Experiences with the 2013-2016 CLEF interactive information retrieval tracks

Petras, Vivien; Koolen, Marijn; Gäde, Maria; Bogers, Toine

Published in:
CEUR Workshop Proceedings

Publication date:
2019

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

Experiences with the 2013-2016 CLEF Interactive Information Retrieval Tracks

Vivien Petras
Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Berlin
vivien.petras@ibi.hu-berlin.de

Marijn Koolen
Humanities Cluster
Royal Netherlands Academy of Arts and Sciences
Amsterdam
marijn.koolen@di.huc.knaw.nl

Maria Gäde
Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Berlin
maria.gaede@ibi.hu-berlin.de

Toine Bogers
Science and Information Studies
Department of Communication & Psychology
Aalborg University Copenhagen
Copenhagen
toine@hum.aau.dk

ABSTRACT

This paper describes our experiences with the interactive IR tracks organized at CLEF from 2013-2016 and aggregates the lessons learned with each consecutive instance of the lab. We end with a summary of practical insights and lessons for future collaborative interactive IR evaluation exercises and for potential re-use scenarios.

KEYWORDS

interactive information retrieval, evaluation, CHiC, SBS, CLEF, book search, information seeking

1 INTRODUCTION

After the INEX (Initiative for Evaluation of XML Retrieval) Interactive Track ended in 2010 [23], there was a gap in interactive information retrieval (IIR) experimentation at the large-scale evaