Ole Harbo

Illustrations

Robert A. Fairthorne (1904-2000) one of the front figures of information science used the frog as his brand, and the illustrations are examples of his use of that brand.

1. A drawing from the SIRE conference 1977 showing the frog trying to grasp the tiger (information science).
2. His postcard thanking for the conference 1977
3. Calendar front page 1978 with his photo of a frog doll

Toward a General Theory of the Book

Srećko Jelušić

University of Zadar, Zadar, Croatia

In this article, we argue that the organization of the contemporary society is unthinkable without publishing. The very beginning and ending of each individual as well as of each state marks a PUBLICATION of birth or death, or in the case of a state by the declaration of its establishment, or its termination. The time that passes between these two extreme situations relies heavily on products of publishing industry. The entire education of an individual as well as the overall functioning of a state also relies on PUBLISHED legal documents. Just as we claim a research paper doesn't exist until it is published, a legal document becomes valid the moment it is published or shortly after the publication. By building and organizing national libraries and national archives societies expect to forever preserve their relevant books and documents[1] as a proof of their existence and as a basis for many future decisions and endeavors. Therefore one can conclude that societies that base their organization on written documents use publishing as a tool for governing the relationships in the society. While ol' Johannes Gutenberg (1398-1468) was printing first sheets in his printing press, he perhaps thought he would still be glorified in 21st century for his movable type invention and that his name will forever be recorded in the history of the book. In his mind he imagined computerized metal printing presses printing at the rate of 10 000 sheets an hour, machine sewn book blocks and machine produced hardcovers. He also knew the destiny of books will be such that some will be banned, and some maybe even burned! But all joking aside, Gutenberg hardly had such thoughts. He was more likely to have been concerned with printing higher volume of sheets, improving sales of his printed materials, and paying off his debts. To most craftsmen and artisans those were the primary concerns, however, not the only ones. Paying rent, suppliers, taxes, supporting the family, concern for their own health because one cannot produce if not fit, were the too often the concerns of craftsmen of the olden days as well as of today, so possibly Guttenberg's as well.

Where can one find a glimmer of inspiration in Guttenberg's situation – an idea that inspired his endeavor? If the fact that he was a goldsmith were true, then perhaps he used the collective knowledge of the masters from whom he learned. It is also possible there were some sort of a relation between, for example, a signet ring

[1] This term in this context refers to all eletronic publications, all web pages, all videos, all recordings etc., in general all documents that were ever publically available in any way.

and the movable type set in the printing tray as an arranged assortment of many signet rings. This may be, but even if so there is no significant inspiration, only experience, trial and error, and then again a trial. Guttenberg's invention came to life and started a technologic flywheel, results of which are still in use today. Since inventions happen within a certain context, it is realistic to assume that the movable type would have been invented regardless of Guttenberg. This would have happened at perhaps slightly later time, but probably at around the same period and somewhere in Europe. Given the high concentration of great minds in 15th century the inevitability of this invention is obvious; Aldus Manutius (1450-1515), Christopher Columbus (1451-1506), Leonardo da Vinci (1452-1519), to mention a few.

In contrast to Guttenberg, as we know from his written legacy Aldus Manutius, chose the printing industry to spread scientific knowledge. Between a career of a university professors and a printer, he chose the latter. Therefore, he was convinced that printing of books was essential for the development of science. He dedicated his scientific and business skills to confirming this hypothesis. He probably had neither the time nor the knowledge to employ scientometrics to determine if his hypothesis were valid or not. However, his hypothesis could at least partially be confirmed in the fact that he created a market, as he was selling the books he printed. This means that an audience of readers existed and that there was interest and need for books that were coming out of Aldus workshop. Printed books were not only coming out from his workshop. Many printing *oficine* sprouted in Venice and in other Italian cities and all over Netherlands, Germany and elsewhere. Thus, it could be said that Manutus's motivation was focused on the idea of enlightenment and, largely in scientific notion. There is *a reasonable scientific assumption* that Aldus Manutius printed books deliberately and with scientific intent. It has already been proven that he was the inventor of the style, format, equipment, editing, printing blocks – abundantly upgrading Guttenberg's discovery. For his print shop he ordered letters Aldina, still named after him today. He used the experience of his predecessors in the best possible way and satisfied the growing need of the society. There we can already see the interaction of his books and the public. Theorists speak of the interaction as an important element of social cohesion. A network of printers and book distributors emerged in Europe and it provided the intensive interaction between the text and the reader.

Above listed concentration of great minds represents only a fraction of those who defined the spiritual history of modern Europe, and who were directly linked to writing, science and publishing. Naturally, it took some time for printing to spread out, however, neither computers nor the Internet caught on in few days, and as a matter of fact they are spreading out still.

We must not skip over, nor forget that at this time in catholic countries inquisition emerges, an organized force of the society that instantly noticed ambivalence of technology, in this case, the printing. Printing technology spread the doctrine of the Church, but at the same rate it spread the Reformation as well[2]. When we talk about technology, and printing is a technology, we talk about is ambivalence. Today's theorists of critical theory speak of ambivalence as one of the basic features of technology. So what does this have to do with inquisition? Printing technology produced far greater number of copies of a given title than was ever possible in scriptoriums, therefore, books could reach greater number of readers. Handwritten books were under the control of the Church, but print undoubtedly liberated the text form such control. Technology of print broke the ecclesiastical framework. Books escaped into a new public, however, the Church introduced lists of banned books as a new mechanism of control. So here we have a formal proof books influence peoples' thoughts, and consequently their actions. And one more thing! When we talked about Guttenberg and Manutius we mentioned inventions and technology. The industry we are discussing developed along with technology and it would have not existed without technology. Going back to the system theorists, we know that they are talking about relationship between technosphere and sociosphere. Considering their interpretation of the ambivalence of technology, they are predicting human destiny to get in to the hands of those who are controlling the technology development. Hofkirchner and Castells believe our future depends on our ability to overcome imposed belief that technology and its consequences cannot be managed. Thanks to the significant relationship between humans and technology (such as between Johann Gutenberg and his presses, and Aldus Manutius and his oficina) structured publishing industry developed thoughout centuries producing printed content holders that were the building bloks of Western civilization. We said Guttenberg was driven only by desire, but Manutious already had a concept, an intention to reach a certain intellectual goal. Perhaps we are unfair toward Guttenberg, but this is only so we can demonstrate a certain graduality in profession development. Today we are also witnesses to graduality that is obvious in development and spreading of computers. Despite of this and the fact that publishing is in fact the infrastructure for most other professions, theoretical aspects of publishing never arouse much attention from scholars.

> "There is a large scholarly literature on the history of publishing and the history of the book, from Gutenberg through to the nineteenth and twentieth centuries, and there are many works which recount the history of particular firms. But in the sphere of scholarly research, the contemporary world of book publishing has, by and large, been ignored." [3]

[2] See more on this in: Chaunu, P. Le temps des Reformes, Librairie Artheme Fayard, 1975.

[3] Thompson, B.J. Books in the Digital Age. Cambridge : Polity, 2005, pg. 3.

Thompson wished to fulfill this void and, relying on the Pierre Bourdieu's work, he developed a theory on publishing fields starting with description of structure of publishing and explaining that certain publishing fields are as diverse as different athletic disciplines. In one discipline a person can be the top athlete, while in another that same person can be a complete loser. Thompson's central concept in describing the publishing fields is *field logic*, a result of different relations that shape participants' and organizations' behavior. Thompson notes that publishing fields one cannot observe out of the context of other social fields, and he gives an example of scientific publishing that cannot be considered outside of the context of academia and research. However, apart from the context approach, in his book he points out that along with the specifics of each field there are four groups of developmental tendencies that affect the advancement of publishing industry: centralization of finances, market structure changes, market and publishing companies globalization, and the impact of new technologies. In effort to find a relationship between publishing and the way society functions as a whole Thompson writes:

> "And it is not just the academy that has come to rely in countless ways on an industry about which it knows very little: the output of academic publishers also makes a vital contribution to the broader sphere of public discussion and debate. We should not underestimate the importance of books in helping to cultavete the kind of critical and informed culture that is essential to what we could call a vibrant public sphere".[4]

Here Thompson critically points to Habermas who, in his book The Structural Transformation of Public Sphere: an Inquiry into a Category of Bourgeois Society, writes on influence of publishing onto the public sphere, but only mentions newspapers as main factors of restructuring of the public sphere in bourgeois society.

> "The *bourgeois public sphere* can be understood as a sphere of private persons assembled to form a public. They soon began to make use of the public sphere of informational newspapaers, which was officiallly regulated, against the public power itself, using those papers, along with the morally and critically oriented weeklies, to engage in debate about the general rules governing relations in their own essentially privetized but publčicly relevant sphere of commodity exchange and labor." [5]

[4] Ibid pg. 11.

[5] Johns, A. Book of Nature and the Nature of the Book. //The Book History Reader, Finkelstein, D, McCleery, A. (ed.). London : Routledge, 2003, pg.69.

In his text The Public Sphere Habermas writes about the transformation of the private sphere into the public one, describing the role of newspapers but not taking into account the role of the books, although he is saying this transformation begun in late middle ages when newspapers did not exist. This is what Thompson considers an oversight. As much as Thompson tries to explain the mechanisms that drive the publishing industry, his focus is primarily on its development and not on the theory of publishing. He focuses on the future of publishing and its role in the society. Thompson is concerned with the development of the society and poses a question, what will the future society be like if the publishing were to go trough radical changes.

Another author, Adrian Johns, in his book Book of Nature and the Nature of the Book discusses Elizabeth Einstein's thesis on print as a method of fixating a text, and writes:

> "The sources of print culture are therefore to be sought in civility as much as in technology, and in historical labors as much as in immediate cause and effects. The *printing revolution* if there was one, consisted in changes in the conventions of handling and investing credit in textual materials, as much as in transformations in their manufacture. The point deserves to be stressed explicitly. I do not question that print enabled the stabilization of texts, to some extent; although fixity was far rarer and harder to discern in early modern Europe than most modern historians assume. I do, however, question the character of the link between the two. Printed texts were not intrinsically trustworthy. When they were in fact trusted, it was only as a result of hard work. (...) At no point could it be counted on to reside irremisibly in the object itself, and it was always liable to contradictions. Those faced with using the press to create and sustain knowledge thus found themselves confronting a culture characterized by nothing so much as indeterminacy. If printing held no necessary bond to truth, neither did it show a necessary bond to falsity or corruption. Each link remined vulnerable to dispute. It is this epistemic indeterminacy that lends the history of the book its powerful impact on cultural history." [6]

The true mediating role of a publisher is revealed here. This role distinguished itself through the centuries and became the guarantee of quality and authenticity of a manuscript, even though at the beginning of printing there were cases of adulterating of even strictly controlled texts such as the Bible. The same author concludes:

[6] Johns, A. Book of Nature and the Nature of the Book. // The Books History Reader. Finkelstein, D, McCleery, A. edts. London and New York : Routledge, 2002, str. 68.

> "Textual corruption of even such closely monitored texts as the Bible actually increased with the advent of print, due to various combinations of piracy and careless printing. Like print itself, piracy therefore had *epistemic* as well as *economic* implications: it affected the structure and content of knowledge."[7]

We could go from a book to a book and in each of them we would find a part of the story about publishing and book making. We could try and measure which product influences public opinion more (or less): newspapers or books, or television, or the Internet. We believe the answer is not in the attempt to explain only publishing practice. The answer to this question in our opinion should be sought in the context of the information society, system theory and media theory. Historically, a book was an absolutely dominating medium, unlike today. To be nostalgic about this fact would be like being nostalgic for the sail ships that were the sole means of sea transportation for the long period of our history.

We mentioned the information society. Naturally, we find the word *information* in numerous phrases. Its usage is somewhat random, as in information society or information specialist, and so on. However, there is no doubt information is the foundation of education and the invention of print made it possible for a large number of people to get information. This in turn created a greater need for literacy.[8] It took a long time for education and literacy to become everyone's, literally everyone's right, regardless of social status, heritage and gender. Although form today's point of view the right to education does not seem like a special social privilege, it took centuries to gain this right. Along with literacy and education, naturally came the right to participate in decision making and participation in public affairs. Therefore literacy, education and the right to vote are in a indissoluble relationship and they make up the foundations of today's society regardless of what we may call it: democratic society, information society, society of knowledge or by some different name. D. Bell used that different name– post-industrial society. When Daniel Bell's [9]book appeared in late seventies of the last century it became apparent that the industrial era is ending. Upon conclusion of the industrial era a new page turns in our history. What Bell announced Pierre Levy explained about twenty years later:

> "There was a fusion of telecommunications, information technology, journalism, publishing, television, film industry and computer games under the wing

[7] See book Dearnley, J., Feather, J. The Wired World, an introduction to the theory and practice of the information society. London : Library Association Publishing, 2001.

[8] Bell, D. The coming of a post-industrial society: a wenture in social forecasting, Heinemann, 1974.

[9] Lévy,P. L'intelligence collective, Pour une antropologie du cyberspace. Paris : Éditions La Découverte et Syros, 1997, pg. 9.

> of multimedia and computer industry, just as the journalists predicted. That is not all, nor the most important. Setting aside certain commercial repercussions, we believe it is necessary to reveal what civilization compromises were made with the emergence of multimedia: new communication systems connections, new mechanisms of regulation and collaboration, unedited manuscripts, unusual intellectual behavior, a change in relations to time and space."[10]

Lévy subscribes to the theory that making political decisions that regulate technology can affect the development of *collective intelligence*, which may affect the World's infrastructure in the future. He believes that there is an urgent need to establish *collective intelligence* because "bureaucratic hierarchies (based on stable text), media monarchies (supported by television and mass media), international monetary networks (using telephone and real time technologies) neither mobilize nor coordinate intelligence, experiences, wisdom and imagination of human beings."[11]

The author explains what is *collective intelligence* and in the rest of the book he goes into this theory in detail:

> "*It is intelligence that is everywhere, that is questioned constantly, coordinated in real time and that contributes to the actual development of competencies.* To this definition we must add: basis and goal of collective intelligence is mutual recognition and enrichment of people, and not glorifying the cult of fetishized communities."[12]

Christian Huitema joins to Pierre Lévy's theories on programming technologies that will as a result have our programmed future. One can assume that the power of technology has not been unleashed until the invention of computers as the reach of technology was not global nor was there a technology that served as infrastructure to all other technologies. Those of us who are apprehensive about the theory which states it is not possible to control new technologies should read Huitema's description of his first involvement in the IAB (Internet Activities Board) meetings in Los Angeles in1991. IAB was a team of hand picked specialists who **supervised the development of the Web** since the eighties of the last century. Huitema participated in the meetings with the pioneers of Arpanet, a research network founded in 1969 by the United States Ministry of Defense where "packet" data exchange technology was invented thus enabling networking of the computers. The meetings were about further development of the Web. He mentions Vint Cerf who in 1973 with Bob Kahn invented the very concept of the

[10] Ibid pg.12.

[11] Ibid pg. 29.

[12] Comp. Huitema, C. Et Dieu créa l'Internet. 5. edition. Paris : Eyrolles, 1999, pg. 1 and further.

Internet, Dave Clark the architect of the Internet, John Postel who managed the web addresses and others. With the strong support of the United States military establishment, these experts while they programmed the network actually programmed a large part of today's information and communication environment.[13]

Along with the scholars we mentioned thus far and today's advocates of the critical theory of self -sustainability stands Niklas Luhmann[14] who, following Ludvig von Bertalanffi's footsteps develops a general systems theory. Wolfgang Hofkirchner goes even a step further. Conclusions based on his discussions on *sociosphere*, *technosphere* and on ambivalence of technology, are applicable to the theory on a place of a book in the context of multimedia today. According to Luhmann[15], a system has the ability to self - regulate, meaning it contains internal dynamics that allow it to always establish its balance. Since we named this chapter an outline of a theory, in place of the concluding remarks we will list the topics we believe should be studied in context of the critical theory of the system, with hopes that perhaps in the future a theory of book will be developed.

Complexity.

In the world of publishing, it became apparent there are no more simple solutions. Printed book took primacy over manuscript, but other competing media appeared. Open access initiative emerged and with it new relations needing regulation. There are numerous stakeholders, more demanding markets, increasing demand for reading skills, life long learning, in short nothing is as it was and everything is a subject to constant change.

Convergence of media and codependency.

Text is still a starting point for all media activity. Starting point of a movie is a script; starting point of an opera is a libretto. In order to make multimedia content it is necessary first to WRITE the text and to determine on which media it should be presented in order for it to merge into a meaningful multimedia entity.

[13] Luhmann, N. Soziale Systeme-Grundriss einer allgemeine Theorie. Frankfurt : Suhrkamp, 1981.

[14] I wrote more in depth about this topic in the book titled Structure and organization of library systems. Zagreb: Filozofski fakultet, 1992.
Luhmann's general theory of systems features the following elements:
- Autoregulation and achieving inernal correlations
- Surrounding
- Differentiation
- Evolution
- Communication

Information politics.

Rulers have always made laws. Some of the laws referred to publishing and broadcasting of decisions, new laws and rules. Today, activities related to communication between the governments and the public are governed by regulations which we refer to, in a strict sense, as the information policy. In the broader sense, the information policies regulate the ICT sector not only on national but also on international level.

Relationship between social sphere and techno sphere

From the invention of lever and stirrups to the invention of computers people are in the constant relation to technology. We believe that Fuchs's and Hofkirchner's contemporary research is going in the direction of explaining the complexities of the relation between the human and the machine, especially in the sphere of idea generation, and that the starting point for the research should be humanism, not technology.

Transdisciplinarity

When it comes to reproduction and dissemination of content the relationship between a human and a machine surpasses all known disciplines. These issues should be studied across all disciplines keeping the focus on the man and his destiny in the universe.

Shift from the industry

The issue should be observed with detachment from the problems the industry is facing, but in the context of the industry and with the attempt to benefit the industry with this research.

Finally, it is necessary to determine the tone further research is to carry. It seems that a man would not be a man if he were not driven by optimism. This is why we will end this chapter with an optimistic view from a scientific authority, and this should be taken as a guiding idea for all further research:

> "The dream of the Enlightenment, that reason and science would solve theproblems of humankind, is within reach. Yet there is an extraordinary gapbetween our technological overdevelopment and our social underdevelopment. Our economy, society, and culture are built on interests, values, institutions and systems of representation that, by and large, limit collective creativity, confiscate the harvest of information technology, and deviate our energy into self-destructive confrontation... If people are informed, active, and commu-

> nicate throughout the world; if business assumes its social responsibility; if the media become the messengers, rather than the message; if political actors react against cynicism, and restore belief in democracy; if culture is reconstructed from experience; if humankind feels the solidarity of the species throughout the globe; if we assert intergenerational solidarity by living in harmony with nature; if we depart from the exploration of our inner self, having made peace amongst ourselves. If all this is made possible by our informed, conscious, shared decision, while there is still time, maybe then, we may, at last, be able to live and let live, love and be loved."[15]

To what extent are these ideas achievable is for each of us to decide.

Translated by Mirta Marošić

Address of congratulating author:

Srećko Jelušić
University of Zadar, 23000 Zadar
Croatia

[15] Castells, M. *End of millennium—The information age: Economy, society and culture*. Vol. 3. Malden, MA : Blackwell, 1998, str. 390.

Revisiting the Concept of the Political Library in the World of Social Network Media

Leif Kajberg

Bagsværd, Denmark

Abstract. In these times, public libraries in many countries have increasingly come under pressure from developments within the information landscape. Thus, not least because of the massive digitization of information resources, the proliferation and popularity of search engines, in particular Google, and the booming technologies of Web 2.0 public libraries find themselves in a very complex situation. After all there seems to be a need for public libraries to reorient their aims and objectives and to redefine their service identity. At the same time search engines, and especially Google, are increasingly becoming under scrutiny. In discussing the survival of public libraries and devising an updated role for libraries in the age of Google and social media, attention should be given to fleshing out a new vision for the public library as a provider of alternative information and as an institution supporting information democracy. Hence, attention is given to public libraries as democratic spaces and the role public libraries can play in hosting and organising electronic discussion forums on, among other things, current political issues. Also, the concept of the political library is revisited. In exploring the opportunities of public libraries as spaces of e-discussion, a couple of Danish projects concerned with involvement of citizens in political and community-related e-discussions are briefly mentioned including a project pursued by the Aarhus Municipal Libraries (Denmark). The Aarhus project entitled "Demokrateket" considers proactive mediation of community information and the creation of physical virtual fora allowing citizens to shape the political agenda.

1 Introduction

Peter Ingwersen's groundbreaking and influential research efforts in the areas of information seeking and information retrieval – initially centering on users' search patterns in consulting card catalogues and mapping and exploring characteristics of reference librarians' search procedures in handling technically-related user enquiries in public library environments (the JIT Project) – marked a shift of emphasis and a new direction of research in our field. In fact, Peter Ingwersen's pioneering research contribution represented a refreshing trend towards more evidence-based R & D in a decade – the 1970s – reflecting a clearly "political" atmosphere characterised as these years were by political and social priorities and agendas within the library profession and discipline. During the subsequent years, empirical and analytical research in the information field expanded and the social and political tone became less outspoken and finally it almost disappeared. Now, however, new winds are blowing, maybe, and the following piece should be seen as an attempt to sketch a context and a range of issues that sort of justify rethinking of public library purpose and the revival of the concept of "the political library".

Digitization of information combined with increasing growth of electronic networks has created new opportunities for providing information resources and services for citizens. New forms of and channels for distributing information and documents within Internet, new tools and opportunities for digitizing our written cultural heritage and making it accessible, new mechanisms for discovering and accessing information and new services and networking forums such as Facebook, Flickr, MySpace, Twitter, YouTube and social tagging provide an opportunity and challenge to the kind of services the public library produces and to the societal role and institutional identity it assumes for itself. Activities and services facilitated by the Internet are increasingly used by various citizen groups. Especially younger generations have embraced the new forms of electronic interaction and adopted Internet as their own media. Thus, harnessing the benefits and challenges of Web 2.0 remains a major challenge to public libraries today.

2 Methodology

The study presented here aims to analyse the ways in which public libraries can strengthen their survival capacity by drawing upon the new Web 2.0 technologies available and develop new roles. An analysis is conducted of selected writings covering such key notions as social software applications, collective intelligence and digital socialism. Also revisited is the dated concept of *the political library*. Based on observations emerging from the analysis, a revised role is outlined for public libraries in the era of digital information and Web 2.0 with a special focus on information democracy, serving as a democratic agora and maintaining the function as a neutral information provider in a Google-dominated commoditized information world. Thus, part of the analysis consists in shedding light on the nature, viability and conditions and opportunities of information democracy within the framework of today's social networking media.

3 Literature Review

The theoretical framework provided for the present study draws on inspiration from Doctor's piece on justice and social equity in cyberspace (Doctor, 1994). This article was published in the early days of the Internet characterised as they were by enthusiasm, euphoria and a fascination of the promising new potentials and possibilities represented by the new global medium and utility. Revisiting the somewhat idealistic ideas, notions, conceptions and projections that arise in the first, pioneering and booming years of the Internet from a contemporary information democracy perspective is one of the objectives of the analysis reported here. Characteristic to the pioneering years of Internet and "the Information Super Highway" is the fairly optimistic and in some respects even idealistic tone. Today, things are more complex and we are seeing the commercialisation of the Internet, "abuse" of the Internet (hacking, theft of money and identities and other types of crime) along with the "hedonistic" take-over (e.g. through the spread of porn) so eminently depicted by Keen (2008).

Current professional literature on the implications of Web 2.0 technologies for libraries and their service provision tends to emphasize the new social software tools and media as information assets to be integrated in existing service offerings. The Web 2.0 social media are typically seen as opportunities and means for supplementing, enhancing and enriching the existing mix of library-related services and facilities. Briefly, Google and interactive technologies such as wikis and blogs are considered new devices in the library service provision toolbox.

However, there are signs that a more critical awareness of Web 2.0 phenomena is beginning to gain ground. Brabazon (2006) has some serious reservations about the whole ideology behind and the peer production practices of Wikipedia and she is very concerned about what Google does to students in pursuing projects and assignments. In a very thoughtful piece, Waller (2009) takes a close look at the relations between Google and public libraries and explores similarities and differences. On the surface of it, Google seems to pursue goals and offer services and products that are parallel to or overlap the kinds of searching assistance and information provision that are core activities in libraries, but in the end the two players in the information arena deviate markedly from each other. The author demonstrates that the *conceptions of information* adhered to by (1) Google as a commercial firm and (2) public libraries as providers of balanced and consolidated information are fundamentally different. The commercial firm and the public agency simply want to do different things. Waller's reflections on the democracy-underpinning role of public libraries in maintaining a balanced and non-commercial information provision are very central to the observations on a redefined role for public libraries in the present paper.

A decidedly pessimistic view of Web 2.0 and interactive social media can be found in Keen's book *The Cult of the Amateur* (2008). The book embodies a frontal attack on what the author sees as the frightening regime of amateurs and a pervasive culture of narcissism resulting from the Web 2.0 revolution. Keen provides a coherent and very critical perspective on the web 2.0 tools and phenomena and demonstrates their manipulating potentials and how they make expert knowledge and expert performance erode and gradually bring about de-professionalization in some respects. Professionals have been replaced by noble amateurs. Keen explores the seamy side of blogs and blogging and addresses the problem of tricksters and fraudulent behavior. He provides examples of dubious editorial practices characterizing Wikipedia and the mediocrity of content provided by contributors. Above all he laments on the downgrading and dismissal of experts and the devaluation of expert knowledge. Keen ends up with a very pessimistic state-of-the-art description and scenario in which he identifies Orwell-like tendencies and points out that the American society is moving into an age of total digital surveillance. Sounding a bit like an old moralizing culture critic, Keen draws attention to a range of critical and pertinent issues affecting all web users.

The published literature on public libraries is very sparse on the implications of Web 2.0 and social networking for the community involvement of public libraries. Actually, very few contributions address the role of libraries in maintaining freedom of information in the Digital Age along with their supportive role in relation to campaigning initiatives, local grassroots activities, the organisation of political debates as well as the provision of alternative, anti-mainstream and anti-elitist information, etc.

In contrast, library literature, especially that part of it, which covers 20th century developments in libraries and librarianship in Australia, UK and USA, provides considerable coverage of the role of libraries in promoting and consolidating democracy. For instance, Waller (2009, p:6) refers to what she calls the "grand tradition" of public libraries in the 1950s with Lionel McColvin, UK as one of the leading figures. According to McColvin public libraries would have a leading role in advancing democracy, in knowledge building and the spread of knowledge and in empowering citizens through the possibility for self-education. However, recent library literature also includes items that focus on libraries and democracy and the societal role of libraries. In his monograph on *Civic Librarianship* McCabe explores the concept of civic librarianship and develops a vision for the mission and purpose of the public library. Civic librarianship differs markedly from the libertarian public library, but it is also very different from the public library of the traditional type, which has often fallen short in fleshing out its basic mission into effectual and tangible strategies for action. McCabe (2001, pp:78-79) sees a broadened role for public libraries and identifies a number of areas, where strategic action is needed:

- Restore the confidence of public librarians and trustees in exercising social authority.
- Renew the public library's historical mission of education for a democratic society.
- Develop the public library as a centre of the community.
- Develop strategies to build communities through public library service.
- Use services and collections to meet social as well as individual needs.
- Strengthen the political efforts of public librarians and trustees.

As can be seen, the suggestions for reforming public libraries in line with the conceptual framework of civic librarianship are of a more general nature and since the book appeared in 2001 there is no treatment of the challenges of e.g. social networking technologies and the way people communicate and organize information-related activities *outside* the library context after the advent of the social web revolution.

Clearly, civic librarianship is meant as an effort to update and expand the role of the public library while keeping the library's historic mission of education for a democratic society. The author's insistence on civic dialogue and social interaction is also of relevance when discussing and defining the role of the public library in times of web 2.0.

Kranich (2001, pp:83-95) explains how libraries help reduce the digital divide, increase access to government information and are fighting against both censorship and private interests to ensure that access to information is as free as possible. The library as civic space creates opportunities for community and dialogue, which she thinks is a very important democratic function as a supplement to information-related and education-centred tasks. In their joint article Canadian library researchers Alstad and Curry (2003) describe how squares and other public spaces are increasingly replaced by company-owned areas such as shopping malls, where people can no longer act as citizens, but are primarily consumers. In order that libraries are to support democracy and serve as public space they should, among other things, change their objectives so that they provide for libraries moving towards a more proactive stance thus making room for lectures and discussion groups. A Danish perspective is provided by Skot-Hansen and Andersson (1994) who carried out a study of libraries as a resource in the local community. As pointed out in the study, for a library to serve as a local driver it should relate actively to the community it belongs to and sharpen its profile in interaction with other institutions, associations and groups. The libraries' social function is also examined in a British study conducted by Matarasso (1998). He concludes that libraries have a great potential to contribute to the development of the local community. In a contribution in the anthology entitled *Libraries and Democracy: the Corner Stone of Liberty* Durrance and others (2001, pp:49-59) explore several American library projects that address web-based community information, which are considered to help strengthen civil society. The libraries' own websites can be used successfully, for example in providing guidance to citizens in pointing to web-based government information resources and be targeted to various minority groups. Also, American libraries have often been a leader in or been active as a partner in the development of virtual local area networks, so-called community networks.

4 Collective Intelligence

The phrase "collective intelligence" has been used for decades and it has become increasingly popular and more important with the advent of new communications technologies. A book by Pierre Lévy (1997) about the computerization of society from a social-theoretical standpoint represents an early approach to *Collective Intelligence*. Overall, there is a catching drive in the book and it is very visionary, but there are obscure passages and incomprehensible observations as well. In the chapter on "the Dynamics of Intelligent Cities" the author explains the idea of a direct, computer-mediated democracy – *a virtual agora* – and he anticipates the application of web 2.0 tools for discussions and decision-making. The author adds that the introduction of what he calls a *real-time mechanism for direct democracy* would facilitate a democratic dialogue.

> "Within the framework of collective intelligence, real-time democracy is the absolute antithesis of the demagoguery of live action broadcasts and the immediacy of crowd behaviour" (Lévy, 1997, p:77).

One reservation that might be voiced here is the reaction or response from those representing the establishment and the political scene: are politicians and those in power today really interested in this kind of direct democracy?

5 The Political Library: Revival of a Concept?

In her thesis on the Political Library with the subtitle "Public Library as a space for citizens' participation and public discourse" Jadinge (2004) discusses the potential public libraries have for actively supporting civic participation and public /discourse. The study seeks to explore the origin of the idea of the political library in a Swedish public library context in the mid-1970s. The author observes that the political library deserves to be taken out of oblivion of mainly two reasons. First, it is an idea that is quite radical (in the general sense of the word!) by today's standards, and it should therefore serve as fuel for a renewed discussion of library ideology and democracy issues, in the field practice as well as in research. The concept of a political library is interesting because it affects some fundamental aspects of library and information activities, such as the neutrality/objectivity issue and the relationship that libraries have to civil society. Secondly, it is relevant to offer a historical perspective to today's library debate. The author's view is that undertaking a comparison between the context of the 1970s and the situation and conditions of the 2000s can be fruitful. As is the case today, democracy problems were frequently and sometimes heavily discussed in the 1970s, but the atmosphere and context was different and attention was focused on how the political library should act so as to maintain the library's neutrality. To be neutral may nevertheless often involve some sense of commitment.

The results of the Swedish study prompt further analysis of the notion of the political library, its relevance today along with its potential for renewing the role of a public library in transition. Today, appraising the generalizability and pertinence of the political library and giving the concept a needed brush-up implies an awareness of the opportunities of web 2.0 tools and applications.

6 Access to Alternative Information

Given that public libraries take the function as provider of alternative, non-elitist and non-mainstream information seriously, there are many situations where the active involvement and service provision of libraries would be relevant and desirable. Illustrative examples are the campaigns and debate sessions preceding elections, referendums, etc. Typically, and this observation could be generalized to many countries, the official information presented to the electorate is biased. Thus, for instance, in Denmark the many referendums relating to Denmark's entry into the Common Market and the EU as well as Denmark's accession to the EU treaties, etc. constitute an illustrative example: there is unequal access to information and lack of funds for distributing alternative information. Frequently, there is a marked lack of alternative information resources reflecting positions other than those held by the establishment and those possessing the political power and the money. There is a need for information that provides alternatives to and challenges the official and dominating messages and viewpoints. The new social network media have partly remedied this situation, but libraries could still play a role here.

7 Facebook as an Information Tool for Local Protest Actions: a Danish Example

In Denmark the controversy over and the fight for the survival of a local railway in a thinly populated area provides an illustrative example of the involvement or lack of involvement of the local public library in a much discussed local matter. For the time being the Western railway, a local railway line in the Western part of Denmark, is in risk of being closed down in that a majority of Regional Council Members want to eliminate the line because it is considered loss-making; it is argued that it is too expensive in terms of operational and maintenance costs and the case is made that buses are a better solution. The prospect of a rural railway line ceasing to exist because of a Regional Council decision evoked strong protests from parts of the local population, created a heated debate and led to the formation of railway protection initiatives. Also, a group on Facebook named "Save the Western Railway" was set up. However, the local library has adopted a fairly passive role in relation to the railway issue. No meetings have been hosted by the library and the only activity organized by the library is the setting up of an exhibition featuring the railway and its history. The Western Railway protection citizens' initiative represents an interesting case illustrating how Facebook is relied on by politically articulated individuals and groups. There are ten thousands of examples of this nature on Facebook. These grassroots activities, campaigns, protest groups and unofficial networks confronting decision-makers and those in power provide examples of how initiatives are born, strategies are developed, individuals get involved and become members of groups, how communication takes place, how various types of information and views are presented and exchanged and how decisions are made, etc. Also illustrated are the exchange of information, views, advice and know-how between various bodies of expertise and those who maintain grassroots initiatives. And last not least: studies of the emergence of grassroots initiatives in a Facebook context – or as they develop within other social networking media – could be designed so as to explore the ways in which libraries respond to, support or ignore groups and initiatives arising and developing within the social networking media.

There are various ways in which public libraries could adopt a more proactive role in relation to Web 2.0 and citizens' campaigns and initiatives. Thus, a Danish project, outlined on the web pages of the Librarians' Union, addresses the role of the public library as a moderator of current political debates, etc. going on in the local community. The library is supposed to provide balanced subject-specific input for discussions progressing in social network media of the Web 2.0 type. You can have people debating current and crucial topics and issues on the Web. But the prerequisite is that you prepare solid background information and that you dare bring up controversies, tender subject and sensitive issues for discussion. Also, you should be ready to go for interaction with other media. On the whole, libraries could adopt a more active democratic role.

8 Libraries as Democratic Agoras

Thus, as illustration, in the Danish Municipality of Odder it has for several years been natural for citizens and politicians to engage in discussions on a variety of issues using web based discussion forums. Last year's municipal elections provided another example of the electronic communication between citizens and local politicians in that more than 400 comments were posted as part of a lively debate between citizens and those standing as candidates for the town council. One of the reasons for the recorded success in raising and maintaining e-debates is that those responsible for hosting and maintaining the debate invest quite a lot of effort in furnishing people with background knowledge on a specific topic or issue. For instance, all town council decisions are described in a journalistic mode on the commune homepage. In addition, video transmissions of sequences selected from, among other things, town council meetings and local civic meetings on key issues are available. It is crucial to bring up tender subjects and sensitive issues for discussion. If you dare not put something on the line and raise a controversy in areas and issues people are very eager about they tend to drop out and ignore debates.

Unfortunately, most local authorities and councillors tend to avoid conflicts and shrink from raising sensitive subjects. Thus, it is obvious for and the initiative rests with the libraries when it comes to providing local residents with opportunities for making themselves heard in public life and as part of a functioning democracy. But public libraries could be instrumental in, or take a role in creating an active democratic communication in matters and issues that are of concern to citizens. However, a task like this cannot be reduced to acquiring and having district plans ready for examination or distributing election campaigning material (flyers, brochures, etc.). It is much more than that. Libraries must dare to act as initiators and those taking the lead. What must not be forgotten in this respect is the interaction with other media. Consideration should be given to involving several target groups and communities. In the context of the 2009 Municipal Election, video-based profiles and portrayals of the candidates for municipal election were made available. At the same time a group was set up on Facebook in the hope that in relying on this vehicle, there would be better possibilities for appealing to and attracting the interest of younger target audiences.

Digital debate is not better than analogous debate, and you cannot say that it is better to discuss on the web than relying on conventional discussion pieces and letters in newspapers or exchanging questions and views at civic or election meetings. But e-debates facilitated by forums such as the Odder Net in the time before and in the run-up to the municipal election could be instrumental in making citizens making an informed decision when casting their votes. At the same time it is noted that quite a few citizens express themselves only on the Web. Obviously, a certain amount of resources are required for setting up an adequate framework for a debate. Thus, the role of the library/library is primarily that of a mediator.

Considerably broader in scope is a draft development project presented by the Aarhus Municipal Libraries and entitled *demokrateket*. The vision underlying the concept of demokrateket is to vitalize societal and community-related challenges to citizens and to create physical and virtual fora that allow citizens to be involved in shaping the political agenda. Thus, demokrateket is intended to develop innovative approaches to the library's communication and mediation of community information as a proactive and interactive activity, which should include users and political players in the physical library environment along with web pages and social and mobile fora. The final project will be unique in that it envisages a shift of the library's efforts provided through a democratic (physical and virtual) space from a reactive and communication-centred role towards a proactive, front-edge and staging role. In taking on its new role, the library should establish and facilitate interactive, independent and direct channels of communication between citizens and their political representatives. In doing so, the library should support free opinion building and active citizenship. The library staff's competences in terms of serving as trendspotters identifying social and political issues and performing the function as moderator of debate-prompting and democratic processes become of central importance in implementing the demokrateket.

Related to the Aarhus project is a previous project undertaken by the public library in Frederikshavn and supported by a grant from the Danish Agency for Libraries and Media. The project, which is completed now, is entitled "The Library as a Democratic Agora" and has as one of its objectives to explore the role of the public library as a "third place (space)" and as one of the cornerstones in Danish democracy. In examining and developing this role, which includes facilitating democratic discourse, a challenging and slightly provocative approach would be adopted. Critical analysis of the findings of the Aarhus and Frederikshavn projects and output from similar democratic discourse projects conducted in library contexts is essential in defining a new role for the public library.

A Danish report on the future role and services of public libraries in the knowledge society appeared in 2010. The report is structured in five parts under the following main headings: Open libraries, Inspiration and learning, The Danes' Digital Library, Partnerships and Professional development. Unfortunately, the report is almost silent on the role of public libraries in democratic processes, in enhancing participatory democracy and in the establishment and monitoring of discussion fora. Occurrences of the term "debate fora" *can* be found and partnerships within the framework of civil society are touched upon as well, but there are very few concrete examples of partnerships representing the civil society and there is no mention whatsoever of groups of citizens committed to a specific issue, associations, grassroots initiatives, political groups, political parties, NGOs, etc.

9 Concluding Observations

For quite a few years basic public library roles and tasks tended to include such service areas as provision of books and other materials, information services, reference work. supporting learning activities, organizing cultural activities and promotion of reading. However, during recent years in some countries efforts have been made to redefine public library purpose – the mission of public libraries – with a view to the role of supporting political debates, campaigns, citizens engaging in social and grassroots issues, "activism", etc. But assuming a sharper role in relating to and supporting citizens' political and community-related activities is not a new phenomenon. Actually – as shown by an illustrative case from the Swedish public library history summarized elsewhere in this paper – in some countries there has been a tradition that public libraries committed themselves to making information resources available in connection with community action and citizens' group-based initiatives of various kinds. And by hosting discussions and meetings. In this context it is worth referring to the UNESCO Public Library Manifesto, which indicates the participation of citizens in civic life as an overall aim of public libraries.

The findings and reflections embodied in the Swedish study of the political library and the results of McCabe's analysis of the concept of civic librarianship provide good starting points for further analytic work. In defining an appropriate role for the public library in the Age of Web 2.0, there is a need for re-examining and partly reviving thoughts and ideas on how libraries could support grassroots initiatives and alternative political viewpoints and analyses. Hence, libraries and librarians need to discuss and clarify their stance towards key issues such as participatory democracy, political participation, empowerment and emancipatory roles.

One can imagine that the libraries are keen not to become completely left behind now that e-democracy is taking root in many contexts and environments. Here, the libraries' role can be – as an extension of efforts geared to reduce the digital divide – to provide part of the community dialogue that is undertaken in municipal websites as "real-life" physical sessions (by organizing such activities as politicians' cafés and the like). Still many people do not use or have access to computers and the Internet, and clearly this situation somewhat limits the suitability, performance and impact of Internet-driven social media as a tool of democracy.

The very interesting issue here is: can the public library redefine its mission? The analytic review of selected readings on Web 2.0 and social media, collective intelligence, digital socialism and the political library has generated some ideas and clues that might be of relevance to the discussion on a changed role for the public library. But can the public library be transformed into an agency that capitalizes on the social media and their innovative applications in supporting democracy, citizen participation in community development and political processes, multiculturalism, etc.? To shed light on this issue more explorative efforts are needed. Thus, in carrying on with the analysis of an updated role for public libraries, it seems obvious to conduct an empirical study that might be approached as interview-based analysis. For example, a study could be designed that aims to identify selected librarians' views of public library roles in the light of Web 2.0.

In discussing new roles for the public library, there are classic library virtues that should be safeguarded including the library's position as a recognized and trusted repository of information and public knowledge. In the times of booming web technologies and social media and commercialization of information and knowledge there is a need for an agency of neutrality and credibility that sort of helps users find out and unmask the increasing amount of bias, distortion, fraud, misuse, cheating and manipulation within the fancy new world of new web-based media and assist them in navigating in today's information universe, which may be less smooth than imagined. A new user educational perspective would certainly be relevant here.

In analysing the conditions and opportunities for information democracy in the sense of Web 2.0, explorative studies are needed to map politically-related information universes, information transfer and information use. The Digital Age with its new social media invites political engagement, but the era of digitization is also an age of despotic political leader styles, persistent and irremovable power structures, spin doctor-driven politics and infotainment. At least these features seem part of the reality in many countries. Power structures are opaque and various sorts of extra-parliamentary opposition groups, NGOs and grassroots initiatives in specific areas face barriers and difficulties in having their message heard. As is well known, because of failures, backlashes and disappointed expectations situations arise that eventually lead to frustration and apathy. The more than meagre results of COP 15 in Copenhagen on the risks and dangers of climate change and global warming come to mind in this respect: about 100,000 committed people walking their planned route in the streets carrying banners and signs and shouting slogans, etc. claiming action on the part of world leaders convening in Copenhagen. They might need a helping hand from libraries.

References

Alstad C., & Curry, A. (2003). Public space, public discourse, and public libraries. *LIBRES Library and Information Science*

Research Electronic Journal, 13,1, Retrieved April 20, 2010 from http://libres.curtin.edu.au/libres13n1/pub_space.htm

Brabazon, T. (2006). The Google effect: Googling, blogging, wikis and the flattening of expertise. *Libri, 56,* 157-167.

Doctor, R.D. (1994). Justice and social equity in cyberspace. *Wilson Library Bulletin*, *68,* 34-39.

Durrance, J.C., Pettigrew, K., Jordan, M. & Scheuerer, K. (2001). Libraries and civil society. In N. Kranich (Ed.), *Libraries and democracy: the cornerstones of liberty* (pp. 49-59). Chicago: American Library Association.

Jadinge, A.L. (2004). (The Political Library. The Public Library as a space for citizens' participation and public discourse). Retrieved September 29, 2010 from http://www.abm.uu.se/publikationer/2/2003/189.pdf

Keen, A. (2008). *The cult of the amateur: how blogs, MySpace, YouTube and the rest of today's user-generated media are killing our culture and economy*. London, Boston: Nicholas Brealey Publishing.

Kranich, N. (2001). Libraries, the Internet, and democracy. In N. Kranich (Ed.), *Libraries and democracy: the cornerstones of liberty* (pp. 83-95). Chicago: American Library Association.

McCabe, R. B. (2001). *Civic librarianship: Renewing the social mission of the public library.* Lanham MD: Scarecrow Press.

Matarasso, F. (1998). *Beyond book issues: the social potential of public libraries*. London: British Library, Research and Innovation Centre, Comedia.

Skot-Hansen, D., & Andersson, M. (1994). *Det lokale bibliotek: afvikling eller udvikling*. Copenhagen: Danmarks Biblioteksskole and Udviklingscenteret for folkeoplysning og voksenundervisning.

Waller, V. (2009.). The relationship between public libraries and Google: Too much information. *First Monday*. 14, 9. Retrieved April 20 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2477/2279

Address of congratulating author:

Leif Kajberg
31 Klirevænget, DK-2880 Bagsværd, Denmark
Email: leif.kajberg[at]gmail.com

Characteristics and Background of a New Paradigm of Information Society Statistics[1]

Bin Lv[1] **& Guoqiu Li**[2]

[1] Shanghai University, Shanghai, China
[2] East China Normal University, Shanghai, China

Abstract: From the 1990s, as a result of the popularization of new information technologies and the carrying out of developmental strategies for the information society by many countries, informatization measurements has turned into information society statistics, covering a broader scope. This article summarizes the characteristics of information society statistics, and analyzes the background of this transformation.

Keywords: Information Society Statistics, Characteristics, Background

1 Introduction

In 2006, Bin &Guoqiu [1] outlined informatization measurement research performed by business organizations, academic institutions and national as well as international organizations. Since 1995, informatization measurement research in Europe and America has evolved quickly. The level of progress in research on these issues, which can be seen from many different viewpoints, suggests the possibility of developing a new field of research: information society statistics.

The article is structured as follows. First, the historical development from informatization measurements to information society statistics is discussed, taking inot account the development of statistical models and indices; and the influence of ICTs on the information society and economy, as well as on policy making. This trend is emphasized from the mid-1990s and onwards. This is followed by a more detailed discussion on the technological and social background of statistical developments, based on basic statistical research theories, such as the long wave theory. A brief concluding section ends the article.

[1] This paper is related to the research project "Study on Informatization Measurement based on the Core Information Ability of ICTS", supported by the China National Planning Office of Philosophy and Social Science (http://www.npopss-cn.gov.cn/, 05BTQ012, 2005-2007).

2 From Informatization Measurement to Information Society Statistics

According to research, the characteristics of information society statistics are as follows.

First, the breadth of information society statistics is wider than that of informatization measurement. Informatization measurementaims at measuring the degree of development, as well as distribution, of information channels, media and technology development in a particular country or region. Before the 1980s, informatization measurements basically appeared as individual cases, such as the informatization index of Japan and the measurement of the information economy by Porat, a case which was widely referred to in China. However, these research efforts received very little attention outside Asia and involved very few countries and institutions. After the mid-1990s, many organizations – from business organizations and academic bodies, to national and international organizations, including the UN –began to pay attention to informatization measurements and to conduct research along these principles [2]. Examples on informatization measurement initiatives are:

- the index of the information society by IDC/World Times;
- the framework for analyzing, as well as the actual analyses of, internet expansion by the Mosaic Group;
- the e-Readiness analyses by Mcconnell International Consultant Inc of USA;
- the first report of e-Readiness by the EIU;
- the WEF/Networked Readiness Index (NRI) by the WEF;
- from 1999 and onwards, the continuously published reports on the electronic and digital economy by the US federal statistical bureau ;
- the index system of information society statistics, DAI, presented in 2003 by the ITU;
- ORBICOM of UNESCO announced the index of information conditions, synthesizing measurements on information density (the total quantity of capital and service of ICTs, including networks and ICTs skill) and conditions for information use (user rates and expense capacity of ICTs).

Second, the theoretical and methodological foundations of informationzation measurement and information society statistics are different. This question involves *what* informatization measurement should or could estimate, i.e. the object of research for informatization measurements. Consequently, it relates to how the information and knowledge society can be understood. In fact, in his book *Theory of Information Society*, Frank Webster pointed out the lack of the precise standards for measurement as the biggest flaws in exsting theories on the information society [3]. For instance, RITE in Japan measured the total quantity of information, whereas Porat measured the information sector (i.e. the information industry and the information profession). Many scholars ahave criticized these two measurement methods for being insufficient [4].

After the 1990s, various countries started to come to agreements on the objects of measurement for analyzing information technology and its proliferation. In fact, today the international definition of the concept of the 'information society' is based on the development of ICTs. Education, knowledge, information and communication are at the core of many efforts towards human progress, and today, ICTs have an enormous influence on nearly all aspects of our lives [5].

Presently, three central aspects of the proliferation of information technology have been generally accepted on an international level: *infrastructure*, *application* and *influence* of information technology. A country, an enterprise or a family will be isolated from the information society without access to computer networks. But the access to computer networks does not automatically imply an effective use of the network. Specific quality and skills are also required. Once the use of computer networks reaches a higher degree, it may have a remarkable influence on various aspects of the society and might bring about changes on economic and societal structures.

Third, the research and practical implementation of statistics on informatization involves developing *official statistics*, or at least, to prepare the production of official statistics. This characteristic is closely related to the advancement of the information society. Before the appearance of the Internet, the information society was primarily discussed by scholars of futurology. Although some international experiments had been performed, such as Porat, the indices developed and, results of the analyses, never truly the systems of official statistics. Partly, this was because of methodological problems. More importantly, though, the development of the information society at the time (prior to 1990) had not achieved a level where there was any need for carrying out official statistics. Thus, before the 1990s, research on informatization measurements was done by academic institutions, whereas national and international organizations were not yet involved.

The situation changed completely in the late 1990s, due to the development of the Internet and the enormous increase in access to cell phones and personal computers. One may observe that the early cases of statistics were carried out by the business organizations such as IDC and Mosaic. But following the report on the knowledge economy by the OECD and the report on the digital economy by American Commercial Affairs, many countries started analyzing and producing statistics on the information society. The most central statistics concerned the information economy. Coming into the 21st century, the progress of information society statistics grew faster, with many international organizations joining in. After the Information Society World Summit, a core set of globally unified indices on the information society were developed, soon to be harmonized with official public statistics in the various countries.

The fourth characteristic is its ostensive *policy meaning*. For example, measurements on the digital gap informed about the disparity between different countries as well as between different communities within countries, and in which direction efforts can be directed to close the gap. Whether we talk of academic research or public statistical rankings, this kind of information society statistics is full of policy features.

First, the whole world has generally considered the index of ICTs as the symbol of a country's competitive advantage and comprehensive national power. Therefore, with the help of information society statistics, countries can compare themselves with their competitors, find out the differences and develop relevant strategies and policies to reduce the differences.

Second, the ICTs indexes can provide evidence to reduce the digital gap between the rich and the poor (countries. Upon entering the new millennium, the most important target of many international organizations such as the UN had already become the elimination of the digital gaps. An essential part of this is the carrying out of effective surveys on the situation. In fact, the digital gap does not only exist between rich and poor countries, but also between different areas or communities in a country. Thus, the measurements of information related issues are vital for understanding the situation and for developing appropriate policies.

Third, ICT related issues are also related to industrial and governmental policies. Telecommunication and telecommunication service policies, innovation and technical policies, education and labor market policies, as well as service trade policies and so on, are all depending on an accurate and comprehensive monitoring and statistical analyses of information technology and the information economy.

3 The Background of Information Society Statistics

Why did so many organizations pay attention to the research of information society statistics after the mid-1990s? How was the mutual agreement on ICT measurements achieved? And what was the background for the characteristics mentioned above?

3.1 Technical backgrounds

The development of information technology can be traced back to the 19th century. ITU is one of the earliest international organizations, with more than 150 years history behind it. But until the 1970s, information technology was still based on mutually independent technologies such as the computer memory, processing, and transmission of data, creating barriers to their application, popularization and further development. After the 1970s, development in the digitalization of information technologies made a *merging* of technologies such as computer memory, processing and communication possible, leading to the ICTs we have known since the 1990s.

Following the appearance of Internet, the development of ICTs entered a new phase. In 1993 the US proposed the plan for a national information super highway, bringing the whole world the potential for constructing new information infrastructures. The development and rapid popularization of the networks made the prospect of the information society at arm's length. Various countries and regions, such as the EU, proposed efforts toward the construction of an information society; and made this an important political goal.

ICTs are, like the electricity in the past, penetrating technologies. Not only does it enable us to do new things, but also tell us how to do it. Only by cherishing digital technologies in our businesses and society, the European countries can achieve their full potential. Not only was the importance of the construction of a knowledge-based future through ICTs [6], an understanding held by various European countries, but it also became important in the process towards increasing mutual recognition other countries and regions as well.

The development and application of technology has not only changed the way of manufacturing products, but also the life style. Furthermore, it promotes the transformation of economies and societies. There are new products, services and industries coming out of the new technologies; and their application in different businesses and industries, as well as governmental agencies speeds up the process of informatization. There is no doubt that the proliferation and application of ICTs and information products became very important after the 1990s. People's acknowledgement of the information society will concentrate on technology. Hitherto, the grandest expression of these tendencies was the "Principle Manifesto", passed at the Information Society World Summit in 2003: "under the advantageous condition, these technologies might become the powerful method to help to enhance productive forces, to promote economic development, to create employment opportunity, to enhance employment possibility, and to improve the quality of life of all people. Information & communication technology may also promote the dialogue between different peoples, between different nations and between different civilizations" [7].

No doubt such a mutual recognition came out of the Information Society World Summit being organized and led by ITU, whose perspective was technical. This kind of understanding is reasonable. Since the 1990s, because of the rapid popularization of the Internet, any theory on the information society or informatization would definitely lack persuasive power if not taking information technology as its main clue. It would be similar to theorizing on the industrialization, without taking the steam engine and electric power into account, technologies without which the industrialization process would be inconceivable.

The present upsurge of information society statistics started in 1995, after the expansion of the Internet started. This is by no means a coincidence. The ICTs –represented by the Internet, and later on, other social media –constitute the technical background of this fast development of information society statistics. That is the real background underneath the rapid proliferation the ICTs, brought to the information society and the ensuing need for estimation of the information society at different levels. Based on the development of new technologies, the studies and practices of information society statistics of the late 1990s and onwards have a distinct identity; and distinguishes itself from former research and practices.

3.2 Social advancements and policy background

After the mid-1990s, the construction of the information society received widespread attention in the economically developed countries, and became a central policy goal in European countries.

In 1993, the US had proposed the plan to construct a national information infrastructure (the information super highway plan), which generated a giant echo in the world. The European Union immediately proposed its own development plan for the information society fearing that the EU would fall behind the US in the competition. The European plan for the information society can be seen as a response to the American plan to develop a national information infrastructure. In 1994, the European Union decided that the technologies related to the information society would be a crucial area of development, and officially included research on ICTs in the scope of the third framework programme plan of the European Commision.

The information society conference organized by the G7 in 1995 was an important milestone in the development of the information society. It was the first time the concept of the information society to discuss from a global point of view. More importantly, the discussions did not remain on a strictly academic level, but transformed the concept into public policy. This conference had a global impact on industrial, newly industrialized and developing countries.

In 2000, members of The European Union proposed the Lisbon Strategy in Lisbon. The goal was: to achive a high growth, higher employment and stronger social inclusion; and ICTs was considered the essential factor. Afterwards, plans such as e-Europe, e-Europe 2005, etc. was presented as well, providing further policy impetus to the development of the information society in Europe.

The UN also made unremitting efforts to promote the whole world becoming an information society, by being an active partner in the preparations of the Information Society World Summit in 2003 and 2005, the first global conference in history using the concept of the information society in the title. Before the conference, various countries and local committees of the UN participated in negotiating a generally accepted consensus on some principle questions. The background of the summit was the radical and ongoing transformation from an industrial to an information society. This information revolution was affecting people's life, ways to study and work, as well as the interaction between government and civil activities in various countries. The conference passed four important documents: the "Geneva Principle Manifesto", the "Geneva Motion Plan", the "Tunisian Mutual recognition", and the "Tunisian Motion Agenda".

It is these social advancements and policy goals that formed the need to carry out measurements of different aspects and elements of the information society. That need came from many strata of society, including international organizations, national policy makers, enterprises as well as individuals, who need various forms of statistics on the information society.

Measuring the development and present situation of the information society is the most important, since it is no longer is a topic of discussion for futurology scientists but is an existing reality. Thus, people living in the information society need to understand it, and in particular, to understand the ICT perspectives and the degree of ICT proliferation and influence.

Next, it is very important for policy makers to understand disparities in ICT development and applications on regional, national and international levels.

Third, new forms of inequalities also need to be considered quantitatively, by measuring differences in the mastering information skills or the ability to effectively to use modern information technology.

3.3 Progress of fundamental research

The theoretical understanding of the information society has changed along with the development of the information society. Frank Webster once pointed out that there were five aspects to consider for understanding the information society: technology, space, occupation, economy and society. After the 1990s, because of the popularization of ICTs, and especially the fast development of the Internet, the recent generation of information technologies has become centre of people's attention again, and also, a core aspect for information society statistics.

3.3.1 Long wave theory and corresponding information technology-economy paradigm

The Long wave theory, proposed by the economist Condra from Czechoslovakia, suggests that economies, at least since the 1800s, are develops cyclically over 40-60 year periods, led by the development of core technologies. This model has provided a useful framework for the research on the information economy and the information society. The long wave of ICTs started in the 1970s or 1980s, lasting at least 20 years. According to long wave theory, each wave, or cycle of development, can be divided into four stages: initiation, growth, transition and the transformation of the model. The development of information technologies and the information economy since the 1990s is analogous to the theories proposed in long wave theory. Therefore, the paradigm of information technology and economy might constitute the framework for information society statistics [8].

3.3.2 The S-curve

The S-curve was proposed by Canadian scholars in 1999[9]. The intention was to explain how the proliferation of ICTs can be divided into several, closely interconnected, stages. This model has come to be widely used for analyzing ICTs and their interaction with economy and society in various countries. These stages can also be understood as degrees of development of informatization.

The *preparatory stage* is the initial step of informatization, with the development of infrastructures, human and economical resources, and abilities of informatization. These are the foundations of informatization.

The *operational phase* equals an intermediate stage of informatization. This stage builds on the requirements and development of the previous stage, i.e. when all preparatory requirements and prerequisites for informatization have been met; and a phase of implementation can start, involving e.g. e-government, e-commerce and so on.

The *influence stage* is the next phase of informatization development, where the informatization has reached a higher level. It has an obvious influence on economy and society, even the potential to change the structure of economy and society. For example, the present share of electronic commerce in comparison to the total amount of transactions is probably no more than 10%, even in the US. This has no decisive influence on the trade structure of the entire society, and even less so, worldwide; nor does it call for a need to redefine transaction behaviors in the trade agreements. However, what would the situation be if the share of e-commerce surpassed 50%?

Aside from the above-mentioned theories, the supply-demand model proposed by OECD for studies in information society statistics is also widely used. Further, Ahokas & Kaivooja proposed the MAS model [10]. It is based on the idea that to analyze technology only is insufficient. In this model, M represents *Motivation*, A represents *Access* and S represents *Skill.* These three factors represent the three central factors in a three-dimensional model. This model is only one out of several examples on attempts to explore and explain these phenomena.

4 Conclusions

The utilization and the proliferation of information technologies, along with the emergence of the information society brought about by the ITC development, have made information society statistics an extremely important research topic. In the process of development, information society statistics has changed into becoming official statistics for policy making on both national and international levels. One underlying reason is the increasingly obvious lack of ability of traditional statistics to analyze ICT and information society related issues, and that information society statistics provides a solution by bringing forward new statistical models for the benefit of economy and society.

Annotation and Reference

[1] Lv Bin, Li Guoqiu (2006), Measurement of information society: a new topic of studies in information society, *Journal of the Chinese Association of Library Science,* January: 18-23. (Text in Chinese)

[2] http://www.bridges.org/files/active/0/ereadiness_whowhatwhere_bridges.pdf

[3] Webster, Frank (2002), *Theories of the Information Society*, Second Edition. New York: Routledge : p.8

[4] Li Guoqiu ,Lv Bin (1996), Concept of information society and its cultural supposition in US, *Library and Information*, (3). This article discussed the insufficiency of the Porat system.

[5] Principle Manifesto" (2003), passed at the Geneva Information Society World Summit, (Eighth article), http://www.itu.int/wsis/documents/index1.html

[6] ISTAG (Information Society Technologies Advisory Group) (2006),SHAPING EUROPE'S FUTURE THROUGH ICT, March 2006. http://www.cordis.lu/ist/istag.htm

[7] "Principle Manifesto" (2003), passed at the Geneva Information Society World Summit, (Nineth article) http://www.itu.int/wsis/documents/index1.html

[8] The paradigm of an information technology-based economy was proposed by Manuel Castells, a sociologist active in the US. Manuel Castells: *The Rise of the Network Society*, Social Science Academic Press (China), 2003.

[9] United Nations Conference on Trade and Development (2003), INFORMATION SOCIETY MEASUREMENTS: THE CASE OF E-BUSINESS,Background paper by the UNCTAD secretariat,TD/B/COM.3/EM.19/2

[10] Ahokas, Ira and Kaivo-oja, Jari (2003), Benchmarking European information society developments. Foresight, 5(1).

Addresses of congratulating authors:

Bin Lv
Department of Library, Information and Archives, Shanghai University
99 Shangda Road, Baoshan District, Shanghai, 200436 China
Email: lvbin[at]staff.shu.edu.cn

Guoqiu Li
Department of Information Business School, East China Normal University
500 Dongchun Road,Shanghai, 200241, China
Email: gqli63[at]sina.com

The Development of a Scientific Field, its Research Output and the Awareness of a Scholar Along its Lines

Bluma C. Peritz

Hebrew University of Jerusalem, Jerusalem, Israel

Abstract. The roots and development of a scientific field and its research output by looking at the contribution of one of its scholars.

1 Introduction

This is not a scholarly paper neither a biographical sketch, but an essay about the development of a scientific field, its research along the lines with the awareness and contribution of a scholar who always looked ahead.

With this idea in mind I began to retrace the history and found that I have to go back to the years when library studies were first formalized.

My account is that of an observer, teacher and researcher coming from a similar academic professional environment and roots as Peter Ingwersen.

2 Roots and Developments

Information science has a history of taking techniques from the past and adapting them to the present needs. Despite it's origins, information science is a social science because it deals with an artifact of man: information. Wersig [1] calls it a "post modern science" because it is driven by the need to develop strategies to solve the problems caused by the classical sciences and technologies.

The study and literature of information science has not developed along the same directions in different parts of the world. The influences, constraints and needs have varied greatly from one area to another. Grover and Glazier [2] have proposed a taxonomy of theory for library and information studies research for intelligent information systems. The rapid growth of scientific research establishment and the development of a sophisticated technology of information storage and retrieval were bound to change completely the study of the processes of information handling.

3 Information Retrieval

In a paper Peter Ingwersen published together with Irene Wormell [3] entitled: Modern Indexing and Retrieval Techniques; matching different types of information needs they wrote: "The main objective of information retrieval is to facilitate the effective communication of desired information between a human generator of information and the human user".

This statement became probably the guiding lines who led to the publication of his book Information Retrieval Interaction [4] a great contribution to the understanding of the subject who became an important textbook.

The early roots of IR research can be traced in the theory of classification and indexing, tools Peter Ingwersen was well equipped with. But, as I mentioned before the developments of information and communication technology has been the major driving force at the development of IR from early small batch-mode IR to today's Web based systems [5]. Ingwersen [6] was the first to define specific indicators for the Web.

Chicago 1995, during the 5th International Conference on Bibliometrics Scientometrics, in a long late late night discussion over lots of beer and coffee Peter raised his concern regarding the still lack of a body of knowledge, a strong theory and models for the development of IR. Information retrieval became one of the most important fields of research in Peter's career, a period of notable achievement. He was there at every step of its development, from content analysis and description of documents to means to improve subject access, to find practical solutions to problems in online IR.

Information retrieval developed with time to include all aspects of processes, systems and techniques to access and retrieve desired information. From Boolean logic and probability for online IR, to cognitive analysis [7], clustering, matching different types of information, expert systems and last, not least, the awareness for user oriented IR. He stressed the importance of users' studies and information seeking behavior, to learn about the problems and especially about the needs. This awareness will lead later on to other important developments and research, like information interaction and context

4 Bibliometrics-Informetrics and more

Peter Ingwersen's "multi" interests, like "multi florous" bearing more than three flowers developed in parallel.

Another field of investigation and research who became during the years closed to his mind was Bibliometrics-Scientometrics. The Royal School became one of the few training institutions to have a Center for Research in Bibliometrics. The research conduct by Peter and his colleagues in this field included many aspects of national and international interests. To mention just a small selection: on the national level: Danish and Scandinavian research; Evaluation of strategic research programs: the Danish environment [8].

On the international level: Publication behavior and international impact; Citations and impact of research; Applying citation analysis to evaluate research programs.

Last but not least, his latest interest and research is a branching of from his very first interest Information Retrieval to Integration of Information Seeking and Retrieval in Context [9].

Although the different fields Peter studied and researched are strongly connected they still have each one of them their own theories, models, techniques and methodology.

Peter, until 120 there is still a long way to go.

5 Emeritus

According to the Oxford English Dictionary the definition for Emeritus is: “Honorable discharged from service” giving as an example retired Professor. Funk Wagnalls’ Dictionary definition is: “Retired from active service but retained in an honorary position” giving as an example Pastor-emeritus. The second definition fits Peter better.

I can hardly see Peter discharged; this is almost impossible giving his many involvements in on-going research projects, professional committees and especially his devotion for teaching and guiding the new generation of professionals in the field.

His contribution to the development of Information Science, Information Retrieval, Evaluation of research and Integration of Information is well documented in the literature and appreciated by his fellow researches, colleagues and students.

For his distinction in research he was awarded the De Solla Price Medal and for his excellence in teaching he was selected best teacher by the American Society for Information Science and Technology.

To like and enjoy Peter you have to know him and his sense of humor.

I wish Peter for the years to come to keep up with his vitality, intellectual curiosity, freshness of mind.

To Peter with endearment and appreciation from an old timer, Berlin 1993 (even before from FID, without being properly introduced).

References

1. Wersig, G. Information Science: The Study of Postmodern Knowledge Usage. Information Processing and Management. 29, 229-239 (1993)

2. Grover, R., Glazier, J.D.: A Conceptual Framework for Theory Building in Library and Information Science. Library and Information Science. 8, 227-242 (1986)

3. Ingwersen P., Wormell, I,: Modern Indexing and Retrieval Techniques: Matching Different Types of Information Needs. In: Information, Knowledge, Evolution: Proceedings of the 44 FID Congress, pp. 79-90. North-Holland, Amsterdam (1989)

4. Ingwersen P.: Information Retrieval Interaction. Taylor-Graham, London (1992)

5. Ellis, D.: Progress and Problems in Information Retrieval. L.A., London (1996)

6. Ingwersen P.: The Calculation of Web Impact Factors. J. of Documentation. 54, 236-243 (1998)

7. Ingwersen P.: Cognitive Perspectives of Information Retrieval Interactions: Elements of a Cognitive IR Theory. J. of Documentation. 52, 3-50 (1996)

8. Ingwersen P. and Larsen B.: Evaluation of Strategic Research Programs: The Case of Danish Environmental Research 1993-2002. Proceedings of ISSI 2005, pp 450-459. Karolinska University Press, Stockholm (2005)

9. Ingwersen P. and Jarvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer, New-York (2005)

Address of congratulating author:

Bluma C. Peritz
Hebrew University of Jerusalem, Jerusalem, Israel
Email: bluer[at]cc.huji.ac.il

Renewals and Affordances in Libraries

Niels Ole Pors

Royal School of Library and Information Science, Copenhagen, Denmark

1 Introduction

This paper is exploratory in nature investigating a phenomenon that has gone unnoticed in the recent professional and academic literature. The phenomenon is renewals of library loans. In the paper, the author looks at the development of renewals in relation to first-time loans and generates hypotheses concerning the phenomenon. These hypotheses could possibly be a basis for further studies and investigations. The number of renewals of materials has increased very rapidly. This has to be seen in the perspective that the total number of circulation appears to be rather stable. The paper focuses on renewals in both public and academic libraries.

The paper discusses the situation in Denmark, but the results ought to be of wider interest because the focus is really on the relationship between user behaviour and technological development

The overall purpose of the paper is to explore intended and unintended effects of technological development in relation to user behaviour. Another purpose is to explore how well an ecological perspective and approach can contribute to explanations of intended and unintended effects. The specific purpose of the paper is to explore changes in renewals and analyse these changes in relation to accessibility, availability, use and perceived misuse of the whole system. This analysis departs from an ecological perspective discussing the merits and the demerits of the system in relation to user behaviour.

This paper takes a macro-ecological view of the traces of library users' behaviour as reflected in the statistical data of renewals. The theoretical perspective is the theory of affordances. This theoretical perspective is appropriate for the analyses of data because it is easy to document intentions behind the system and it is also rather easy to identify perceptions and unintended use of the system. In reality, the paper is concerned with issues about how the interaction with an advanced information system affects behaviour and how users react to the affordances inherent in the system.

It is old news that digital access to the databases of libraries has changed the way users behave in relation to both physical and digital collections. However, several interesting topics have gone rather unnoticed both in research and in the professional debate. One of these topics is investigated in this paper. One can look at renewals from different perspectives. One perspective is of course that it is a service improvement for the users giving them more time to study documents and it also increases their convenience. Another perspective is what it means in relation to availability of documents for the browsing user - be it digital or physical browsing. Renewals are also interesting in relation to different types of materials and media.

2 Affordances

Affordance theory can be considered as a kind of subset of a broader ecological perspective. Williamson [1] introduced it in the library and information science literature claiming that the approach would give a richer and more detailed understanding of information behaviour in a full context. The ecological approach can of course be found in many disciplines. It is for example often used in the management literature studying strategic planning and development. Givens and Sadler [2] have developed the concept in library and information science with inspiration from ecological psychology and give a rather simple definition of affordances and affordance theory. Their starting point is the simple observation that objects have embedded several affordances. Some of these are intended by the designer or creator, others are not. Users of the object can perceive this affordances either as intended by the designer or they can interpret the affordances in the object differently using the object in a way not intended by the designer. In their paper they give many examples of these possible discrepancies between designer intention and user perceptions of objects and the appropriate way to use the object. It is in this perspective about intentions and perceptions.

In the present paper, the affordance perspective is combined with an ecological perspective. An ecological perspective is especially appropriate in this context because we look at the all the libraries in the nation as a system in relation to renewals. Renewals are a facility all libraries have on their local web-pages. All users can investigate the collection in their favourite libraries and they can make reservations and renewals unless other users have requested the document and they can perform some other tasks like cancellations of requests and similar. The possibility for renewal of documents has always existed, but earlier one had to take the document to library to loan it again it. Now all these operations can be conducted from home or from any computer. There is no doubt that the overriding intention is to give users a service improvement – in other words to make library activities more convenient.

The ecological psychology works primarily on an individual level but the perspective offered in this paper is based on an assumption that it is meaningful to use the approach on groups of users of a cooperating library system.

3 Methodology and data

A literature search concerned with renewals in the library literature did not reveal new research or even a professional debated concerning the issue. This is interesting because the profession has always been occupied with the number of loans and renewals constitute an increasing part of the loans.

The data has been collected from 2000 to 2009 from the Danish Library Statistics [3]. The data consists of the number of renewals and the number of interlibrary loans for every single library in the country. For some of the years, we have more detailed data dividing the data into types of media and materials..

The paper is built on longitudinal data covering 10 years from both public and academic libraries. For the year 2007 and 2008, detailed data has been collected for renewals in every single public library system in the country. The detailed data give information about the distribution of renewals on the different types of materials or media and renewals among children and adults.

This paragraph starts with the longitudinal data illustrating the development of renewals in libraries.

4 Loans and renewals

This part of the paper gives an overview of the development and trends in interlibrary loans and renewals. Further, this part also contains a more detailed exploration and discussion of the trends and factors influencing or interacting with them. The Danish National Library Authority introduced bibliotek.dk in 2000. It is a unique service. It is simply a database including collection data from all and all types of libraries in Denmark. Every citizen has access to the database and have the right to order a book, a CD, a DVD or other kind of materials delivered to a library of own choice. It means in principle that every citizen has the right to lend materials from every library in the country, totally free of charge. The library authority has established a system of transport that means that the library system is connected on a daily basis by a transport system that delivers materials from one library to another. The bibliotek.dk system has many options for the user. They can sort according to different criteria and they can renew their loans. Bibliotek.dk is seamlessly integrated with the local online catalogues and it gives the user several pathways into both the local collections and the national collection.´

One of the effects of the introduction of this service is that interlibrary loans and interlibrary borrowing increase and especially loans from academic libraries to public libraries are prominent in this process. One can argue that the national collections of media are used more and by a more differentiated group of people

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total loans	73	72.5	71.7	72.2	72.7	74.7	73.7	72.8	71.6	74.5	76.9
Renewals	7.5	10.7	12.3	14.3	16.4	17.9	19.6	21.3	22.7	26.3	28.6
First-time loans	65.5	61.8	59.3	58	57.2	56.8	54.1	51.6	49	48.2	48.3
% Renewals	10	15	17	20	23	24	27	29	32	35	37

Table 1: Loans and Renewals in Danish Public Libraries in Millions

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Downloads	1.4	2.8	3.6	5.5	7.1	10	11.2	11.9	13.1	16.4
Renewals	0.1	2	2.5	3.5	3.5	3.9	3.1	3.1	3	3.1
First-time loans	4.1	3.9	3.9	3.6	3.5	3.3	3.1	2.8	2.7	2.6
Total loans	4.2	5.9	6.4	7.1	7	7.2	6.2	5.9	5.7	5.7
% Renewals	2	34	39	49	50	54	50	53	53	55

Table 2: Downloads, Loans and Renewals in 16 Academic Libraries in Millions

due to the technological development in the information technology. However, the discussion here will limit itself to a discussion concerning renewals.

This paragraph is a descriptive one presenting the development and trends in the amount of renewals in the Danish library system.

Table 1 gives an overview of loans and renewals in the total Danish public library system. It is evident from the table that the total circulation consisting of first-time loans and renewals has been rather stable during the last 10 years. However, the proportion of renewals has increased from 10 % to 37 % in the period, indicating that more than one third of the circulation is a renewal.

The number of first-time loans has decreased every year since 1999 till 2008. In 2009 there is a small increase in first time loans, but overall the decrease in the period is rather significant. Overall, the trend is clear. It should also be noted that the composition of loans change during the period. Newer media like audio books, music and films on DVD constitute an increasing proportion of the total circulation.

The only reason that loans at first glance appear to be rather stable is the increase in renewals. The increase in renewals is dramatic.

Table 2 shows the same type of trends in the academic library sector. The table includes data from 16 of the biggest academic libraries in the country. The number of downloads is also included as it gives a more detailed picture of behaviour.

The increase in downloads is remarkable and it can be considered surprising that loans of physical materials declines rather slowly. Due to renewals we see nearly the same picture as in public libraries. The total number of loans of physical materials is rather stable. However, the proportion of circulation that is renewals is now more than 50 %.

Collection	Loans	Average loan period	Opening days	% of collection in circulation
100.000	200.000	25	250	20
100.000	200.000	35	250	28
100.000	300.000	25	250	30
100.000	300.000	35	250	42

Table 3: Worked Example demonstrating the Availability of Documents in relation to Average Loan Periods. Circulation-ratio and Number of Opening Days

It is not possible to renew books, journals or other types of media that other users have reserved. This applies to both types of libraries.

It is evident that the immediate availability of physical documents decreases as an effect of the increase in renewals. This point can be illustrated by a simple example. The example only demonstrates what happens to the immediate availability when a library send part of its collection to other libraries and its users engage heavily in renewal activities. This is of course not new and it relates directly to the discussion of ownership, access and availability [4]. Renewals imply that the user browsing in the physical library and the user browsing in an online catalogue will notice that fewer documents are immediately available. Both factors are important. We do not have any indication of a decreasing number of visitors to either public or academic libraries. This probably means that it is difficult to argue that immediate availability is less important than it was before.

Several factors influence the availability. The amount and distribution of requests is important and it is also pertinent to acknowledge that requests can be in the form of browsing implying a less focused need. The length of the loan period is also important in relation to availability and the size and composition of the collection including the number of copies of single documents are further important factors [4].

In table 3, some of these factors are included. The table outlines with a worked example the effect of an increase in the average loan period. The assumption is that renewals increase the average loan period.

The formula used is simply the number of loans * the average loan period divided by the number of opening days. This will give the proportion of documents in circulation or not immediate available on an average day.

It is important to stress that it just is an illustration demonstrating the effect of the circulation ratio and the average loan period.

In the next paragraph we will explore the phenomena of renewals in more depth.

5 Analysis and Discussion

The analysis departs from a very detailed set of data of renewals. For every single public library we have obtained the number of renewals for every type of material. In previous paragraphs the paper explored the longitudinal and distribution of the average proportion of renewals. For the years 2007 and 2008 we have the detailed information and it will be explored in this paragraph as a starting point for the discussion.

We will start to look at the distribution of the proportion of renewals in relation to the circulation according to type of material. We only look at materials for adults but the data for library materials for children is quite similar and there appears to be no difference in renewal behaviour between adults and children.

This table simply shows the proportion of renewals in relation to circulation and it is evident that audio books and films on DVD are renewed a much less than the other categories of materials. This is probably due to the fact that these categories are becoming increasingly popular and many of the public libraries have introduced flexible loan periods (1 week or 2 weeks) for the most popular categories.

However, we also see that an increase in renewals can be found in relation to all categories. The correlation coefficient between renewals in 2007 and 2008 is 0.94 indicating a very high correlation. However, we do see a rather marked difference in renewals in the different public libraries.

Table 4 shows average proportions of renewals. It is possible to look at these proportions in relation to every public library. Denmark has 98 library systems. Each municipality is by law required to run a library service.

	2007	2008
Books	34	37
Serial	32	35
Audio books	23	24
Music	31	35
Film (DVD)	25	30
Multimedia	31	36
Other	31	35
Total	32	36

Table 4: The Proportion of Circulation coming from Renewals in relation to Type of Document in Public Libraries in 2007 and 2008 in %

In 2007, the range of proportion of renewals went from 12 % to a maximum of 68 %. The similar figures for 2008 are a minimum of 14 % and a maximum of 62 %. 8 public library systems had in 2008 renewals proportion of more than 40 and in 2008 it was 11.

The public libraries with the highest proportion of renewals in relation to the circulation are nearly all situated in Copenhagen and the suburban area and in or in the vicinity of the other big cities in Denmark. The big cities have of course all universities, college universities and a number of professional schools. At the bottom of the list we find public libraries situated in smaller towns and on islands or areas with a significant rural characteristic. This reflects that the users of bibliotek.dk more often than other users are situated in the Copenhagen areas and they are more often than other people students or participating in different educational activities. This is of course also an indicator of the use of the local library’s online catalogue. It is an indicator because several user studies have demonstrated similar characteristics among the heavy users of online catalogues and further, because the interaction between the national system and the local system is rather seamless leading the user from one system to the other depending on the character of request or activity [5].

From here, the rest of the paper consists of a more hypothetical discussion due to the lack of previous investigations into the matter.

The increase in renewals in both types of libraries is probably just an effect of the technological possibility and ease of pressing the renewal bottom. More and more materials people use the databases and get exposed to the opportunity. If we do look at it in the light of affordance theory or in the light of an ecological perspective, it will probably be safe to argue that the designers behind the system had an intention of contributing to a kind of service improvement giving the users an easy and convenient way to renew materials. It has probably not been the intention behind the system design to design an interface that would be important a factor contributing to a decreasing availability of documents.

A hypothesis is that the design is influenced by an intention of service improvement for the single user. It has probably not been the intention that users would renew as much as they do. This service improvement is perceived by the user and the users obviously employ the opportunity very much. What is not perceived by the user is the interaction of the different parts of the library system that relate to other users' needs and unspoken requests, the effect on availability and maybe even the attraction of either physical collection or the perception of the online catalogue.

It is also interesting that the system design gives the user a possibility to look at all the documents they have on loan and they are with one click able to renew all of them – unless of course they are reserved by other users. This is very convenient but we really do not know why users renew as much as they do and we do not even know how it affects the average loan period.

The technological developments work together and interact. The majority of both public libraries and academic libraries have introduced services like email notification or sms notifications to users when the loan period is about to end. This service is often promoted by the single library and the intention is probably to increase the service quality and to get the material back to the library. However, it could be perceived by the user in a not intended way as a trigger to renew all the material simply to avoid a fine for bringing back material too late. A not intended side effect of this is that the income of libraries decreases as fines has contributed to income generation.

The affordance perspective could also be used to analyse a similar phenomenon like interlibrary loans. This is also an activity that increases. One of activity features of this activity is that many users place order literature or reserve it both in their own library and in other libraries. The effect of this is of course a diminishing availability and a rather huge number of materials not picked up [6].

It is evident that there is a real lack of studies investigating both the effects of renewal behaviour on the cooperating library system in terms of loan periods, accessibility and availability but there is also a need to look deeper into the phenomenon investigating users' perception of the activity, including their knowledge, motives and interpretations of their activity. Further, here is room for more usability oriented studies that elicit information about the design issues in the human – computer interaction.

6 Conclusion

It is obviously that the possibility to renew loaned documents can be considered a service improvement for the users at least in terms of convenience. However, it diminishes the immediate availability of documents. This applies to both physical and digital access. What the consequences are of this diminished availability is totally unknown and further research must be conducted.

The rather dramatic increase in renewals signifies new information behaviour among at least a group of users. The reasons for this increase is probably the fact that more and more people use the libraries' database and become more aware of the possibilities for organising a library visit. but also for using the system to renew. This is probably connected to another service improvement. It is the fact that nearly all public libraries notify by email or by sms the users that the loan period is near the end. A user would normally get this notification 4 -5 days before the end of the loan period. It can easily be hypothesised that at least some of the users act on this information by renewing documents on loan. It has as a matter of fact the side effect that the income of the libraries go down because they will after this service improvement lack some of the income generated by overdue loans and the resulting fines.

Hopefully, it is obvious that the perspective of affordances is a fruitful perspective on the renewal activity as it structures the discussion into areas concerning intentions and perception. However, further studies could easily employ a multitude of theoretical perspectives including persuasive design, social capital and theories about the effect of technological change and how it is taken up by different groups.

It is important to stress that this paper only has scraped the surface of a complex phenomenon that really matters in the interaction between library systems and users. The paper indicates areas that possibly could be beneficial investigate more..

References

http://www.bibliotekogmedier.dk/biblioteksomraadet/statistik/biblioteksstatistik/ (The Danish Library Statistics)

Williamsson, K. Discovered by chance: the role of incidental information acquisition in an ecological model of information use. In: Library and Information Science Research 20 (1) 23-40. (1989)

Sadler, E. & Given, L.M.: Affordance theory: a framework for graduate students' information behaviour. In: Journal of Documentation 63(1), 115-141 (2007)

Buckland, M. Book Avialiability and the Library User. New York. Pergamon Press (1975)

Larsen, K. Brugerundersøgelse af bibliotek.dk. Ballerup. DBC. Oktober (2008)

Pors, N.O. 844 fjernlånsbrugere. København..København Kommunes Biblioteker. 2002.

Address of congratulating author:

Niels Ole Pors
Royal School of Library and Information Science,
2300 Copenhagen S, Denmark
Email: nop[at]iva.dk

The Historic Context Dimension Applied in the Museum Domain

Mette Skov & Brian Kirkegaard Lunn

Royal School of Library and Information Science, Aalborg, Denmark

1 Introduction

Our contribution[1] to the Peter Ingwersen festschrift takes as point of departure the "Nested model of context stratification for IS&R" introduced in Peter Ingwersen's and Kalervo Järvelin's monograph "The Turn: Integration of Information Seeking and Retrieval in Context" published in 2005. In the present paper, focus is on the historic context dimension in the nested model. A historic view on the museum domain illustrates how the historic context dimension can influence an information seeking and retrieval (IS&R) situation. Accordingly, the aim of the paper is to acknowledge the importance of including a historic context dimension in studies of IS&R.

2 The concept of context – approaching the 'unruly beast'

As researchers have moved away from a de-contextualised and system-centred view, context has become a hot topic both in information science as well as in other scientific disciplines (Dervin, 1997). The advent of, e.g., the "Information Seeking in Context" and "Information Interaction in Context" conferences has stimulated studies and theoretical discussions of context. It is generally agreed that information seeking and information retrieval (IR) are inherently interactive processes, which occur within multiple, overlapping contexts that inform, direct, or shape the nature of interaction (Cool & Spink, 2002, p. 605). In other words, how people access information is highly dependent on the context of their interaction and this context is influenced by a range of factors.

In spite of the growing attention given to context in information science no uniform definition exists of what the concept entails and which elements are important to information behaviour. In Dervin's (1997) analysis of various contextual approaches, she labels context as an 'unruly beast' because of the difficulties in characterising and

[1] The present paper is based on Skov (2009)

defining the concept and gaining methodological control over it. As a result the concept is used and defined variously. Part of the confusion is of paradigmatic nature.

Depending on the paradigmatic approach, context is conceptualised and analysed differently. The divide can be seen as a continuum. At one end, the objectified or positivistic approach sees context as yet another analytical factor that should be taken into account along with other factors (Dervin, 1997; Talja, Keso, & Pietiläinen, 1999). In this logic, context has the potential of being virtually anything that is not defined as the phenomenon of interest. The goal is to identify which aspects impact or relate to the phenomenon at hand and then adapt information systems based on these inputs. At the other end of the continuum, the interpretative or social-cultural approach does not understand context as an independent variable. Instead it sees context as a carrier of meaning without which any possible understanding of human behaviour becomes impossible (Dervin, 1997; Talja et al., 1999). In addition, every context is by definition different and thus generalisation in the traditional sense is impossible (Dervin, 1997). In the latter view, context is too complex and we cannot know which contexts are important. It is important to note that along this context continuum a variety of approaches exists.

In practice, context in library and information science studies usually refers to any factors or variables that are seen to affect individuals' information seeking behaviour such as work roles, work tasks, problem situations, communities and organisations with their structures and cultures (Talja et al., 1999). Contexts are multi-dimensional in that they can be described by a variety of attributes (Dervin, 1997). Examples of attributes that have been used to describe contexts include time, place, types of participants, history of interaction, the tasks motivating the interaction and the technical possibilities of the information systems. In addition context can be viewed as the socially defined settings, within which different *situations* take place (Allen, 1997).

Approaching context is seen as a way to add value, or 'digging deeper' in Dervin's (1997) terminology, when exploring information seeking behaviour, and it help us to understand complex relationships among contextual factors and human information behaviour.

3 The nested model of context stratification for IS&R

The discussion above shows how contexts within which a person seeks information are influenced by various factors. At another level of analysis, understanding context within the IR interaction itself is important (Cool & Spink, 2002). They describe four salient levels which to a large extent correspond to Wilson's (1999) information behaviour model: 1) the information environment level within which information behaviours take place. At this level concrete examples might be institutional, organisational or work task settings; 2) the information seeking level focuses on the goals a person is trying to achieve or a problem resolution task that influences the IR interaction level. At the second level, information use, resolution of ASK, or stages in a search process may be addressed; 3) the IR interaction level of context explores the user-system interaction within search sessions; *and* 4) the query level explores the linguistic level of context in association with system performance.

A similar multi level approach to understanding context is illustrated in Ingwersen and Järvelin's (2005, p. 281) nested model of context stratification for IS&R, introduced as part of the integrated framework. Compared with Cool and Spink (2002), the context stratification model adds two additional levels of context. Firstly, the relationships between single information objects, see dimension two. Secondly, the historic context, see dimension six. The nested model includes the following six dimensions of context illustrated in figure 1 below:

1. Intra-object structures
2. Inter-object contexts in form of relationships between single information objects, like citations or references.
3. Interaction (session context)
4. Social, systemic, media, work task, conceptual, emotional contexts
5. Economic techno-physical and societal contexts (infra-structure)
6. The *historic context* operates across the context stratification as an additional dimension. This temporal form of context influences all IS&R processes.

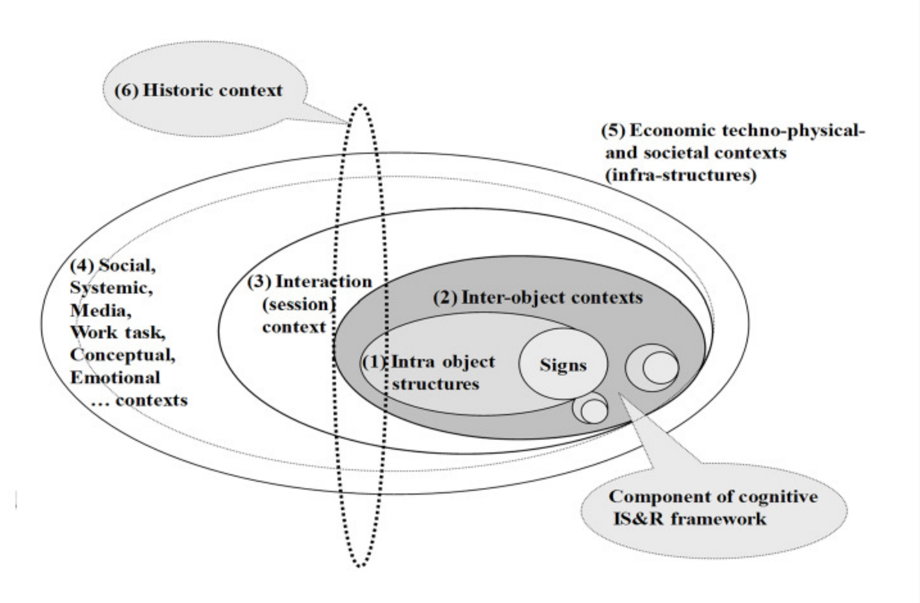

(6) Historic context
(5) Economic techno-physical- and societal contexts (infra-structures)
(4) Social, Systemic, Media, Work task, Conceptual, Emotional … contexts
(3) Interaction (session) context
(2) Inter-object contexts
(1) Intra object structures
Signs
Component of cognitive IS&R framework

Fig. 1. Nested model of context stratification for IS&R. Source: Ingwersen and Järvelin (2005, p. 281).

In order to exemplify the potential influence of the historic context dimension, the following section takes a historic view on the museum domain.

4 The historic context dimension: from the order of classification to cultural complexity

The museum as an institution has undergone a long development reaching from the earliest museums of private collections to present day's rich variety of museums to be found in castles, boats, farms, dungeons etc. as well as in more traditional museum buildings. These changes are reflected in the museum's changing purpose and role through time and not least in relation to the role of the visitor. In this light, it is interesting to consider a historic perspective on knowledge organisation and classification of museum objects. A historic perspective provides valuable insight as it offers a broader context for understanding today's knowledge organisation and indexing in museums. At the same time, since many museums possess collections assembled and described many years ago, the reminiscences are still visible and relevant in today's work with cultural heritage collections.

In her book on museums and the shaping of knowledge, Hooper-Greenhill (1992) illustrates how interpretation of museum objects is shaped by the historic and cultural context within which they are displayed. Hooper-Greenhill's analytical work explicitly addresses elements of knowledge organisation and indexing. Based on Foucault's (1970; 1974) theories on three major *epistemes*[2] Hooper-Greenhill applies a long-term historic view. She describes how there have been radical shifts in what meanings have been made, how collections have been put together, and how theses collections, once constituted, have been used. *The renaissance episteme* covered the early years of museums and was characterised by a complexity of interpretation and similitude, and things being read for their hidden relationships to each other. Resemblance was never stable and consisted of endless relationships. There was, therefore, "… no real substance, and no means of verification. Legend, stories, and material things all offered possibilities for discovering likenesses and relationships. None could be discarded, as all were potentially 'true' " (Hooper-Greenhill, 1992, p. 15). *The classical episteme's* founding structure was that of order, through measurement and the drawing-up of hierarchical series. A two-dimensional table of classification emerged, and all natural things were arranged and grouped into families on the basis of their visible features. A thing became an object through its visible features. In the *modern*

[2] Hooper-Greenhill analytical work is based on Foucault's (1970; 1974) theories on three major *epistemes*: the renaissance, the classical and the modern episteme. Foucault defines episteme as the unconscious, but positive and productive set of relations within which knowledge is produced and rationality defined (Foucault, 1974, p. 191).

episteme objects are constituted through organic, historic links, through stories, and trough people: "In the modern age, knowledge is no longer shaped by the secret, enclosed, circulating structures of the Renaissance *episteme*; nor by the flat, classificatory table of difference of the classical *episteme*; now knowledge is structured through a three-dimensional, holistic experience which is defined through its relationship to people" (Hooper-Greenhill, 1992, p. 214). Or in other words, the *context* of the artefact started to become important.

Hooper-Greenhill's analysis of the history of museums shows that the principles of selection, putting together a collection,- and classification have radically changed over the years. In 1992 Hooper-Greenhill, building on Foucault (1970), wrote that the end of the modern age was close to an end signalling the final of the modern museum. Ten to fifteen years later, Anderson (2004) and Cameron and Kenderdine (2007) do not hesitate to call it a paradigmatic shift, and Anderson describes and characterises the reinvented museum. In the reinvented museum the human, social, and cultural context of the museum artefact is central, leading to multiple viewpoints and potentially numerous perspectives (Anderson, 2004). Further, the reinvented museum does not represent a voice of authority but aims instead at being responsive to visitor needs and two-way communication. One may also say that the context of the visitor likewise becomes important, and that the museum is audience focused instead of collection driven. This process of rethinking the museum – reinventing the museum – can be said to symbolize the "…general movement of dismantling the museum as an ivory tower of exclusivity and toward the construction of a more socially responsive cultural institution in service to the public" (Anderson, 2004, p. 1).

Knowledge of the shift from the traditional museum to the reinvented museum is important, also from an IS&R perspective, since the shift has influenced the area of study. Specifically, three areas should be pointed to. Firstly, the shift has led to changing perspectives on description and indexing of museum objects: "In the past, the object on display was accompanied with a label that fixed it in a monolinear frame of reference. A chair was 'Oak, Seventeenth Century'; a gun was identified by its firing capacity; […]. The human, social, and cultural contexts of theses artefacts were rendered invisible by these strategies. Now the many frames of reference that can contextualise material things are displayed along with the things themselves" (Hooper-Greenhill, 1992, p. 204-205). This quote illustrates a development in museum exhibitions closely connected to the shift from the traditional museum to the reinvented museum outlined above. In the reinvented museum artefacts are contextualised by human, social, and cultural references and these contextualised references must be captured in the registration and indexing process. The quote also illustrates a change from an order of classification to cultural complexity. Evidence of this change is reflected in the work tasks of the museum professionals today (Skov, 2009) and shows how the historic context dimension influences, e.g., the organisational context (in context dimension four).

Secondly, the role of the museum visitor has changed and the information seeking behaviour of today's virtual museum visitors must be viewed in this perspective. Hooper-Greenhill describes how, at the birth of the public museum, a division was drawn "…between the private space where the curator, as expert, produced knowledge (exhibitions, catalogues, lectures) and the public space where the visitor consumed those appropriately presented products" (1992, p.200). However, this cleft between the museum professional and the visitor has started to close. The opening process can be seen quite literally when visitors are invited behind the scenes on 'open days' and also in relation to the concept of open storage. Less literally but equally important, are efforts to support direct visitor participation and involvement (e.g., Black, 2005) along with the shift from viewing museums as collection-driven institutions to viewing them as visitor-centred (Anderson, 2004).

Finally, new technology can be seen to play an important role in the reinvented museum. New technology provides new ways of communication and interaction with visitors. It is being used to reach new audiences; approach specific groups of users via tailored communication and as one way to fragment the meaning of the artefact by introducing many perspectives, voices, and points of view. Museums have moved a long way from the birth of the public museum to today's reinvented museum.

5 Concluding remark

Ingwersen and Järvelin (2005) introduce a historic context dimension in their "Nested model of context stratification for IS&R". This paper aimed at illustrating the relevance of the historic context dimension using the museum domain as an example. An analysis of the shift from the traditional museum to the reinvented museum through the perspective of a historic context shows, how especially three areas related to IS&R are influenced. Firstly, the shift has led to changing perspectives on description and indexing of museum objects, which relates to both the inter-object context (context dimension two) and museum employees' work tasks (context dimension four). Secondly, the role of the museum visitor has changed and thus the information seeking behaviour of today's virtual museum visitor. This primarily interrelated with how museum visitors interact with an online museum system (context dimension three). Thirdly, new technology in museums provides new ways of communicating and interacting with visitors (context dimension four). Summing up, this paper illustrates how the historic context drives and shapes the current IS&R situation.

References

Allen, B. L. (1997). Information needs: A person-in-situation approach. In P. Vakkari, R. Savolainen & B. Dervin (Eds.), *Information seeking in context* (pp. 111-122). London: Taylor Graham.

Anderson, G. (2004). *Reinventing the museum: Historical and contemporary perspectives on the paradigm shift*. Lanham: Altamira Pr.

Black, G. (2005). *The engaging museum: Developing museums for visitor involvement*. London: Routledge.

Cool, C., & Spink, A. (2002). Issues of context in information retrieval (IR): An introduction to the special issue. *Information Processing & Management, 38*, 605-611.

Dervin, B. (1997). Given a context by any other name: Methodological tools for taming the unruly beast. In P. Vakkari, R. Savolainen & B. Dervin (Eds.), *Information seeking in context: proceedings of an international conference of research in information needs, seeking and use in different contexts. Tampere, Finland, 1997* (pp. 13-38). London: Taylor Graham Publishing.

Foucault, M. (1970). The order of things: An archaeology of the human sciences.

Foucault, M. (1974). *The archaeology of knowledge*. London: Tavistock Publications.

Hooper-Greenhill, E. (1992). *Museums and the shaping of knowledge*. London: Routledge.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Berlin, Germany: Springer.

Talja, S., Keso, H., & Pietiläinen, T. (1999). The production of 'context' in information seeking research: A metatheoretical view. *Information Processing & Management, 35*, 751-763.

Skov, M. (2009). The reinvented museum: Exploring information seeking behaviour in a digital museum context. Copenhagen: Royal School of Library and Information Science. XII, 312 p.

Wilson, T. D. (1999). Exploring models of information behaviour: The 'uncertainty' project. *Information Processing and Management, 35*(6), 839-849.

Addresses of congratulating authors:

Mette Skov
Royal School of Library and Information Science
Fredrik Bajers Vej 7K, DK-9220 Aalborg East, Denmark
Email: ms[at]iva.dk

Brian Kirkegaard Lunn
Royal School of Library and Information Science
Fredrik Bajers Vej 7K, DK-9220 Aalborg East, Denmark
Email: bki[at]iva.dk

Beloved Mentor of New Generation of Scholars

Peiling Wang[1] **& Iris Xie**[2]

[1] University of Tennessee Knoxville, Knoxville, USA
[2] University of Wisconsin-Milwaukee, Milwaukee, USA

Peter's Guidance and Support

Peter is an internationally renowned researcher in the field of information science. His contributions to the field go beyond his prolific research outputs and leadership in building research communities. As one of the first generation of pioneers in cognitive IR interaction, Peter has generously mentored and supported many budding researchers formally and informally. Like most mentors, Peter attends his protégés' presentations, provides critical comments, reviews their draft proposals and manuscripts, co-chairs program committees, co-authors publications, co-lectures at courses and workshops, and serves as an external reviewer for tenure and promotion. What is most significant to the new generation of researchers and well appreciated by many is that we found our works are cited in Peter's books, chapters, and papers -- not just as a passing note rather as a part of his synthesizing the current research. As a luminary of the field, his willingness to pay attention to tyros' publications and promote their good works is truly exemplary given the long-observed phenomenon of scholarship: "a junior citing upward, but being ignored by the (cited) grandees" (Cronin & Shaw, Identity-creators and image-makers, *Scientometrics*, 2002, p. 44).

Over the years, Peter spent considerable time and efforts, carved from his busy schedules, evaluating and writing external reviews for tenure and promotion dossiers of faculty members outside of Denmark. We now also frequently chair or serve our own schools' tenure and promotion committees as an obligation, or serve as external reviewers of other universities per invitations by the directors or deans of the faculty members being considered for tenure or promotion. It is extremely difficult to find qualified, willing external reviewers as the time and work of the external reviewers are mostly unrecognized and unrewarded. Typically, universities do not share specifics of these external reviews with the candidates, nor do they reveal the names of the external reviewers being invited. Sometimes, the reviewers are given a very tight deadline. Often, external reviewers are not notified of the tenure and promotion decisions. At the best, the external reviewers would get a thank-you note upon receiving the review letters. Being senior faculty members involved in tenure and promotion decisions, we now have the responsibility to read and discuss the external assessment letters as a part of the process. In comparison, Peter's review letters are uniquely thorough and in-depth. It is obvious that he does tenure and promotion evaluations as seriously as his own research projects. His evaluations include literature review, methodological design, presentation of results, and discussion. Thus, his letters show clearly that he has read the works closely by the candidate and his assessment of the candidate's contribution to the field is grounded in the current literature. Another uniqueness of his assessment is that Peter not only provides qualitative evaluation of the candidate's works, but also conducts his own citation analysis of these works to make quantitative evaluation of the candidate's influence in the field. In some cases, he even applied new citation analysis methods in assessing the candidate's works as appropriate.

Kudos to Peter for his caring support of young scholars!

Peiling's Tales of Peter

I first read Peter's many influential papers and the 1992 book (now free access at http://vip.db.dk/pi/iri/index.htm), when I was a doctoral student at the University of Maryland about twenty-three years ago. My professors chose his works as required readings because of the importance of his research areas, and the breath and depth of his research. At first, I was totally intimated by these articles as a new doctoral student from another field in a new country. Gradually, I experienced many Gestalt *aha!*-moments in understanding the cognitive view of IR and found myself drawn to the area of cognitive information behaviors. Peter's work has made a significant impact on my research and academic career. He pioneered a user-centered theoretical framework with innovative empirical design to observe users in real information-seeking settings. He is the driving force to integrate two important but divided research domains: system-oriented IR and use-centered IR interaction. My doctoral research observed real academic users' relevance judgment and information use during a research project following Peter's cognitive approach.

My opportunity to finally meet Peter came during the 1995 ASIS Annual Meeting. We had a brief encounter through my ASIS Doctoral Forum mentor professor Nick Belkin. In the following year, 1996, Peter came to my presentation at the first ISIC conference, Tampere, Finland. I was so thrilled that he would come to hear a novice researcher; but I was more nervous and worried about my ability to answer his questions as I had seen him raising tough and challenging questions in earlier sessions. However, I soon was at ease and assured that he thought my research was solid. He congratulated me for having selected an important research area with a sound methodology and suggested further analyses I could and should do with the collected data to generate additional results.

Since then, Peter has become my important mentor. In the following years, at various conferences, seminars, workshops, and Nordic Doctoral Research courses, I had many opportunities to meet and work with him. I witnessed firsthand his razor-sharp intellect and wisdoms. I learned a great deal of his effective pedagogy. My experience in the teaching setting with him as a listener or a speaker has been without parallel. His passion for the material and dynamic delivery made long lectures a truly joyous tour of knowledge wonderland. He has a great sense of humor. Peter was our 2009 Lazerow Memorial Lecture speaker. Just when he started his lecture, the slides on the screen disappeared. After he paused a few seconds, he figured out how to bring back the LCD. "Aha, that [touch screen] went to sleep." He charmed his audience: "please don't get too much into sleep, because I cannot touch you all." He then moved on with the lecture as the audience burst out laughing.

Peter always provides frameworks and models as basis to design his teaching materials and conduct research. In teaching he not only explains new concepts crystal clear, but also foresees the various difficulties students may have. His interest in students' learning and attention to their cognitive needs are evident in his scaffolding method that accommodates diverse students' backgrounds without simplifying the material. I have since adopted the scaffold strategy and modeled my teaching largely after the techniques I saw in his teaching with a great success.

My recent collaboration with Peter on the topic adaptivity in IR interactions has taught me how to effectively bring people together. Peter initiated the collaborative work, quickly built a team of mixed backgrounds, and integrated seemingly unrelated ideas. At one point, I was really worried if this team could get anywhere within a very tight deadline. Peter led the team gracefully through the difficult process. In the end, we were very proud of the product; mostly we have all enjoyed working together under Peter's leadership. I will definitely adopt his well-structured managerial style, his openness to arguments and debates, and his flexibility in adapting and making changes, even it means to throw out a chunk of the produced material.

It will not be an overstatement that Peter has been the single greatest influence on my research and teaching. He exemplifies the researchers of higher caliber, the teachers of high standards, and the mentors of inspiration to which I strive to hold myself as a researcher, an educator, and a mentor. I will always remember his kindness, encouragement, and gentle guidance.

On a personal level, first, I had the opportunity to visit his home and meet his auntie during my lecturing at the Nordic Doctoral Research Courses in 2003. Peter invited me and his advisees to his Malmø home. Sitting at the big desk in his study with his advisees, we had a greatest time talking and laughing. The study is located in a quiet spot with the walls occupied by bookshelves and quite a few archaeological artifacts. I truly enjoyed Irene's and his hospitality. One day after the course, Peter took Pia and me to visit his auntie Henni who treated us with tea and cookies. Henni was battling cancer with a steadfast will which I also saw in my late mother who battled and died of cancer in 2000. Peter updated me often on Hennie's conditions. I had hoped to visit Henni again to tell her how much I admired her strength, but she passed away on November 23, 2005, at the age slightly shy of 87. Second, during their 2003 trip to China, Peter and Irene met my family. It was quite a new experience for the family to host foreign guests. I was so amazed about how they communicated well with my father in spite of the language barriers. My father who passed away in 2006 was very fond of them and talked about their visit often whenever I spent time with him.

Iris's Tales of Peter

Just like Peiling, Peter's prominent papers and books have significantly influenced my own research interests and academic career. Although I had the opportunities to attend his presentations and meet him at conferences, I did not have close personal interactions with Peter until 2006. When I was writing my book "Interactive Information Retrieval in Digital Environments" during 2006 and 2007, I planned to include some key macro and micro models of interactive information retrieval into the book. After careful research, I sent out several email requests to the authors asking for their electronic versions of the models. Peter not only was the first author to send his model within a few hours, but also showed his care and support towards a young scholar's work in his email reply. He provided PowerPoint version of the models so I could easily edit, copy and paste into my text. He went a step further to discuss the copyright issues of the models. I kept a copy of his email and considered it a vivid example of his genuine heart and warm style, and an appropriate insert for this Festschrift article:

> Dear Iris, I know who you are both personally and from the research literature. I am delighted that you can make use of our models and I enclose the required ones as a PowerPoint presentation with three slides. Then it is easier to edit (if you wish) and copy as an image to your manuscript. By the Way: I hold myself the copyright for the 1992 book, Information Retrieval Interaction. So you have my permission to use the figure in your work, provided that you refer back to the monograph, as you have done in your letter of request below. I really appreciate it - I send you my best wishes for your work - Yours Peter Ingwersen

His email did make a difference in enlightening my writing process while I was quite stressed out and frustrated during that time.

When I had problems in obtaining copyright of his models from the publisher, my publisher was not able to offer helpful suggestions to solve the problem. After learning that, Peter sent a revised model and offered detailed instruction about how to modify another model as well as how to present the copyright information in the book. These instructions help me greatly prepare for the copyright information for Peter's models as well as others' models included in my book.

Peter also suggested a meeting at ASIST conference in Milwaukee to further discuss these models and any questions that I might have. Before our meeting, I was wondering whether he was an approachable person just like his emails. One day before our meeting, Peter spotted me talking with my colleague in the lobby of the conference hotel; he came to introduce himself. His smile and easy-going manner made me feel relaxed and comfortable. During our meeting when we chatted, I felt as if I was talking to a longtime friend. In addition to the discussion of his own models, Peter reviewed my "planned and situational model" and provided insightful comments and constructive suggestions. Adopting Peter's suggestions, I made appropriate changes to my model. Moreover, Peter also revealed background information of some of the related interactive IR models as well as their limitations. Of course, our conversations went beyond research topics of our field; Peter also shared his unique thoughts on global warming and other social issues.

In a word, Peter is the role model for me to pursue my research interests, teach and mentor my students, and interact with my colleagues.

Peter's Footprints in China

As a scholar with international reputation, he also spent extensive time in presenting his research to researchers and students around the world. The year 2003 had a disastrous beginning when the SARS epidemic was spreading in Asia and made daily world news headlines. By summer, SARS was under control with only occasional cases that were managed effectively. In China, it was an unusually quiet period when almost no foreigners were seen on the usually crowded busy streets of Shanghai. Peter had agreed to give lectures during his trip to Beijing for ISSI conference arranged before the SARS epidemic. The organizers were concerned and suggested to postpone the seminar because many international events to be held in Asia were either canceled or moved to a safer place during that period. Peter responded that nothing would stop him unless the travel restriction was not lifted. He then came to China, with his wife Irene, among the few brave souls. He and Irene honored their commitments and delivered their lectures in Shanghai (August 12-17) and in Taiwan (August 18-23). Then they went Beijing for ISSI conference (August 24-29). The hosts were really proud of Peter and Irene, and the audiences were truly inspired. These lectures were filled with laughter and interactions! Many years later, whenever we visited Shanghai and Taiwan, we often ran into the people who had met Peter formally and informally and spoke highly about his lectures and his unique charisma.

Since his first trip to China in 2003, Peter has returned to China a few more times with his wife Irene giving lectures and serving Advisory Board of the International Collaborative Academy of Library and Information Science (ICALIS), Wuhan University. On his way to SIGIR conference in Singapore in 2008, Peter volunteered to present at five universities in China including: East China Normal University in Shanghai, College of Computer and Information Science at South West University in Chongqing, Information Management School, Wuhan University, Henan Normal University, and Institute of Scientific and Technical Information of China. Before his trip to China, Sichuan province was hit by an 8.0 Ms earthquake at a site not far from Chongqing, where one of his destinations, the South West University, is located. In his email, Peter was very worried about the colleagues there. Without being intimated by the post earthquake problems, Peter, along with Irene, decided to travel as planned to visit and give lectures at the South West University. The faulty members and students were very much touched and truly grateful.

Instead of preparing one lecture for all the universities on this trip, Peter prepared more than five different presentations tailored to the specific audiences. His topics ranged from "Research on Information Retrieval in Context," "Information Science in International Perspective" to "Science Indicators for Research Evaluation." His presentations received excellent feedback and high remarks. The audiences learned the most updated research as well as his own research of 40 years in information science. During his visit, he was elected to the International Advisory Board of the School of Information Management in Wuhan University. In addition to lectures, Peter spent substantial time with scholars and students in these universities to hear and discuss about their research topics and projects, teaching, and learning. To many young scholars and students, Peter offered suggestions for their career development.

Peter loves China. China has been on his radar screen. Whenever a natural disaster or problem occurred in China, we will get Peter's messages asking about our families and colleagues there.

At Peter's Retirement

Dear Peter:

As you said, life is multidimensional and not all about work. As you retire from your current position, we wish you the best in your new adventures and enjoy your many other talents: singing operas, competing in tennis matches, and traveling around the world.

Congratulations on being the inaugural Professor Emeritus in the newly named Det Informationsvidenskabelige Akademi! We hope you continue to mentor a new generation of scholars and contribute to and influence the field by occasional teaching and research involvement. We would love to hear from you often on your thoughts and advices.

We wish to see you soon and often.

“We only part to meet again.” -- John Gay, English Poet, 1685-1732

Addresses of congratulating authors:

Peiling Wang
School of Information Sciences, University of Tennessee Knoxville
1345 Circle Park Drive, Knxoville, TN 37996, USA
Email: peilingw[at]utk.edu

Iris Xie
School of Information Studies, University of Wisconsin-Milwaukee
3210 N Maryland Ave, Milwaukee, WI 53211, USA
Email: hiris[at]uwm.edu

Contextual Perspectives in Knowledge Management and Information Retrieval

Gunilla Widén

Åbo Akademi, Åbo, Finland

Abstract: Knowledge work is increasingly important in today's organizations and managing knowledge as a resource is underlined. In Knowledge Management research contextual aspects are often focused. The aim in this paper is therefore to discuss contextual perspectives in KM and their possible contribution to the holistic perspective of information seeking, retrieval, and use. There are at least three important contextual perspectives that affect information activities in an organization. These are the information culture, the social and the electronic environment. Research projects looking closely at the social context (e.g. social capital) and IR-engineering would be important in the future. The digital arena with social and interactive tools as represented in the Web 2.0 environment would be interesting to integrate to the discussion of effective knowledge systems.

Keywords: Knowledge Management, Information Retrieval, Contextual perspective, Social Capital

1. Introduction and background

"Modern work is increasingly knowledge work. Access to recorded information or human sources is essential." [1, p 323]. With this statement in mind it is obvious that knowledge is a key resource today and the manageability of knowledge continues to be interesting. Since the 1990s knowledge as a resource has been underlined and discussed. Knowledge is seen as a critical factor for organisational performance, a driving force of creativity and innovations, and a key resource for intelligent activities [2 , 3].

Access to information and knowledge is a key to effective management. Also sharing and use of the accessed knowledge demands its own capabilities and skills. In organizations these capabilities are usually collected to Knowledge Management (KM) activities. There are many possible aspects involved in KM and many possible ways to define the concept. In KM-research information activities are often anchored in a social context on an organisational level. This is also what Ingwersen & Järvelin [1] underline in their book The Turn where they point out the need for integrating different aspects and viewpoints of information seeking, retrieval, and use. An important direction of this integrated view is context which has also been an important element in different information behaviour models underlining social and environmental aspects [4 , 5].

While context and environmental elements are so important in KM research the aim in this paper is to discuss how contextual perspectives in KM *can contribute to the integrated view of information seeking and information retrieval.* To discuss this aspect KM as a process is first presented, then the contextual aspects of KM, the social and environmental challenges are discussed. Finally a reflection of how these aspects/insights can contribute to the integrated view of IS/IR is put forward.

2. Knowledge Management

Knowledge Management is a process that supports companies to retrieve, organize, store, and share information and expertise which they need for different kinds of business activities [3]. This is a challenge because knowledge is an intangible asset and it is difficult to exactly define knowledge. Ingwersen [6] defines knowledge as internalized information and often understood as a blend of explicit and tacit elements [7 , 8] which means that there are many types of knowledge at different levels of the firm. This view is often the starting point when discussing KM where the fact that information and knowledge for an organisation is highly specific is underlined. Every organisation must define information and knowledge in the light of their activities and goals [9, p 119]. The importance of the contextual perspective is focused where human and social aspects are stressed in combination with knowledge organisation and contents.

KM often involves different perspectives such as information technology (IT), information systems, processes, people and contents of which information behaviour, information seeking and information retrieval are important parts. It is clear that with that many perspectives it is challenging to develop effective KM in an organisation. The aim is to develop an integrated view where knowledge, people, processes and IT form a manageable entity [9, p 119]. Organisational knowledge management means supporting people so that they can use and share what they know and in the end support strategic information management. These processes should be connected to the human aspects, communication, learning, networking, and the social environment, in order to achieve effective information sharing and utilisation [10].

The ongoing development of technology is also changing the nature of KM all the time. This is one of the most rapidly changing aspects of KM. Traditionally, it has been stressed that IT serves as an important infrastructure helping staff to perform better. Technology in connection to KM has often been about system analysis and design [11]. But besides the fact that technology is an important infrastructure, the digitalisation of information creates new possibilities for collecting, distributing, and presenting information. This also means that technological and social forces are interwoven. Networking and digitalization are facilitating the creation of an entirely new paradigm of KM [12]. The development of IT-based information activities in an organisation such as IR is a highly important part of KM.

3. KM in context

The growing interest in managing people and their information interactions puts new demands on the insights into the social environment that constitutes the platform for human action. To study the information sharing process as such does not give the more complex insights into e.g. the motives for sharing information for a specific purpose. In recent years, many studies have actually emphasised the information seekers in their social and cultural context, i.e. information practices [13 , 14 , 15 , 16].

Information and knowledge in any organisation is individual and specific for each environment. KM is contextualised by the organisation, and includes information as an object and a user construct [17]. It is important for organisations to understand what the drivers for KM are in their particular environment while there are no standardized rules for implementing KM in organisations [17 , 18]. Lately the web environment has promoted many new challenges and possibilities to organisational KM and will therefore be discussed here as well.

3.1. Information culture and communication climate

The actual information use in the workplace is shaped by its environment, which is built of institutional, organisational, and personal elements [19 , 20 , 21 , 22]. It has been shown that information, as a resource in an organisation should be supported by an open and active information culture. There are different kinds of information cultures in organisations and it is important to know what kind of culture exists in the company in order to adjust the information work and planning accordingly. Open, changeable companies can more easily shape an active information culture. This is so because the integration of processes and functions are working well. That means that the planning of the processes, in particular, is strongly integrated and a company is able to create overall solutions for the whole company. Because of the integration of the processes, a more active communication between the units is created [22].

Overcoming the cultural barriers to sharing information and knowledge has more to do with how one designs and implements one's management efforts into the culture than with changing the culture [19]. The visible dimensions of culture such as values and missions, structures and stories, should be balanced with the invisible dimensions of culture such as unspoken sets of core values. This means that the awareness of cultural dimensions and widely-held core values helps to link knowledge sharing efforts with the common interest. Visible connections between knowledge sharing and practical goals are then possible to make and the holistic view of persons, working closely, techniques, and technology is important [19 , 23 , 24].

3.2. Social aspects

Closely to cultural elements are the social aspects while social interaction forms a part of organizational culture. However, the focus is different underlining how human interaction and networking work in a particular organization. It has been shown that the social capital framework is a usable navigation tool in picturing the information and knowledge sharing processes. Social capital is something that is part of every organization and group but that has not always received attention for effective management. The theories of social capital are related to the key contemporary phenomena and success factors, such as uncertainty, knowledge creation, and innovativeness. Using the social capital framework, it is possible to reveal what motivates employees to share information [25] and open up the understanding of how organizational information profiles are built and how they function. Social capital is a context of dynamic environments for information sharing, continuously changing, but following patterns and rhythms. It opens up relational, structural, and content dimensions related to sharing. It gives an overall picture to organizational and collaborative action and improves the understanding of the role of knowledge sharing in diverse institutional contexts and results in better information control [26 , 27].

The interaction between organisational and individual aspects must be focused. Information sharing arises from a constant mix of organisational and individual motives, and factors like purpose, timing, and availability play an important role as enablers and barriers to sharing [28 , 29]. It is important to highlight the relationship between information sharing and work role or work tasks in this context [e.g. 30]. Through work tasks it is possible to define a closer context of organisational information behaviour. The task goals, processes and information seeking are bound together and explain patterns for information seeking and use.

Social aspects of knowledge sharing are, however, not easy and straightforward to manage. There is a wide range of specific social factors that influence people's willingness to share knowledge. Among other aspects, it has been acknowledged that interpersonal trust and commitment to the group or organization is of importance [31 , 32].

3.3. Digital context / E-context

As Ingwersen & Järvelin states there are new challenges to IR engineering while means of access and sources are increasingly becoming electronically networked and formalized in systems. "This integration of e-generation, e-access, and e-use makes IR engineering complex – but not unmanageable. The question for IR engineering is: which additional variables from the immediate contexts does one wish to include in a controlled relationship with one another." [1, p 323]

The digital world is a growing area and many organisations are active within both a digital and a traditional framework, e.g. using digital libraries. Many organisations are also completely situated in a digital environment such as e-businesses. Regardless of whether the organisation is a combination of virtual and traditional settings or completely an e-organisation, the electronic environment puts special demands on the management of organisational processes such as KM. The digital environment brings both possibilities and challenges to this field.

E-businesses are often more complex than traditional businesses. KM can be used as a tool to overcome this complexity through managing the knowledge base and making relevant knowledge accessible. E-businesses have often more knowledge to manage and leverage than traditional businesses. This is because of the expansion of reach and expansion of customers and suppliers. In the e-business environment, technology can generate much information that develops into knowledge (e.g. customers' search behaviour). It becomes important with KM initiatives to ensure that the information that is generated by Internet browsers can be maintained and used effectively [18]. Collaboration is also an important feature of KM in e-business environments demanding platforms for collaboration and knowledge sharing across geographical and organisational boundaries [33]. According to a study by Plessis and Boon [34], KM is important in e-businesses because it increases efficiency by providing 24-hour access to knowledge. KM activities facilitate integration between disparate groups or departments within an e-business. The customer and supplier base increases quickly, leading to much knowledge to manage and a need for KM systems (i.e. technology).

Web 2.0 is an emerging part of managing organizational knowledge. Web 2.0 tools can be seen as social and interactive, supporting knowledge sharing and the development of social capital. Here both the digital and social contexts are combined. The dimensions of social capital are visible in social software as well as in advancing the structural, relational, and content dimensions. Managing social capital for effective knowledge sharing is a complex process, but Web 2.0 lends some concrete support in this organizational challenge. The social and interactive web brings new ways to support KM, a new way of organizing knowledge and knowledge sharing, a new context to take into consideration also in IR-engineering. There is evidently a need for an interactive discussion where social capital aspects, knowledge sharing theories, and Web 2.0 techniques are combined.

4. Social and organisational environment

How can these contextual aspects of KM research contribute to IR-area? It has been stated that what constitutes information and knowledge for any organisation is highly specific. Every organisation needs to define it in the light of what it wants to achieve [9 , 18]. There are at least **three important perspectives** that make the KM different in different environments. These are the information culture, the social aspects as an emerging perspective, and the e-context.

1. There is a need to define each organisation's information culture to shape the information activities such as the development of IR-systems accordingly. The importance of the internal environment and a communication strategy and open information culture in implementing and committing to KM has been underlined [22 , 35]. An active information culture can have many advantages. Communication of information becomes more effective in the company, also affecting customer relations [22].
2. Social dimensions affecting the management of information and knowledge is a growing area of interest, and aspects like group identity and social capital are fields that are emerging aspects in KM initiatives. This is a local and complex process, which means that it is difficult to find a common practice in managing it [26]. Social capital is a variety of different entities in the social structure that facilitates the certain action of the individuals who are in the structure [36]. Similarly as in the case with KM, the entities of social capital are different in different environments and it is important to picture the social practices from many angles and in many contexts. When defining organisations through the social capital dimensions, it could be easier to manage the knowledge if it is known if the structural or relational dimensions are strongly emphasised in the organisation.
3. Although the digital or electronic environment is seldom totally separate from the corporate or public organisation, it poses specific challenges to KM. It is important to integrate the KM-initiatives and e-environment initiatives into a coherent picture. Managing information and knowledge in any organisation today involves e-perspectives. KM in digital environments are often connected to technological aspects although KM in digital environments is not only about the most suitable information and knowledge systems. The great amount of information to be processed and a single platform giving access to this amount of information and the multiple partners in the organisation are important differences from traditional environments. Here Web 2.0 brings together social and interactive aspects in a virtual environment.

Additionally it is relevant to reflect on the relation between the organisation and the external environment and information about the environment must be retrieved in order to adjust the organisational activities effectively. Many studies have investigated managerial information seeking skills and a lot of work in this area is still to be done.

5. Concluding remarks

Information seekers are cognitive actors in a social, organisational and cultural context [1, p 261]. These aspects influence the activities of each individual in their different activities connected to managing information and knowledge, such as information retrieval, seeking and use. There is a constant development of the information environment, always generating new demands on KM, systems development, and IR-engineering.

We have seen that KM activities must be in line with both the external environment as well as the internal organisational environment. Special characteristics in the external environment in combination with the complex reality that is represented in the internal environment are underlined. A growing understanding of networks, motives for sharing, learning aspects are important. *The awareness of the cultural dimension, what constitutes the organizational information culture, is crucial to define in order to develop all information activities (e.g. IR) accordingly.*

The organizational context is changing all the time, there are new technological innovations, a changing external environment, and a continuous mix of social relations affecting all organizational activities. To fully understand and utilize the knowledge processes in any organization, the focus should be on social and socio-technological conditions interacting with the knowledge processes [37]. Knowledge work is seen as supportive of social capital [38]. Positive effects of having a high level of social capital is the availability of intellectual and knowledge resources through networks and through the relationships between individuals and social units. However, managing social capital and knowledge processes are challenging areas. One of the problems knowledge work and management faces is the difficulty in getting people to share their tacit knowledge. Access to what people know is not always depending on an effective IR-system. However, it is an important part of it and it is important to know the information environment to adjust IR-work to support the whole information and knowledge base of an organization. *Through the social capital perspective we can focus on structural, relational and content dimensions and define the role of IR-engineering to support these dimensions of knowledge sharing.*

Many tools can be used to support this holistic approach. Web 2.0 is one important development where the social and interactive techniques support the efforts to bridge different knowledge sharing processes. Developing the social capital (namely, networks, relations, and cognition) of an organization may well be supported by Web 2.0 tools which enable knowledge processes to function in and between individuals, different units, and social networks. Furthermore, Web 2.0 techniques afford the possibility of increasing one's knowledge base, exploring one's thoughts, supporting and disproving one's ideas. A new culture of voluntary, contributive, and collaborative participation is emerging [39]. Web 2.0 techniques demand a huge amount of motivation from the individual to be able to adapt to the interactive tools. Trust is an important enabler to motivating and using social technologies. *This kind of participatory context, where interactivity and information production are underlined, is what users expect today. This is a challenge and delivers also new possibilities for IR-engineering.*

There are many possible angles to focus in research, combining KM-perspectives and IR-systems development. Research projects looking closely at the social context (e.g. social capital) and IR-engineering would be important in the future. The digital arena with social and interactive tools as represented in the Web 2.0 environment would be interesting to integrate to the discussion of effective knowledge systems.

References

[1] Ingwersen, P. Järvelin, K. The turn : integration of information seeking and retrieval in context*,* Springer, Dordrecht (2005)

[2] Wiig, K.M. Knowledge Management: where did it come from and where will it go? Expert Systems With Applications 13 (1) 1-14 (1997)

[3] Gupta, B.; Iyer, L.S.Aronson, J.E. Knowledge management: practices and challenges. Industrial Management & Data Systems 100 (1) 17-21 (2000)

[4] Ingwersen, P. Cognitive perspectives of information retrieval interaction. Journal of Documentation 52 (1) 3-50 (1996)

[5] Wilson, T.D. Models in information behaviour research. Journal of Documentation 55 (3) 249-270 (1999)

[6] Ingwersen, P. Information retrieval and interaction*,* Tyler, London (1992)

[7] Nonaka, I. Dynamic theory of organisational knowledge creation. Organization Science 5 (1) 14-37 (1994)

[8] Polanyi, M. Personal knowledge: towards a post-critical philosophy*,* University of Chicago Press, Chicago (1958)

[9] Orna, E. Information strategy in practice*,* Gower, Aldershot (2004)

[10] Widén-Wulff, G., et al. Knowledge Management and Information Management. In LIS Education in Europe: Joint Curriculum Development and Bologna Perspectives*,* Kajberg, L. Lörring, L., Eds.(121-132). The Royal School of Library and Information Science, Copenhagen (2005)

[11] Srikantaiah, T.K. Knowledge Management: a faceted overview. In Knowledge Management for the information professional*,* Srikantaiah, T.K. Koenig, M., Eds.(1-17). Information Toady, Medford, NJ (2000)

[12] Pottenger, W.M.; Callahan, M.R.Padgett, M.A. Distributed Information Management. Annual Review of Information Science and Technology 35 79-113 (2001)

[13] Solomon, P. (ed. Information mosaics: patterns of action that structure. Exploring the contexts of information behaviour. Proceedings of the 2nd International Conference on Research in Information Needs, Seeking, and Use in Different Contexts, ed. Wilson, T.D. Allen, D.K. Taylor Graham, London, 1999 (1999)

[14] Hall, H. Borrowed theory: applying exchange theories in information science research. Library & Information Science Research 25 287-306 (2003)

[15] McKenzie, P.J. A model of information practices in accounts of everyday-life information seeking. Journal of Documentation 59 (1) 19-40 (2003)

[16] Hyldegård, J. Collaborative information behaviour: exploring Kuhlthau's information search process model in a group-based educational setting. Information Processing & Management 42 (1) 276-298 (2006)

[17] Kirk, J. Information in organizations: directions for information mangement. In Introducing Information Management: an Information Research Reader, Maceviciute, E. Wilson, T.D., Eds.(3-17). Facet, London (2005)

[18] Plessis, M.d. Drivers of knowledge management in the corporate environment. International journal of information management 25 193-202 (2005)

[19] McDermott, R. O'Dell, C. Overcoming cultural barriers to knowledge sharing. Journal of Knowledge Management 5 (1) 76-85 (2001)

[20] Widén-Wulff, G. Informationskulturen som drivkraft i företagsorganisationen, Åbo Akademi University Press, Åbo (2001)

[21] Widén-Wulff, G. Information as a resource in the insurance business: the impact of structures and proce sses on organisation information behaviour. New Review of Information Behaviour Research 4 79-94 (2003)

[22] Widén-Wulff, G. Business information culture: a qualitative study of the information culture in the Finnish insurance industry. In Introducing Information Management: an Information Research reader, Maceviciute, E. Wilson, T.D., Eds.(31-42). Facet, London (2005)

[23] Bhatt, G.D. Knowledge management in organizations: examining the interaction between technologies, techniques, and people. Journal of Knowledge Management 5 (1) 68-75 (2001)

[24] Park, H.; Ribière, V.Schulte, W.D.J. Critical attributes of organizational culture that promote knowledge management technology implementation success. Journal of Knowledge Management 8 (3) 106-117 (2004)

[25] Widén-Wulff, G. Ginman, M. Explaining knowledge sharing in organizations through the dimensions of social capital. Journal of Information Science 30 (5) 448-458 (2004)

[26] Davenport, E. Snyder, H.W. Managing social capital. Annual Review of Information Science and Technology 39 517-550 (2005)

[27] Widén-Wulff, G. Tötterman, A.-K. A social capital perspective on collaboration and Web 2.0. In Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends, Dumova, T. Fiordo, R., Eds.(101-109). Information Science Reference, Hershey, PA (2009)

[28] Sonnenwald, D.H. Pierce, L.G. Information behavior in dynamic group work contexts: interwoven situational awareness, dense social networks and contested collaboration in command and control. Information Processing & Management 36 461-479 (2000)

[29] Solomon, P. Discovering information in context. Annual Review of Information Science and Technology 36 229-264 (2002)

[30] Byström, K. Hansen, P. Work tasks as units for analysis in information. In Emerging Frameworks and Methods. Proceedings of the Fourth International Conference on Concepts of Library and Information Science (CoLIS4), Bruce, H., et al., Eds.(239-251). Libraries Unlimited, Greenwood Village (2002)

[31] Byrne, R. Employees: capital or commodity? Career Development International 6 (6) 324-330 (2001)

[32] Newell, S. Swan, J. Trust and inter-organizational networking. Human Relations 53 (10) 1287-328 (2000)

[33] Mudge, A. Knowledge management: do we know that we know? Communication World 4 (1) (1999)

[34] Plessis, M.d. Boon, J.A. Knowledge management in eBusiness and customer relationship management: South African case study findings. International journal of information management 24 73-86 (2004)

[35] Mei, Y.M.; Lee, S.T.Al-Hawamdeh, S. Formulating a communication strategy for effective knowledge sharing. Journal of Information Science 30 (1) 12-22 (2004)

[36] Coleman, J.S. Social capital in the creation of human capital. American Journal of Sociology 94 (supplement) 95-120 (1988)

[37] Orlikowski, W.J. Knowing in practice: enacting a collective capability in distributed organizing. Organization Science 13 (3) 249-273 (2002)

[38] Cohen, D. Prusak, L. In good company: how social capital makes organizations work, Harvard Business School Press, Boston (2001)

[39] Brady, M. Blogging: personal participation in public knowledge-building on the web. 2005, University of Essex, Chimera Institute for Social and Technical Research, Colchester (2005)

Address of congratulating author:

Gunilla Widén
Information Studies, Åbo Akademi,
Fänriksgatan 3 B, 20500 Åbo, Finland
Email: gunilla.widen[at]abo.fi

The Six Episodes of Professor Peter Ingwersen's Academic Achievements

Mei-Mei Wu

National Taiwan Normal University, Taipei, Taiwan, ROC

Upon receiving the invitation letter from Professor Ingwersen's associates, Birger, Jesper & Fredrik, I cannot believe Professor Peter is retiring from his full professorship as I believe that teaching and research is his most favourite task. In a short moment, the memory of discussions of the information retrieval interaction dialogue, user-librarian negotiation, and the concept of "episodes" of professional talks pop up. To salute to an always talking, hard working professor, I hereby present six episodes to honour his academic achievement.

From 1980 to 2010, a total of 30 years, with every five years as a slot, there are six blocks. In the sequence of the six blocks, his academic publications counts as 5, 7, 11, 26, 27, and lately 13, making the six episodes the a shape of flying line. Professor Ingwersen won numerous awards and honours as listed in Table 1 including 1992, 1993, 1994, 2002, 2003, 2005, 2007, 2009, 2010 falling in episodes 3, 5 and 6. In particularly, the 2003 ASIST Award for Research in Information Science and 2007 ASIST Outstanding Information Science Teacher Award described his as both gifted as excellent researcher as well as good teacher. In Episode 4, he started to launch the new line of research, Webometrics. I wonder if he keeps working, what will be the next topic for this always innovative mind?

Sources:

http://www.iva.dk/omiva/medarbejdere/default.asp?cid=684&tid=4#forskningsomr
http://www.norslis.net/contactpersons/peteringwersen.html
http://www-clips.imag.fr/mrim/essir03/presentateurs_essir.html

Address of congratulating author:

Professor Mei-Mei Wu
Graduate Institute of Library & Information Studies
National Taiwan Normal University, Taipei, Taiwan, ROC
Email: meiwu[at]ntnu.edu.tw

	1980-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010
Academic activities	1982-84 ESA Research Fellow at European Space Agency (ESA-IRS), Frascati, Italy 1984- associate professor at the Royal School of Librarianship	associate professor at the Royal School of Librarianship, worked with IRM and design of specialized information services and systems for industry	1991 received Ph.D. from Copenhagen Business School, Department of Informatics	1996 member of the editorial boards of several central LIS journals 1997 appointed Visiting Professor (Docent) at Åbo Akademi University, Finland 1997-98 initiated the field of Webometrics together with the late Thomas C. almind 1998-2000 co-chaired a DANIDA sponsored PhD-network in South Africa 1996-2000 leading partner of the Centre for Informetric Studies	1999-2002 Research Professor at the Dep. of Information Studies, Royal School of Library and Information Science	2006 Full Professor in Information Retrieval at the Dep. of Information Studies, Royal School of Library and Information Science
Honours			1992 Anne V. Marinelli Lecture Series, Texas Woman's University, Denton, USA 1993 Jason Farradane Award (UK) 1994 ASIS, New Jersey Chapter Distinguish Lectureship Award		2002 Lazarow Memorial Lecture, The Information school, University of Washington, Seattle, USA, sponsored by Thomson-Scientific 2003 ASIST Award for Research in Information Science 2005 Thomson Award; Derek de Solla Price Medal	2007 ASIST Outstanding Information Science Teacher Award 2009 Lazarow Memorial Lecture, University of Tennessee, Knoxville; ASIST Los Angeles Chapter Contributions to Information Science & Technology Award (CISTA), USA 2010 D.Phil. Honoris Causa, University of Tampere, Finland
Piublications	5	7	11	26	27	13
Research fields	• user-librarian negotiations • information search procedures • cognitive point of view • man-machine interaction • online search facilities • psychological aspects of information	• subject access • information systems design • information retrieval • quantitative analysis • indexing and retrieval techniques • information needs	• information and information science • information retrieval interaction • Ranganathan in the perspective of advanced information retrieval • information needs and semantic entities • human approach	• cognitive perspectives for information retrieval • interactive information retrieval systems(information retrieval interaction) • informetric analysis • webometrics • methodological issues • information science integration • international visibility and impact • organising digital information	• webometrics • cognitive information retrieval • information retrieval interaction • users in context • international visibility and impact • the boomerang effect • cognitive perspectives of document representation • polyrepresentation • information seeking and retrieval • web filtering	• integration of information seeking and retrieval • polyrepresentation • polyrepresentative exploratory search systems • Using citations for ranking in digital libraries • information retrieval • information behavior • Informatics • Scientometric and Webometric methods

Table 1. Professional profile for Professor Peter Ingwersen

Information Sciences in Croatia: A View from the Perspective of Bibliometric Analysis of two Leading Journals

Tatjana Aparac & Franjo Pehar

University of Zadar, Zadar, Croatia

Introduction

From the very beginings of publication *Vjesnik bibliotekara Hrvatske (Croatian Librarian's Herald)* and *Informatologija (Informatology)* took the role of the main channels for the transfer and dissemination of information and knowledge in the field of information sciences[1] in Croatia. The primary driver for the journal originators, editors and authors continues to be the transfer and dissemination of knowledge and information within professional community whose academic and scientific status was not officially recognized until the 1960s and the late 1970s. Due to mentioned reasons along with their pioneering role, the contribution of two journals is even more significant. Therefore, we will follow the process of profiling in the work of two journals based on the course of their publishing of scientific and professional articles. By analyzing the themes of the articles we will also try to unravel the main concerns and interests of local professionals, as well as examine the length of their "common denominator" in the area's subdisciplines. The overall goal is to gain a (better) insight into the degree of development in the field and to examine points of convergence between the theory and practice of Information Sciences.

Field of Information sciences is in the process of constant development, growth and strengthening of its own theoretical foundations. Despite the occasional conflicts, and questioning of the field's maturity, professional expertise and grounding of information professionals is indisputable today. Literature that deals with information institutions and services, such as libraries and archives, museums, documentation facilities and related establishments, predominantly addressed the issues of the profession, organization and institution management, while it only occasionally seeks theoretical explanations of the field's practical activities.

[1] It is commonly accepted in Croatia to use the plural to determine the field that consists of several subdisciplines, such as Library science, Archival Science, Museology, Informatology, Lexicology etc.

Number of published papers in which Information sciences, or some of its disciplines, are examined as scientific branches is much smaller, so the efforts to prove field's scientific grounding is somewhat obscured.

The origin and development of Information Sciences in Croatia

Starting from the so-called Zagreb School of the late 1960's, Croatia's Information sciences included a variety of disciplines that dealt documentation, communication, classification systems, library science, bibliology, archivistics, museology, lexicology and many other disciplines.[2]

In the study of the occurrence and development of Information sciences in Croatia, two key factors stand out:

Dual origin of the field of Information sciences that has evolved from traditional practices of information institutions, particularly archives, libraries and museums as well as influences from other subject areas.

Information science as a new scientific discipline and a professional field formed during the second half of the 20th century in a specific socio-political environment and modeled after international trends in information theory and practice.

The terminology used in the preparation of this paper is based on tradition of so called Zagreb School and established boundaries of scientific areas, fields and branches. More specifically, in the 1980's the University of Zagreb formally established the field of Information Sciences which included a variety of disciplines engaged in "systematic study of the emission processes, collecting, selecting, evaluating, processing, organizing, storing, structuring, retrieval, transmission, distribution, interpretation, use and protection of information, as well as social communication in all its forms."[3]

It is necessary to emphasize here that a relatively small number of theoretical papers on the origins and nature of Croatia's Information Sciences has been published up to date, while the systematic historical and empirical research of the field is almost non-existent.[4] Therefore, this paper will offer an overview of development trends in information sciences based on analysis of papers published in two journals: *Informatologija* and *Vjesnik bibliotekara Hrvatske*.

[2] Maroević, I. Uvod u muzeologiju (Introduction to the Museology). Zagreb: Filozofski fakultet Sveučilišta, Zavod za informacijske studije Odsjeka za informacijske znanosti, 1993. Str. 93. Zagreb: Zagreb University, Department of Information Sciences, 1993. P. 93

[3] Pravilnik o znanstvenim i umjetničkim područjima, poljima i granama. (Scientific and Arts areas, fields and disciplines Act), URL: http://www.nn.hr/clanci/sluzbeno/2005/1500.htm. (3. 8. 2009.)

[4] Valuable contributions to the development of the theory of Information Sciences in Croatia could be find in works by T. Aparac, A. Horvat, I. Maroevic, M. Mikačić, M. Tuđman, T. Šola.

Institutionalization of Information sciences

The process of social and cognitive institutionalization of Information Sciences in Croatia can be traced at three levels: 1) professional associations, 2) central institutions, and 3) academic institutions which have implemented programs for the education of information professionals. Scientific research and professional activities, as well as publishing and the development of educational programs are encouraged at all three levels. The problem, however, lies in fragmentation and the fact that their efforts were uncoordianted.

Particularly important role in the process of social institutionalization of some subdisciplines of Information sciences have been played by an 'umbrella' institutions: National and University Library, the Croatian State Archives, Museum and Documentation Center as well as the Referral Centre. The core of professionals gathered under the roof of national institutions initiated the founding of professional associations that established connections with international organizations, developed a rich publishing activity that spanned from distributing translations of renowned foreign authors to issuing their own monographs and serials, organizing numerous national and international symposiums, conferences etc.[5]

Bozo Tezak and his **e-t-ak-s-a** concept had a particularly important role in the development of Information Sciences in ex-Yugoslavia and Croatia. B. In explaining his approach Tezak uses the term Informatology to mark the area which includes the theory and practice of emission, transmission, accumulation, selection and absorption of information. Influence of Tezak and his network were crucial for the decision to establish the Referral Center at the University of Zagreb in 1967. The Referral Center was conceived as an institution whose goal was to ensure wide cooperation within the University of Zagreb (the only university in Croatia at the time) and to draw together the universities and scholarly organizations in creating an effective information infrastructure, as well as to encourage participation in international scholarly and professional activities.[6] According to Tezak's ideas, role of the Referral Center was not limited to integrating internal and external university cooperation, but also to integrate Croatia's fragmented and uncoordinated IS community. Important advocate of these ideas was suppose to be a journal called *Informatologia Yugoslavica*, launched in 1969. In the course of its development from "infrastructural institution for science and technology to

[5] First professional association, the Croatian Library Association, was founded in 1940 and the Croatian Museum Society was established in Zagreb only six years later. Archivists have a decade later, in 1954, established the Croatian Archival Association, while the documentalists and communicologists establish their professional associations only during 1990's.

[6] Cf. Težak, B. Referral Center of the University of Zagreb. // Informatologia Yugoslavica. 1, 1-4(1969), 68.

independent research organization in the field of information science", Referral Center is transformed into scientific research unit in 1988 – it becomes the *Institute of Information Science*.

Launch of postgraduate studies in documentation and special librarianship in 1961 at the Faculty of Natural Sciences in Zagreb marks the beginning of formal training and integration of Information Sciences into the existing academic structure. Besides this, starting from the 1976/1977 a new program was offered to educate information professionals pursuing undergraduate studies in librarianship at the University of Zagreb, while the study of socio-humanistic informatics is offered one year later. In 1985 both studies become a part of the Department of Information Sciences (Division of Librarianship, Archivistics, Museology and General Informatology).[7]

Since the 1989 University of Zagreb's Department of Information Sciences establishes a separate research unit called *Zavod za informacijske studije* (*Institute of Information Studies)* which aims to organize a continuous and systematic scientific research in the field of Information Sciences and to ensure regular publishing of the works of teachers and employees of the Department. Starting from the academic year 1994/1995 the Department runs postgraduate studies in Information Sciences (sections of Archival Studies, Informatology, Library Science, and Museology).

This created preconditions for a systematic and organized education of information professionals in other university centers in Croatia, especially at the University of Osijek and University of Zadar, which since the year of 1998 and 2003 carried out systematic training for both regular and part-time students. They offered both undergraduate and graduate programs in the field of Information sciences, as well as doctoral programs which started in 2009.

Croatian journals in information science

The appearance and development of Croatia's professional Information Sciences journals are a direct result of the efforts of professional associations, central national information institutions and academic departments to promote scientific research and professional work, as well as to publish activities in the field. One might even argue that these institutions and associations represent the pillar of the national system of scientific and technical communication in the information sciences.

Most Croatian journals in Information Sciences (Table 1) are directed towards promoting content that contributes to the development of theory and / or practices of individual branches gathered within the common scientific field.

[7] CF. Horvat, A. Povijest Odsjeka (The History oft he LIS Department) 2005. URL:http://www.ffzg.hr/infoz/web2/index.php?option=com_content&task=view&id=3&Itemid=4&PHPSESSID=4bd02ccfed24aac238e0af8f68470d41 (01. 06. 2009.)

Journals	Pusblisher	First Published (Year)
Arhivski vjesnik	Croatian State Archives	1945-
Vjesnik bibliotekara Hrvatske	Croatian Library Association	1950-
Muzeologija	Museum Documentation Center	1953-
Informatologia Yugoslavica (continud working under the name of *Informatologia*)	Referral Center, Center for the Study of Librarianship, Documentation and Information Science and International Permanent Exhibition of Publications (1969-1988) Institute of Information Science, a joint post-graduate study of Information Science (1989-1991)	1969-1991
Informatologia	Institute of Information Science (1992-1994) Croatian Information-Documentation Referral Agency (1994-1995) Croatian Communication Association (1995 -).	1991-
Bilten informatica museologica (stopped its publications, it continues working under the name of *Informatica museologica*)	Museum Documentation Center	1970-1973
Informatica museologica	Museum Documentation Center	1973-
Journal of Information and Organizational Sciences	Faculty of Organization and Informatics, University of Zagreb	2002-
Libellarium	Department of Library and Information Science of the University of Zadar	2008-

Table 1. Croatia's scientific and professional journals in the field of Information Sciences

To date, unfortunately, a detailed analysis concerning the process of generation, transmission and use of information by the Croatian information journals has not been conducted. Recent research carried out for the purpose of doctoral thesis[8] examined two journals using the bibliometric method – *Vjesnik bibliotekara Hrvatske* and *Informatology*. *Vjesnik* is the representative of both the librarianship and related information services, while *Informatologija* was launched in order to cover general and specific problems of the entire scientific field of Information Sciences at home and abroad.

[8] Pehar, F. Communication role of journals in the field of information science: bibliometric analysis of the Vjesnik bibliotekara Hrvatske and Informatologia. Zagreb: University of Zagreb, 2010.

Vjesnik bibliotekara Hrvatske (Croatian Librarian's Herald)

Vjesnik bibliotekara Hrvatske, which despite occasional delays in publication of individual issues gets continuously published since 1950, granted the majority of Croatia's librarians their first encounter with professional literature. This happened at a time when international literature was particularly difficult to access and when domestic sources on Library and Information Science were next to non-existent. From its earliest beginnings *Vjesnik* has taken on the role of the main distributor of (new) knowledge and information regarding the field of Library and Information Science, as well as from other related disciplines and activities. In addition to scientific and professional articles the journal has brought news and reports from the works of local and international professional associations.

Over the years *Vjesnik gradually* introduced some innovations in the form and content guidelines. During the last decade a slight shift in programming direction and the journal's concept can be noted because technical papers and other professional reports are increasingly accompanied by the results of scientific research.

Informatologia

The scholarly journal *Informatologia Yugoslavica* (journal changes its name to *Informatologia* in 1991) is closely associated with B. Tezak. As the founder and the journal's first editor, Tezak was credited with all of its characteristics and development trends until 1980.

According to Tezak's idea *Informatologia* was supposed to serve as an "experimental and a working instrument" for researchers from the Referral Center and the PhD students, who would use it to pass on the new cognitive achievements of Information Siences. It would raise public awareness of information activities in Croatia and the former Yugoslavia, while the papers of foreign authors would give local professionals an opportunity to acquaint themselves with the latest world trends. One of the main motivators behind the launch of *Informatologia* can be found in Tezak's fascination with the idea of creating the Croatian system of scientific information and its inclusion in the worldwide network of scientific information. This system was rooted on strong development and application of computers and computer technology as a basic precondition of development of scientific and educational work. His vision of scientific information system is based on the abovementioned e-t-ak-s-a complex and holistic understanding of Information Sciences.

The results of the bibliometric analysis

Exploring the characteristics of communication in the field and the role of the two journals in the creation, transfer and use of information, this study applied standard bibliometric indicators by which it explored the formal features of the two journals, the features of scientific productivity and collaboration, as well as content characteristics of the published papers. Citation analysis based on a selected sample of reference items from two journals was used in studying the degree of development of formal features of scientific communication in the field of Information Sciences. Furthermore, the analysis encompassed intellectual and social structures and relationships among authors, institutions, countries and publications. The analysis of the references used in the two journals examines the mutual influences of the two journals who are representatives of different subdisciplines within the same field. It also aims to determine the limits of the field of Information Sciences and the degree of permeation with related disciplines (interdisciplinarity). Based on the results obtained, we tried to determine the profile and characteristics of the used knowledge source (literature), as a collection of the most influential authors.

The study was conducted on a sample of papers published in *Vjesnik* in the period from 1950 to 2005, and *Informatologia* from the year of 1969 to 2005. During the study period *Vjesnik* had a total of 47 volumes, that is 71 issues, while *Informatologia* published 37 volumes and a total of 70 issues. Although *Vjesnik* published its first issue nearly 20 years before *Informatologia*, the overall difference is negligible and it consists of only one additional volume by *Vjesnik*. Reasons for a negligible difference in the number of published volumes and issues were determined by analysis of publication irregularities and delayed issuing of two journals. Despite the fact that during the past decade both journals had problems with irregular publications, this trend is still considerably more prominent in *Vjesnik* than in *Informatologia*.

From a total of 2501 papers published by *Vjesnik* in the period from 1950 until 2005, a sample of 550 scholarly and scientific papers were selected for our investigation. The sample of selected papers from *Informatologia* contained 710 papers selected from a total of 1093.

One quarter of papers from *Vjesnik* and 11.1% of papers from *Informatologia* failed to cite any literature. The largest number of papers included bibliography at the end of the document; however these lists consisted mostly of six to ten (6–10) references.. In other words, more than half of the surveyed papers included references consisting of ten or less. On the average, *Vjesnik* authors cite 12.8 references per paper, and *Informatologia* cites 11.5. More than four fifths (94.7%) of *Vjesnik* and *Informatology* papers are equipped with at least one summary commonly written in the Croatian language; however, more than two thirds of the papers have two summaries one of which is in a foreign language. However, more than two thirds of the papers have two author abstracts one of which is in a foreign language

Half of the *Informatologia* papers are published in Croatian language, while more than one third are in English. Research papers are published bilingually, in English and original language, from the very beginnings of the journal. The aim of the bilingual publication of the papers is undoubtedly to increase their visibility and to raise exposure of journal's finest articles. More than 90% of papers published in *Vjesnik* are in the Croatian language. The share of foreign authors in Vjesnik was 19.9%, while local authors enjoyed 79.4%. During the same period *Informatologia* recorded a significantly higher proportion of foreign authors, that is 41.3%.

Investigation into the characteristics of authors and authorship, publication activities and cooperation in the period from 1950 until 2005 for *Vjesnik* and 1969 until 2005 for *Informatologia* has produced very interesting results. Both journals have recorded growth in the number of authors over time. *Informatologia*, compared with *Vjesnik*, grew slightly faster and its author base almost tripled in the period from the mid-1980's until 2005. *Informatologia*'s editorial policy relieves the stagnation observed during the war years (1990-1995) by implementing new program guidelines, opening itself to new research topics and areas, as well as "recruiting" a number of authors outside the field of Information Sciences. In the same period *Vjesnik* recorded a similar trend of declining numbers of authors, while new authors were attracted primarily by the papers composed at the annual conference of the Croatian Library Association. We find that the above mentioned information is a very useful indicator that speaks of growing interest in topics from the field of information science and it also testifies about the vitality of community that overcomes a crisis with far-reaching consequences.

Bibliometric profile of authors publishing in the two journals was based on the results of analysis concerning publication activity of authors, institutions, countries and cities. Nearly four fifths of authors from both journals published one or two papers at most. The share of authors with one or two published papers is slightly higher in *Vjesnik* than in *Informatologia*. Statistical tests indicate significant differences in the author profile between *Informatologia* and Vjesnik. Two-thirds of authors from *Informatologia* are from Croatia and one third of them are from Zagreb. Authors associated with academic institutions (70%) are the most productive, with the proviso that more than half of the papers (54%) come from authors whose academic institutions are outside the field of Information Sciences. It is simply impossible to miss and avoid pointing out a relatively weak representation of academic authors from the field of Information Sciences; *Vjesnik* includes only 13% of them and *Informatologia* a mere 16%. The absence of major contributions of authors affiliated to these types of institutions, who in most similar bibliometric studies of information science journals are among the most productive group of authors, is certainly disturbing information that requires the special attention of researchers and additional investigation.

Social changes in the environment were reflected in the two journals which, among other things, recorded the steady growth in productivity and numbers of female authors. Specifically, the first publication of the *Vjesnik*'s reference period reported more male authors, while a period of transformation of librarianship into the profession/discipline in which the female authors have a dominant role begins during the 1980's. Two-thirds of *Informatologia* authors are males and the number of female authors has increased significantly in relation to the first decade of publication.

Journal cross-authorship analysis has been conducted in the search for possible intersections within two journals, as well as related communities. The results show that only 28 out of 846 authors have published at least one papers in both journals. No case of cross-authorship has been established for the most productive authors in the *Informatologija* sample*,* while the most productive authors of the *Vjesnik* published their papers in *Informatologija.* The recorded degree of cross-authorship speaks in favor of a polarization within research groups operating in the field of information sciences whose primary focus are channels operating in a narrower scientific subdiscipline.

The sample of both journals is dominated by single author papers. The subsample of *Vjesnik* holds 85% and *Informatologia*'s share is 71%. However, it is necessary to point out that in the first period of *Vjesnik*'s publications (1950-1968) not one single multi-author paper has been recorded. This data along with the increase of multi-author papers from the mid-1980s (in the sample of both journals) speaks in favor of maturation of information sciences and gradual acceptance of the key features and achievements of modern science and research. During this period the established collaborative coefficient for *Vjesnik* is 0.12. This indicator illustrates a predominantly individual and insular character of research coming from the authors of *Vjesnik*. On the other hand, the authors gathered around *Informatologia* were much more collaborative in the creating and publishing of their works. This is vividly exemplified by the measured collaborative coeffcient which amounts to 0.40. The reason for such sigificant differences among the journals can be explained by a fact that greater numbers of authors coming from academic institutions are joining the *Informatologia*'s subsample. In fact, three quarters of multi-author papers are signed by the authors coming from these institutions. Given the data on the institutions and the countries from which the multi-author works come from, it was found that intramural collaboration among authors coming from the same countries and same institutions was the dominant trend. If we consider that until 1968 *Vjesnik* had no collaborative papers, but how they constitute almost one fifth of the papers in the second reference period as well as a third of *Informatologia's* works, it is possible to argue that this data speaks in favor of improved levels of scientific research in the field of information sciences.

Content analysis of citing papers from *Vjesnik* and *Informatologia* was conducted with the goal of establishing prevailing research trends in the Information Sciences and identifying the possible common grounds with other scientific disciplines. An analysis showed that more than two thirds of papers from *Vjesnik* belongs to one of the four most frequent topics:

a. organization, storage and retrieval of information, with special emphasis on cataloging, information surrogates, classification and indexing systems (22.4%),
b. various kinds of library and information activities and services, with special emphasis on special collections, as well as the activities and services supported by technologies (17.5%)
c. information services (16.0%), with special emphasis on papers that deal with specific types of information institutions, and
d. the history and development of information institutions, with special emphasis on studying the history of libraries and other types of information institutions (14.0%).

Results of content analysis sugegest an overstated historicity of librarianship as a practical activity and an almost complete historical amnesia of librarianship as a scientific discipline. In addition, if we add a relatively weak representation of authors from institutions outside the information/documentation sector, it can be argued that we are dealing with monodsciplinary journal that attracts a community of practitioners and researchers strongly focused on research problems in the narrower field of Library and Information Science (LIS) and the broader field of Information Sciences.

Only two subject categories of works from the *Informatologia* realized a share higher than 10%. These are papers dealing with different aspects of communication sciences (22.5%), with special emphasis on political and linguistic studies of media and communications, as well as papers that deal with different aspects of information and communication technology (18.7%), especially using ICT in learning and other information systems. Two mentioned categories of works, along with papers that deal with the problems of the information society (6.8%), scientific communication (6.5%) and theoretical/empirical research in the field of Information (and Communication) Sciences (5.9%), account for almost two thirds of works in the observed sub-sample. In the last observed period, *Informatologia* had opened itself up to papers from other fields of social sciences (7.3%), espcially in the fields of economics and education, but also to the papers coming from disciplines in the field of technical sciences. The opening to other areas is truly testified by the fact that more than half of works are attributed to authors coming from academic institutions outside the area of Information (and Communication) Sciences. Strong interdisciplinary connections and the opening of the journal to authors from other disciplines is a major challenge for the editors and reviewers who have the difficult task of opening themselves to the outside influences, while also preserving the disciplinary boundaries of Information Sciences.

The results of reference analysis show another important aspect of this research. The paper started from the assumption that using various citation indicators in a sample of selected references would allow identification of the basic characteristics of scientific communication, as well as socio-cognitive structure and development of Information Sciences in Croatia. The distribution of reference for the years studied indicates that their number is constantly growing. The highest number of references was recorded during the last four observed periods. Examination of the 2512 references led to finding that the authors of the *Vjesnik* use monographs (27.8%) equally frequently as journals (26.8%). A similar trend was observed on a sample of 3435 references from *Informatologia*. Slightly less than half (42.3%) of their references were given to monographs, while the share of journal references was 26.5%.

The average age of references in *Vjesnik* was 23.3 years, half of the references from the observed sequence were 9 years old or less, whereas Price's index (PI), i.e. the number of reference publications that are not older than five years, was 34.6%. Citing half-life indicates that the authors of the mid-1980s began using recent literature faster than their predecessors. Changes in chronological characteristics of *Vjesnik*'s references, especially its move towards more frequent use of up to date literature in the last period, are indicators of the transformation and clearer approach of *Vjesnik* to characteristics of other social sciences journals. The average age of references from *Informatologia* was 10.7 years. Price's index was 46.3% and citing hal-life was significantly more stable than the values recorded in *Vjesnik*. Based on the above mentioned information it is possible to conclude that the authors of *Informatologia*, unlike their Vjesnik colleagues, used more recent literature and that chronological characteristics show that *Informatologia is* a journal very similar to other social sciences journals.

The largest numbers of cited publications in *Vjesnik* were published in the Croatian language (40.6%), while the share of references in English amount to 36.0%. Other languages take up less than one fourth of the total number of references (23.4%). Analyses of the country origin of the cited publications suggest that the journal, and accordingly, the entire professional community relied on the influence of European (82.9%), specifically the Central European Library and Information Science tradition (55.7%). Most of cited publications were published in Croatia, Great Britain, Germany and the United States. Compared to *Vjesnik*, analysis of languages cited in *Informatologia* showed some significant differences. More than half of the cited publications are published in English, and only one fifth of them are in the Croatian language. The biggest impact of English literature was recorded in the period from 1969 until 1985 when its share stood at 72.5%. If we add that the authors most frequently used literature from the U.S., Croatia, Britain and Germany, we can conclude that the community of practitioners and scholars gathered around *Informatologia* developed under the twofold influence of both Anglo-Saxon and Central European tradition of information science(s), but were also affected by other scientific disciplines from those geographical areas.

Classification analysis of cited references sought to determine the disciplinary affiliation of references from *Vjesnik* and *informatology*, that is, to measure the degree of subject dispersion and interdisciplinarity in Croatia's Information Sciences as reflected by the two journals. However, one should highlight the methodological limitations of the analysis carried out on the subsample of cited references given to journals. In this way, we found that the authors of papers published in *Vjesnik* rarely use literature from other subject areas. More than three quarters of the references are a so-called disciplinary self-citations directed at journals in the field of Information (and Communication) Sciences, and half of them are directed at library science journals. Based on this information we can say that the narrower field of LIS, represented by *Vjesnik,* is an area with low degree of subject dispersion, and that it relies mainly on the literature from its own field. Although *Informatologia* also used the works from the field of Information (and Communication) Sciences, their share amounts to slightly more than one third of the total number of references. The remaining part of cited references was published in journals of other scientific disciplines. The biggest impact, aside from the journals in Information (and Communication) Sciences, were made by social sciences (23.9%), technical sciences (12.4%) and humanities (8.4%). When these findings are supplemented by the information that part of authors are from institutions outside the area of information sciences, we can say with great certainty that the broader area of Information Sciences (for whom *Informatologia* is the only and the most important representative) includes a high degree of subject dispersion and relies heavily on the literature outside its own areas.

The analysis of frequently cited publications and authors in *Vjesnik* and *Informatology* was conducted with a goal of determining the most influential authors, journals and other types of publications in Information Sciences in Croatia. The list of core information sciences journals was established using Bradford's law of scattering. Although the analysis captures a relatively low absolute number of citations, it is clear that *Vjesnik* provided a certain degree of consensus on the most important literature in which the authors based their research. One, out of total of 188 cited journal titles, received almost one-fifth of citations. It is *Vjesnik*, which along with four other journals represents the first zone, and core of journals in information sciences taking one third of citations. Data on the high level of self-citation testifies to *Vjesnik*'s role and place in the national librarianship as well as the wider information community. Among other types of cited publications, a major role is played by the serials aimed at transferring the official information from the Republic of

Croatia, foreign governments and domestic, as well as international (professional) organizations. A high proportion of official/legal documents and regulations, and manuals required to work in the library, further suggests the primary professional focus of *Vjesnik*. Results of the same analysis conducted on *Informatologia* indicate that there was no consensus on the most important literature of Information Sciences. It was found that the distribution of cited journals according to zones does not follow the predictions of Bradford's law. This happens because of a small number of titles in the second zone, and the excessive dispersion of journals in the third, e.g. the peripheral zone. The largest number of citations are pointing to *Informatologia* (8.0%). The results of citation of other types of publications were compromised by the "insular" nature of citations, that is, four cited publications have received large number of citations from only one study. Therefore, we cannot view it as a consensus, but the decision of the individual author. Such findings call for increased caution in analyzing and interpreting the results of citation analysis. The low threshold of citedness and a relatively modest results in citation analysis make the implementation of consistent and comprehensive data interpretation very difficult. However, presented results are a good starting point for further research and monitoring of the citations in the field of Information Sciences. Without a systematic monitoring and analysis of used literature it is impossible to assess the potential effects of scientific work and create a strategic plan in the development of Information Sciences as a scientific discipline.

The vast majority of the authors received only one citation and even highly cited authors accomplished relatively low absolute values, which further impedes the complete interpretation of the data. The data, however, was still used to generate a list of most frequently cited authors based on the analyzed samples from two reference journals. By comparing the list of the most productive and most cited authors, some interesting overlaps were discovered. From the perspective of two of the analyzed journals, authors whose names are on both of these lists have realized the greatest impact on the development of Information Sciences in Croatia.

Conclusion

In this paper we have tried to give an overview of the most important results of the bibliometric analysis of two Croatian journals in the field of information science. We pointed to the possibility of applying bibliometric methods to explore socio-cognitive structure, investigate the basic knowledge and research topics, as well as to ponder changes within the system of scientific communication in the field of Information Sciences. Furthermore, we managed to produce a bibliometric profile on the process of creation, transfer and use of data in the field of

Information Sciences by following the activities of two very important journals whose work spanned over half a century. This study, unlike most similar studies, is based on the actual collection of selected papers/references and is completely independent of available citation indexes. We believe that this work can be a basis for future studies aimed to create a clear picture of the nature of communication and scientific contributions of Information Sciences in Croatia.

Addresses of congratulating authors:

Tatjana Aparac
Department of Library and Information Science
University of Zadar, 23000 Zadar, Croatia
Email: taparac[at]unizd.hr

Franjo Peha
Department of Library and Information Science
University of Zadar, 23000 Zadar, Croatia
Email: fpehar[at]unizd.hr

Bibliography of Professor Peter Ingwersen

1970s

Ingwersen, P. (1973). *Lånernes brug af kartoteket*. København: Danmarks Biblioteksskole.

Ingwersen, P., Skov, A. & Pejtersen, A.M. (1978). *Informationskundskab: Emners beslægtethed, brugerkommunikation, informations-søgning*. København: Danmarks Biblioteksskole.

Ingwersen, P. & Kaae, S. (1979). User-librarian negotiations and information search procedures in public libraries: Analysis of verbal protocols. Presented at: *The 3rd International Research Forum in Information Science (IRFIS 3), Oslo, August 1-3, 1979.*

Ingwersen, P., Kajberg, L., & Kaae, S. (1979). Letter from Copenhagen. *Journal of Information Science. Principles & Practice*, 1(1), 63-66.

Timmermann, P., Bermann, T., Lau, B., Risager, K., Wille, N.E., Ingwersen, P. & Johansen, T. (1979). A Study of the User-Librarian Negotiation Process. In: T. Berman et al. (Eds.), *Library User Instruction and the Teaching of Research Methodology*. Report on a seminar held at Roskilde University Centre, Denmark, November 5-7, 1976. Roskilde: RUC, 112-129.

1980s

Ingwersen, P., Johansen, T. & Timmermann, P. (1980). User-librarian negotiations and search procedures: A progress report. In: O. Harbo et al. (Eds.), *Theory and application of information research: Proceedings of the Second International Research Forum on Information Science, Copenhagen, August 3-6, 1977*.

Ingwersen, P., Kaae, S., Timmermann, P & Johansen, T. (1980). *User-librarian negotiations and information search procedures in public libraries: Analysis of verbal protocols: Final research report (ERDS ED-211051)*. Copenhagen: Royal School of Librarianship.

Ingwersen, P. (1982). Bibliotekaren kan roligt indrømme sin uvidenhed. *Bogens Verden*, (1).

Ingwersen, P. (1982). *ESA/IRS – Det europæiske informationscenter: Rapport til Danmarks biblioteksskole*. København.

Ingwersen, P (1982). Search procedures in the library analysed from a cognitive point of view. *Journal of Documentation*, 38(3), 165-191.

Ingwersen, P. (1983). Information in Italy. *Journal of Information Science*, 6(2/3), 91-94.

Ingwersen, P. (1983). Online man-machine interaction facilities: A cognitive view. Presented at: *IRFIS 5: Fifth International Research Forum in Information Science: Representation and Exchange of Knowledge as a Basis of Information Processes*. Heidelberg, September 5-7, 1983.

Ingwersen, P. (1984). A cognitive view of three selected online search facilities. *Online Review*, 8(5), 465-492.

Ingwersen, P. (1984). Gli aspetti psicoligici della ricerca dell'informazione. *Bibliotecario*, (1), 33-46.

Ingwersen, P. (1984). Information technology: Which applications? *Social Science Information Studies*, 4(2-3), 185-196.

Ingwersen, P. (1984). Psychological aspects of information retrieval. *Social Science Information Studies*, 4(2-3), 83-95.

Glistrup, E., Skot-Hansen, D., Bråe Olesen, T., Pors, N.O., Skov, A., Sparrevohn, K., Thorhauge, J., Hinge-Christensen, E., Ingwersen, P. & Pedersen, W. (1985). Bibliotekaruddannelsen på vej ud af dødvandet? *Bibliotek 70*, (3, 4), 101-106, 132-136.

Ingwersen, P. (1986). Cognitive analysis and the role of the intermediary in information retrieval. In: R. Davies (Ed.). *Intelligent information systems: Progress and prospects*. New York: Ellis Horwood/Halsted Press, 206-237.

Ingwersen, P. (1986). Interaction in information systems: A review of research from document retrieval to knowledge-based system by Nicholas J. Belkin & Alina Vickery: Book review. *Journal of Documentation*, 42(3), 197-200.

Ingwersen, P., Kajberg, L. & Mark Pejtersen, A. (1984): *Information technology and information use: Towards a unified view of information and information technology*. London: Taylor Graham, 1986.

Ingwersen, P. & Wormell, I. Improved subject access, browsing and scanning mechanisms in modern IR. In: H. Krasner (Ed.), *The proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*. Pisa, September 1986. Pisa: University, 68-75.

Ingwersen, P. (1987). Brugerbetjening – Informationsbehov – Søgeinterview – System-design. *Biblioteksarbejde*, (21/22), 12-70.

Belkin, N.J., Borgman, C.L., Brooks, H.M., Bylander, T. Croft, W.B., Daniels, P., Deerwester, S., Fox, E.A., Ingwersen, P. Rada, R., Sparck Jones, K., Thompson, R. & Walker, D. (1987). Distributed expert-based information systems: An interdisciplinary approach. *Information Processing and Management*, 23(5), 395-409.

Ingwesen, P. (1988). *Folketingets emneordssystem*. København: Danmarks Biblioteksskole.

Ingwesen, P. & Wormell, I. (1988): Means to improved subject access and representation in modern information retrieval. *Libri*, 38(2), 94-119.

Ingwersen, P. (1989). *Analyserapport: Spørgeskemaundersøgelse*. København: Danmarks Biblioteksskole.

Ingwersen, P. (1989). Forskning i information retrieval. *MIC-nytt*, (5), 7-10.

Ingwersen, P. & Wormell, I. (1989). Modern indexing and retrieval techniques matching different types of information needs. In: S. Koskiala & R. Launo (Eds.), *Information, Knowledge, Evolution: Proceedings of the 44th FID Congress*. North Holland, Helsinki, August ,1988, pp.79-90.

Ingwesen, P. & Wormell, I. (1989). Modern indexing and retrieval techniques matching different types of information needs. *International Forum on Information and Documentation*, 14(3), 17-22

Ingwersen, P. & Rist Bork, J. (1989). Quantitative analysis of Danish chemical research production: A study of chemical abstracts. In: *Information og innovation: Beretning fra 7. Nordiske Konference for Information og Dokumentation 28.-30. august 1989 på Århus Universitet, Danmark*, 107-120.

1990s

Ingwersen, P. (1990). *Klynger og klyngeprincippet*. København: Danmarks Biblioteksskole.

Ingwersen, P. (1990). *Rapport over test af tidsforbrug ved emneindeksering i Folketingets database for udvalgsakter*. København: Danmarks Biblioteksskole.

Ingwersen, P. (1990). *Regler for indeksering af bilag og udvalgsdokumenter*. København: Danmarks Biblioteksskole.

Ingwersen, P. (1990). *Sagsområder i Folketingets udvalg*. København: Danmarks Biblioteksskole.

Ingwersen, P. & Wormell, I. (1990). Databases as an analytical tool in research management: A case study. In: B. Cronin (Ed.), *The knowledge industries: Levers of economic and social development in the 1990's: Proceedings of an international conference held at the Inter-University Centre for Postgraduate Studies, Dubrovnik, Yugoslavia, May 29- June 3, 1989*. London: Aslib, 205-216.

Ingwersen, P. & Wormell, I. (1990). *Informationsformidling i Teori og Praksis*. København: Munksgaard, 1990.

Ingwersen, P. (1991). *Intermediary Functions in Information Retrieval Interaction*. Copenhagen : Faculty of Business Administration, Institute of Informatics and Management Accounting, Copenhagen Business School (PhD dissertation).

Ingwersen, P. (1992). Conceptions of information science. In: P. Vakkari & B. Cronin (Eds.), *Conceptions of Library and Information Science*. London : Taylor Graham, 299-312.

Ingwersen, P. (1992). Fra informationskundskab til informationsvidenskab: Kandidatuddannelsens baggrund og status. *Over Broen*, (11), 16-23.

Ingwersen, P. (1992). Information and information science in context. *Libri*, 42(2), 99-135.

Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham, 1992. (Japanese translation, Tokyo: Toppan, 1995; Korean translation, Seoul: Bibliographic Information Processing Society, 1998.)

Ingwersen, P. (1992). Information retrieval research and development. *Information Management Report*, (December), 15-17.

Ingwersen, P. (1992). Library & information science in perspective. *Information Management Report*, (March), 12-17.

Ingwersen, P. & Wormell, I. (1992). Ranganathan in the perspective of advanced information retrieval. *Libri*, 42(3), 184-201.

Belkin, N.J., Ingwersen, P. Pejtersen, A.M. (Eds.) (1992). *SIGIR '92: Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM Press.

Ingwersen, P. (1993). Communication and negotiation edited by Linda L. Putnam & Michael E. Roloff: Book review. *Library Quarterly*, 63, 380-385.

Ingwersen, P. (1993). The cognitive viewpoint in IR. *Journal of Documentation*, 93(1104), 60-64.

Ingwersen, P (1994). Information science as a cognitive science. In: H. Best et al. (Eds.), *Informations- und Wissensverarbeitung in den Sozialwissenschaften*. Opladen: Westdeutscher, 23-56.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: Elements of a cognitive theory for information retrieval interaction. In: *Proceedings of the 17th ACM-SIGIR Conference, Dublin, July, 1994*. London : Springer Verlag, 101-110.

Ingwersen, P. (1994). Systemudvikling i et in-house miljø: Folketingets emneordssystem som case-studie. *Biblioteksarbejde*, (41), 5-24.

Ingwersen, P. (1994). The cognitive perspective in information retrieval. In: *International Federation for Information and Documentation: 47th FID conference and congress: Finding new values and uses of information. Sonic City Omiya, Saitama, Japan, October 5-8*. Sonic City Omiya, Saitama: FID, 7-14.

Ingwersen, P. (1994). The human approach to information science and management: The framework and prospects underlying the new Danish MSc programme. *Journal of Information Science*, 20(3), 197-208.

Ingwersen, P. (1995). Information and information science. In: A. Kent (Ed.), *Encyclopedia of Library and Information Science*, Vol. 56, supplement 19. New York : Marcel Dekker, 137-174.

Ingwersen, P. (1995). Information and information science in context. In: J. Olaisen et al. (Eds.), *Information Science: From the Development of the Discipline to Social Interaction*. Oslo: Scandinavian University Press, 69-111.

Ingwersen, P. & Willett, P. (1995). An introduction to algorithmic and cognitive approaches for information retrieval. *Libri*, 45(3/4), 160-177.

Fox, E. A., Ingwersen, P. & Fidel, R. (Eds.) (1995). *SIGIR '95: Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press.

Hjortgaard Christensen, F. & Ingwersen, P. (1995). Fundamental methodological issues of data set creation online for the analyses of research publications. In: M.E.D. Koenig & A. Bookstein (Eds.) *Fifth International Conference of the International Society for Scientometrics and Informetrics: Proceedings*. Medford: Learned Information, 103-112.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3-50.

Ingwersen, P. & Borlund, P. (1996). Information transfer viewed as interactive cognitive processes. In: P. Ingwersen & N.O. Pors (Eds.), *CoLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in perspective, October 13-16: Proceedings*. Copenhagen : The Royal School of Librarianship, 219-232.

Ingwersen, P. & Pors, N.O. (Eds.) (1996). *Information Science: Integration in Perspective: Proceedings of the 2nd. International Conference on Conceptions of Library and Information Science (CoLIS 2), Oct. 13-16, 1996*. Copenhagen: The Royal School of Librarianship.

Almind, T.C. & Ingwersen, P. (1996). *Informetric analysis on the world wide web: A methodological approach to "internetometrics"*. København: Danmarks Biblioteksskole.

Hjortgaard Christensen, F. & Ingwersen, P. (1996). Online citation analysis: a methodological approach. *Scientometrics*, 37(1), 39-62.

Ingwersen, P. (1997). European research letter: Europe and information science. *Journal of the American Society for Information Science*, 48, 1139-1141.

Ingwersen, P. (1997). *The Central International Visibility of Danish and Scandinavian Research 1988-1996: A General Overview of Science & Technology, the Humanities and Social Sciences by Online Publication Analysis*. (CIS Report 5.3) Copenhagen: The Royal School of Librarianship.

Ingwersen, P. (1997). Tomas Crone Almind – in memoriam. *Biblioteksskole nyt*, (4), 15.

Ingwersen, P. & Hjortgaard Christensen, F. (1997). Data set isolation for bibliometric online analysis of research publications: Fundamental methodological issues. *Journal of the American Society for Information Science*, 48(3), 205-217.

Ingwersen, P., Hjortgaard Christensen, F. & Wormell, I. (1997). Online determination of the Journal Impact Factor and its international properties. In: B. Peritz & L. Egghe (Eds.), *Proceedings of the 6th International Conference of the International Society for Scientometrics and Informetrics, Jerusalem, June 1997*. Jerusalem: Hebrew University, 45-56.

Ingwersen, P., Hjortgaard Christensen, F. & Wormell, I. (1997). Online determination of the journal impact factor and its international properties. *Scientometrics*, 40(3), 529-540.

Almind, T.C. & Ingwersen, P. (1997). Informetric analysis on the World Wide Web: A methodological approach to "webometrics". *Journal of Documentation*, 53(4), 404-426.

Borlund, P. & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3), 225-250.

Ingwersen, P. (1998). Applied logic. *Journal of Documentation*, 54, 354-355.

Ingwersen, P. (1998). Research brief: The calculation of web impact factors. *Journal of Documentation*, 54, 236-243.

Ingwersen, P. (1998). The Calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Ingwersen, P. (1998). *The international visibility of Danish and Scandinavian research 1988-1996: A general overview of science and technology and the social sciences by online publication analysis.* (CIS Report 5) Copenhagen: Royal School of Library and Information Science.

Borlund, P. & Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In: B.W. Croft et al. (Eds.), *Proceedings of the 21 st ACM SIGIR Conference on research and development of information retrieval. Melbourne, Australia*. New York: ACM Press, 324-331.

Ingwersen, P. (1999). Cognitive Information Retrieval. *Annual Review of Information Science & Technology*, 34, 3-52.

Ingwersen, P. (1999). *Det Strategiske Miljøforskningsprogram: Midtvejsevaluering af publiceringsaktiviteten for ni forskningscentre: 1993-95*. Copenhagen: Centre for Informetric Studies/Royal School of Librarianship.

Ingwersen, P. (1999). Online Indicators of Danish Biomedical Publication Behaviour 1988-96: International Visibility, Impact, and Co-operation in a Scandinavian and World Context. *Research Evaluation*, 8 (1), 39-45.

Ingwersen, P. (1999). The role of libraries and librarians in organising digital information. *Libri*, 49, 11-15.

Ingwersen, P. & Wormell, I. (1999). Publication behaviour and international impact: Scandinavian clinical and social medicine 1988-96. In: *Proceedings of the 7th International Conference on Scientometrics and Informetrics*, July 1999, Colima, Mexico, 222-233.

Ingwersen, P. & Wormell, I. (1999). Publication behaviour and international impact: Scandinavian Clinical and Social Medicine 1988-96. *Scientometrics*, 46(3), 487-499.

Borlund, P. & Ingwersen, P. (1999). The application of work tasks in connection with the evaluation of interactive information retrieval systems: Empirical results. In: *Mira '99 : Final Mira Conference on Information Retrieval Evaluation, 14th-16th April 1999*. Glasgow: Electronic workshops in computing, 1-17.

Jørgensen, H. L., Prætorious, L. & Ingwersen, P. (1999). Udvikling af medicinske artikler 1989-1998: En undersøgelse af Danmark i forhold til de øvrige medlemmer af Den Europæiske Union. *Ugeskrift for Læger*, 161(46), 6339-6343.

Lykke Nielsen, M. & Ingwersen, P. (1999). The word association methodology: A gateway to work-task based retrieval. In: *Mira '99 : Final Mira Conference on Information Retrieval Evaluation, 14th-16th April 1999*. Glasgow: Electronic workshops in computing, 17-27.

2000s

Ingwersen, P. (2000). Editorial: Introduction to the special issue from the Royal School of Library and Information Science, Denmark. *Journal of Documentation*, 56(1). 1-4 (Special issue from the Royal School of Library and Information Science, Denmark).

Ingwersen, P. (2000). The cognitive information structures in information retrieval. In: I. Wormell (Ed.), *ProLISSA: Progress in library and information science in South Africa: Proceedings of the first biannual DISSAnet conference: Southern African LIS research in progress 2000*.

Ingwersen, P. (2000). The international visibility and citation impact of Scandinavian research articles in selected social science fields: The decay of a myth. *Scientometrics*, 49, 39-61.

Ingwersen, P., Larsen, B. & Wormell, I. (2000). Applying Diachronic Citation Analysis to Ongoing Research Program Evaluations. In: B. Cronin & H.B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Medford: Information Today & American Society for Information Science, 373-387.

Belkin, N.J., Ingwersen, P. & Leong, M-K. (Eds.) (2000). *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM SIGIR July 24-28, 2000, Athens, Greece*. New York: ACM Press.

Cosijn, E. & Ingwersen, P. (2000). Dimensions of Relevance. *Information Processing & Management*, 36, 533-550.

Jacobs, D. & Ingwersen, P. (2000). A Bibliometric Study of the Publication Patterns in the Sciences of South African Scholars 1981-96. *Scientometrics*, 47, 75-93.

Järvelin, K., Ingwersen, P. & Niemi, T. (2000). A user-oriented interface for generalized informetric analysis based on applying advanced data modelling techniques. *Journal of Documentation*, 56, 250-278.

Ingwersen, P. (2001). The web impact factor. In: M. Davis & C.S. Wilson (Eds.), *8th International Conference on Scientometrics and Informetrics: ISSI 2001, Sydney, July 16-20, 2001: Proceedings: Vol. 1*. Sydney: The Bibliometric and Informetric Research Group, 12-13.

Ingwersen, P. (2001). Users in context. In: M. Agosti et al. (Eds.), *Lectures on Information Retrieval.* (Lecture Notes in Computer Science: 1980) Bonn: Springer, 157-178.

Ingwersen, P. (2001). Visibility and impact of research in psychiatry for North European countries in EU, US and world contexts. In: M. Davis & C.S. Wilson (Eds.), *8th International Conference on Scientometrics and Informetrics*: ISSI 2001, Sydney, July 16-20, 2001: Proceedings: Vol. 1, Sydney: The Bibliometric and Informetric Research Group, 265-274.

Ingwersen, P. & Larsen, B. (2001). *Guidelines for løbende scientometriske vurderinger af dansk sundhedsvidnskabelig forskning: Rapport til Statens Sundhedsvidenskabelige Forskningsråd*. København: Danmarks Biblioteksskole.

Ingwersen, P., Larsen, B. & Noyons, E. (2001). Mapping national research profiles in social science disciplines. *Journal of Documentation*, 57(6), 715-740.

Ingwersen, P., Larsen, B., Rousseau, R. & Russell, J. (2001). The publication-citation matrix and its derived quantities. *Chinese Science Bulletin*, 46(6), 524-528; 700-704. (Also available in Chinese)

Larsen, B. & Ingwersen, P. (2001). Synchronous and diachronous citation analysis for information retrieval. Generating a boomerang effect from the network of scientific papers. In: M. Davis & C.S. Wilson (Eds.), *8th International Conference on Scientometrics and Informetrics: ISSI 2001, Sydney, July 16-20, 2001: Proceedings: Vol. 1*. Sydney: The Bibliometric and Informetric Research Group, 355-368.

Björneborn , L. & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1), 65-82.

Ingwersen, P. (2002). *Analyse af dækningsgraden 1998 i science citation index for anvendte tidsskrifter samt analyse af førsteforfatterandelen inden for dansk sundhedsvidenskab: Rapport afgivet til statens sundhedsvidenskabelige forskningsråd.* København: Statens Sundhedsvidenskabelige Forskningsråd.

Ingwersen, P. (2002). Cognitive perspectives of document representation. In: H. Bruce et al. (Eds.), *Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4).* Greenwood Village: Libraries Unlimited, 285-300.

Ingwersen, P. (2002). Visibility and impact of research in Psychiatry for North European countries in EU, US and world contexts. *Scientometrics*, 54(1), 131-144.

Ingwersen, P. & Jacobs, D. (2002). South African research in selected scientific areas: Status 1981-2000. In: T. Bothma & A. Kaniki (Eds.), Andrew: *ProLISSA : Progress in Library and Information Science in Southern Africa: Proceedings of the second biennial DISSAnet Conference in Southern Africa, 2002*. Glenstantia: Infuse, 77-92.

Bruce, H., Fidel, R., Ingwersen, P. & Vakkari, P. (Eds.) (2002). Emerging Frameworks and Methods: *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4)*. Greenwood Village: Libraries Unlimited

Larsen, B. & Ingwersen, P. (2002). The boomerang effect: Retrieving scientific documents via the network of references and citations. In: *SIGIR 2002: Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*. New York : ACM Press, 397-398.

Thorlund Jepsen, E., Seiden, P., Björneborn, L., Lund, H. & Ingwersen, P. (2002). WebTAPIR: scientific information retrieval on the world wide web. In: H. Bruce et al. (Eds.), *Emerging frameworks and methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4)*. Greenwood Village: Libraries Unlimited, 309-312.

Ingwersen, P. (2003). *Dækningsgraden i Scince Citation Index af forskningen internationalt indenfor de sundhedsvidenskabelige universitetsfakulteter og HS, samt i de Centrale Medicinske Databaser 1998: Revideret rapport udarbejdet til Statens Sundhedsvidenskabelige Forskningsråd*. København: Danmark Biblioteksskole.

Ingwersen, P. (2003). Information seeking research. In: J. Feather & P. Sturges (Eds.), *International Encyclopedia of Information and Library Science*. (2 ed.) London: Routledge, 301-303.

Ingwersen, P. (2003): Users in IR context. In: *European Summerschool in Information Retrieval: ESSIR '03: Student text booklet*. Grenoble: Laboratoire IMAG, Université de Grenoble, 103-121.

Ingwersen, P., Bojsen-Møller, T. & Bruun, R. (2003). *Det strategiske miljøforskningsprogram: Evaluering af ni forskningscentre 1993-98: Slutevaluering*. København: Danmarks Biblioteksskole.

Björneborn, L. & Ingwersen, P. (2003). Cheshm-andâzhâyi bar web-sanji. Ettela Resani [Journal of Information Science], 19(1/2), 64-86.

Larsen, B., Ingwersen, P., Skram, U. & Viby-Mogensen, J. (2003). Scandinavian Anaesthesia research 1981-2000 *DASINFO*, 11(4), 31.

Larsen, B., Lund, H. Andresen, J.K. & Ingwersen, P. (2003): Using value-added document representations in INEX. In: *INEX 2003 workshop proceedings: December 15-17, 2003, Schloss Dagstuhl, International Conference and Research Centre for Computer Science*. London: INEX, 62-72.

Skram, U., Larsen, B., Ingwersen, P. & Viby-Mogensen, J. (2003). Anaesthesia research in Northern Europe 1981-2000: Visibility and impact in EU context. *European Journal of Anaesthesiology*. 20(suppl. 30), 195.

Ingwersen, P. (2004). Cand.scient.bibl. Thomas Crone Almind hædret internationalt. *Biblioteksskolenyt*, (6), 14-15.

Ingwersen, P. (2004). Det strategiske miljøforskningsprogram: Forskningsaktivitet og international gennemslagskraft. In: *Miljøforskning*. Århus : Det Strategiske Miljøforskningsprogram, 8-25.

Ingwersen, P. (2004). Highlights of a career in information science. *Bulletin of the American Society for Information Science and Technology*, 31(6), 6-8.

Ingwersen, P. & Belkin, N.J. (2004). Information retrieval in context: IRIX: workshop at SIGIR 2004. *SIGIR forum*, 38(2).

Ingwersen, P. & Björneborn, L. (2004). Methodological issues of webometric studies. In: H.F. Moed et al. (Eds.), *Handbook of Quantitative Science and Technology Research : The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht : Kluwer, 339-370.

Ingwersen, P. & Jacobs, D. (2004). South African research in selected scientific areas: Status 1981-2000. *Scientometrics*, 59(3), 405-423.

Ingwersen, P. & Järvelin, K. (2004). Context in information retrieval. In: T. Bothma & A. Kaniki (Eds.), *ProLISSA: Progress in Library and Information Science in Southern Africa: Proceedings of the third biennial DISSAnet Conference in Southern Africa, 2004*. Pretoria: Infuse, 301-310.

Ingwersen, P. & Järvelin, K. (2004). Extending information seeking and retrieval research toward context In: P. Ingwersen et al. (Eds.), *Information retrieval in context: SIGIR 2004: IRiX Workshop*. Sheffield: Sheffield University, 6-9.

Ingwersen, P. & Järvelin, K. (2004). Information retrieval in contexts. In: P. Ingwersen et al. (Eds.), *Information Retrieval in Context: IRiX: ACM-SIGIR Workshop 2004 Proceedings*. Sheffield: Sheffield University, 6-9.

Ingwersen, P & Lynge, E. (2004). Dækningsgraden i Science Citation Index af dansk sundhedsvidenskabelig forskning 1998. *Ugeskrift for Læger*, 166(40), 3493-3497.

Ingwersen, P., Thorlund Jepsen, E. & Borlund, P. (2004). Scientific web publications: Characteristics and retrieval constraints. In: M. Hummelshøj (Ed.) *Knowledge and change: Proceedings of the 12th Nordic Conference for Information and Documentation*, September 1-3, 2004, Hotel Hvide Hus, Aalborg, Denmark. Aalborg : Royal School of Library and Information Science, 35-43.

Ingwersen, P., van Rijsbergen, K., Belkin, N.J. & Larsen, B. (Eds.) (2004). *Information retrieval in context: IRIX : ACM-SIGIR workshop 2004 proceedings*. Sheffield: Sheffield University.

Björneborn, L. & Ingwersen, P. (2004). Towards a basic framework for Webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.

Järvelin, K. & Ingwersen, P. (2004). Extending information-seeking and retrieval research towards context. T. Bothma & A. Kaniki (Eds.), In: *ProLISSA : Progress in Library and Information Science in Southern Africa: Proceedings of the third biennial DISSAnet Conference in Southern Africa, 2004*. Pretoria : Infuse, 311-324.

Järvelin, K. & Ingwersen, P. (2004). Information seeking research needs extension towards task and technology. *Information Research*, 10(1).

Jepsen, E.T., Seiden, P., Ingwersen, P., Björneborn, L. & Borlund, P. (2004). Characteristics of scientific Web publications: preliminary gathering and analysis. *Journal of American Society for Information Science & Technology*, 55(14): 1239-1249.

Jørgensen, H. L., Larsen, B., Ingwersen, P. & Rehfeld, J. (2004). Forskningsaktiviteten for kandidater med ph.d- eller dr.med.-grad fra de sundhedsvidenskabelige fakulteter 1995-1997. *Ugeskrift for Læger*, 166(6), 479-489.

Skov, M., Pedersen, H., Larsen, B. & Ingwersen, P. (2004). Testing the principle of polyrepresentation. In: *Information retrieval in context: SIGIR 2004 IRiX Workshop*. Sheffield : Sheffield University, 47-49.

Skram, U., Larsen, B., Ingwersen, P. & Viby-Mogensen, J. (2004). Scandinavian research in anaesthesiology 1981-2000: Visibility and impact in EU and world context. *Acta Anaesthesiologica Scandinavica*, 48, 1006-1013.

Ingwersen, P. (2005). Integrative framework for information seeking and interactive information retrieval. In: Fisher, K.E. et al. (Eds.). *Theories of Information Behavior*. Medford: Information Today/ASIS&T, 215-220.

Ingwersen, P. (2005). Selected variables for IR interaction in context: Introduction to IRiX SIGIR 2005 Workshop. In: *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*. Copenhagen : Department of Information Studies, Royal School of Library and Information Science, 6-9.

Ingwersen, P. (2005). Webometric research. *Information Research Watch International.* (February), 2-3.

Ingwersen, P. & Järvelin, K. (2005). Information retrieval in context – IRiX: SIGIR workshop report. *ACM SIGIR forum*, 39(2).

Ingwersen, P. & Järvelin, K. (2005). The sense of information: Understanding the cognitive conditional information concept in relation to information acquisition. In: F. Crestani & I. Ruthven (Eds.), *Context: Nature, impact and role: 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005 Glasgow, UK, June 4-8, 2005: Proceedings*. Berlin : Springer, 7-19.

Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Berlin: Springer. (Chinese translation, Scientific and Technical Documents Publishing House, 2007.)

Ingwersen, P., Järvelin, K., Belkin, N. & Larsen, B. (Eds.) (2005). *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX): Salvador, Brazil, 19 August, 2005.*

Ingwersen, P., & Larsen, B. (2005). Evaluation of strategic research programs: The case of Danish environmental research 1993-2002. In: P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005: the 10th international conference of the International Society for Scientometrics and Informetrics*. Stockholm: Karolinska University Press.

Ingwersen, P., & Larsen, B. (Eds.) (2005). *Proceedings of ISSI 2005: The 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24-28, 2005*. Stockholm: Karolinska University Press.

Ingwersen, P., White, H.D. & Schlemmer, B. (2005). Introducing the Derek de Solla price awardees of 2005: Peter Ingwersen, Howard D. White. *ISSI Newsletter*, 1(2), 9-12.

Larsen, B. & Ingwersen, P. (2005). Cognitive overlaps along the polyrepresentative continuum. In: A. Spink & C. Cole (Eds.), *New Directions in Cognitive Information Retrieval.* Oxford : Oxford University, 43-60.

Lund, H., Larsen, B., Voel Jensen, R. E., Golub, K., & Ingwersen, P. (2005). Capturing context for web filtering in the humanities. In: *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX).* Copenhagen : Department of Information Studies, Royal School of Library and Information Science.

Ingwersen, P. (2006). Context in information interaction: Revisited 2006. In: *ProLISSA 2006: Proceedings of the Fourth Biennial DISSAnet Conference.* Pretoria: University of Pretoria, The I-School, 1-10.

Ingwersen, P. (2006). Webometrics: Ten years of expansion. In: *Proceedings of the International Workshop on Webometrics, Informetrics and Scientometrics and 7th COLLLNET Meeting, 10-12 May 2006*. Nancy: INIST/CNRS, 4-8.

Larsen, B. & Ingwersen, P. (2006). Using Citations for Ranking in Digital Libraries. In: G. Marchionini (Ed.) *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, June 11-15, 2006*. New York: ACM, 370.

Larsen, B., Ingwersen, P. & Kekäläinen, J. (2006). The Polyrepresentation continuum in IR. In: I. Ruthven et al. (Eds.), *Information interaction in context: International Symposium on Information Interaction in Context, IIiX 2006*. New York/Copenhagen: ACM Press/Royal School of Library and Information Science, 148-162.

Lund, B., Schneider, J.W. & Ingwersen, P. (2006). Impact of relevance intensity in test topics on IR performance in polyrepresentative exploratory search systems. In: *Proceedings of the ACM-SIGIR workshop on evaluating exploratory search systems (ESS 2006).* Seattle: University of Washington, 42-46.

Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N. J., Tombros, A., & Vakkari, P. (Eds.) (2006). *Information Interaction in Context: International Symposium on Information Interaction in Context, IIiX 2006 : Copenhagen, Denmark, 18-20 October 2006: Proceedings*. New York/Copenhagen: ACM Press/Royal School of Library and Information Science.

Skov, M., Larsen, B. & Ingwersen, P. (2006). Inter and intra-document contexts applied in polyrepresentation. In: I. Ruthven et al. (Eds.), *Information Interaction in Context: International Symposium on Information Interaction in Context, IIiX 2006*. New York/Copenhagen: ACM Press/Royal School of Library and Information Science, 163-170.

Ingwersen, P. (2007). Degree of granularity in publication profiles for mapping national research similarity. In: *Taking CiteSpace to Science: Workshop of 11th International Conference of the International Society for Scientometrics and Informetrics, ISSI, 2007*. Madrid : Centre for Scientific Information and Documentation (CINDOC) of the Spanish Research Council, 15-16.

Ingwersen, P., Hjortgaard Christensen, F. & Wormell, I. (2007). Online determination of the journal impact factor and its international properties. In: T. Braun (Ed.), *The Impact Factor of Scientific and Scholarly Journals: Its Use and Misuse: A selection of papers reprinted from the journal Scientometrics*. Budapest: Akadémiai Kiadó, Scientific Journals Business Centre, s. 373-384.

Ingwersen, P. & Järvelin, K. (2007). On the Holistic Cognitive Theory for Information Retrieval: Drifting outside the cave of the Laboratory Framework. In: S. Dominich & F. Kiss (Eds.), *Studies in Theory of Information Retrieval*. Budapest: Foundation for Information Society, 135-147.

Ingwersen, P. & Larsen, B. (2007). Evaluation of strategic research programs: The case of Danish environmental research 1993-2002. *Research Evaluation*, 16(1): 47-58.

Ingwersen, P., Larsen, B., Rehfeld, J. & Jørgensen, H. (2007). Scientometriske metoder til måling af forksningsaktiviteten og gennemslagskraft. *Klinisk Biokemi i Norden*, 19(3), 10-23.

Ingwersen, P., Ruthven, I. & Belkin, N.J. (2007). First international symposium on information interaction in context. *Sigir forum*, 41(1), 117-119.

Ingwersen, P., Schneider, J.W., Scharff, M. & Larsen, B. (2007). A national research profile-based immediacy index and citation ratio indicator for research evaluation. In: D. Torres-Salinas & H. Moed (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*, ISSI, 2007. Madrid : Centre for Scientific Information and Documentation (CINDOC) of the Spanish Research Council, 864-865.

Hyldegård, J. & Ingwersen, P. (2007). Task complexity and information behaviour in group based problem solving. In: *Information Research*, 12(4). Special supplement (Proceedings of the Sixth International Conference on Conceptions in Library and Information Science, Borås, Sweden, 13-16 August, 2007).

Larsen, B., Björneborn, L. & Ingwersen, P. (2007). The 12th Nordic Workshop on Bibliometrics and Research Policy in Copenhagen. *ISSI Newsletter*, 3(3), 41-43.

Schneider, J. W., Larsen, B. & Ingwersen, P (2007). Comparative study between first and all-author co-citation analysis based on citation indexes generated from XML data. In: D. Torres-Salinas & H. Moed (Eds.), *Proceedings of 11th International Conference of the International Society for Scientometrics and Informetrics, ISSI, 2007*. Madrid : Centre for Scientific Information and Documentation (CINDOC) of the Spanish Research Council, 696-707.

Ingwersen, P. (2008). A context-driven integrated framework for research on interactive IR. *Document, Information & Knowledge*, 126(6), 44-50.

Ingwersen, P. & Järvelin, K. (2008). Informationssøgning set i det Integrerede Kognitive Forskningsperspektiv. *Dansk Biblioteksforskning*, 4(2): 17-32.

Ingwersen, P. & Järvelin, K. (2008). On the integrated cognitive theory for information retrieval: Drifting outside the cave of the laboratory framework. *Revista General de Información y Documentación*. 18, 381-402.

Jørgensen, H.L., Larsen, B., Ingwersen, P. & Rehfeld, J.F. (2008) Forskningsaktiviteten for speciallæger i klinisk biokemi. *Ugeskrift for Læger*, 170(36), 2798-2802.

Papaeconomou, C., Zijlema, A.F. & Ingwersen, P. (2008). Searchers' relevance judgments and criteria in evaluating Web pages in a learning style perspective. In: P. Borlund et al. (Eds.), *Information Interaction in Context: Proceedings of the second IIiX Symposium on Information Interaction in Context*. (ACM International Conference Proceedings series) New York: ACM Press, 123-130.

Skov, M. & Ingwersen, P. (2008). Exploring information seeking behaviour in a digital museum context. In: P. Borlund et al. (Eds.), *Information Interaction in Context: Proceedings of the second IIiX Symposium on Information Interaction in Context*. (ACM International Conference Proceedings Series) New York: ACM Press, 110-115.

Skov, M., Larsen, B. & Ingwersen, P. (2008). Inter and intra-document contexts applied to polyrepresentation. *Information Processing & Management*, 44, 1673-1683.

Ingwersen, P. (2008). Information Science. In: W. Donsbach (Ed.), *The International Encyclopedia of Communication*. London : John Wiley & Sons, 2261-2263.

Ingwersen, P. (2009). Brazil research in selected scientific areas: Trends 1981-2005. In: B. Larsen & J. Leta (Eds.), *Proceedings of ISSI 2009: The 12th International Conference of the International Society for Scientometrics and Informetrics: Rio de Janeiro, Brazil, July 14-17, 2009*. Rio de Janeiro: Bireme/Paho/WHO and Federal University of Rio de Janeiro, 692-696.

Ingwersen, P. & Chavan, V.S. (2009). Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, 10 (Suppl 14).

Ingwersen, P. Lund, B. & Larsen, B. (2009). Data fusion according to the principle of polyrepresentation. *Journal of American Society for Information Science & Technology*, 60(4), 646-654.

Iribarren-Maestro, I., Ingwersen, P. & Larsen, B. (2009). Research assessments by synchronic and diachronic citation impact: A case study of Carlos III University of Madrid. In: B. Larsen & J. Leta (Eds.), *Proceedings of ISSI 2009: The 12th International Conference of the International Society for Scientometrics and Informetrics: Rio de Janeiro, Brazil, July 14-17, 2009*. Rio de Janeiro: Bireme/ Paho/WHO and Federal University of Rio de Janeiro, 487-491.

Schneider, J.W., Larsen, B. & Ingwersen, P. (2009). A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. *Scientometrics*, 80(1), 103-130.

Ingwersen, P. (2010). Scientometric and Webometric methods. *Document, Information & Knowledge*, (1), 4-11.

Järvelin, K. & Ingwersen, P. (2010). User-oriented and cognitive models of information retrieval In: M. Bates (Ed.), *Encyclopedia of Library and Information Science*. 3 ed. London: Taylor & Francis, 5521-5534.

Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In: C. Gurrin et al. (Eds.), *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010*. (Lecture Notes in Computer Science; 5993) Berlin: Springer, 627-630.

Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P. & van Rijsbergen, K. (2010): Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework. In: *Proceedings of IIiX 2010 – 3rd Information Interaction in Context Symposium, New Brunswick NJ, USA, August 18-22, 2010*. (forthcoming).

Lioma, I., Larsen, B., Schütze, H. & Ingwersen P. (2010): A Subjective Logic Formalisation of the Principle of Polyrepresentation for Information Needs. In: *Proceedings of IIiX 2010 – 3rd Information Interaction in Context Symposium, New Brunswick NJ, USA, August 18-22, 2010*. (forthcoming).

Ingwersen, P., Bogers, T., Larsen, B., Lykke, M. & Lund, H. (2010): Assessors' Search Result Satisfaction Associated with Relevance in a Scientific Domain. In: *Proceedings of IIiX 2010 – 3rd Information Interaction in Context Symposium, New Brunswick NJ, USA, August 18-22, 2010*. (forthcoming).

Lykke, M., Ingwersen, P., Bogers, T., Larsen, B., Lund, H. (2010): Physicists' information tasks: structure, length and retrieval performance. In: *Proceedings of IIiX 2010 – 3rd Information Interaction in Context Symposium, New Brunswick NJ, USA, August 18-22, 2010*. (forthcoming).

Elleby, A. & Ingwersen (2010): Publication Point Indicators: A Comparative Case Study of two Publication Point Systems and Citation Impact in an Interdisciplinary Context. *Journal of Informetrics* (forthcoming).

Tabula Gratulatoria

Finally, it is our pleasant duty to announce the names of those colleagues and friends of Peter who were not able to contribute to this volume, but who have expressed their wish to congratulate Peter Ingwersen on the occasion of his retirement and appointment as Professor Emeritus. Herewith, we kindly acknowledge best wishes expressed by

Dag W. Aksnes
Ragnar Audunson
Ricardo Baeza-Yates
Lennart Björneborn
Harry Bruce
Katriina Byström
Erica Cosijn
Rickard Danell
Mari Davis
Mark Dunlop
Efthimis Efthimiadis
Wolfgang Glänzel
Jette Hyldegård
Henrik L. Jørgensen
Joemon Jose
Carol Kuhltau
Haakon Lund
Mona Madsen
Gary Marchionini
Ragnar Nordlie
Nils Pharo
Ari Pirkola
Stephen Robertson
Reijo Savolainen
Balázs Schlemmer
Gunnar Sivertsen
Dorte Skot-Hansen
Annette Skov
Eero Sormunen
Kirsten Strunck
Sanna Talja
Anastasios Tombros
Elaine Toms
Pertti Vakkari
Conception Wilson
Irene Wormell

The Janus Faced Scholar

This Festschrift honours Professor Peter Ingwersen on his retirement from the Royal School of Library and Information Science (RSLIS) and concomitant appointment as the first Professor Emeritus at the Royal School.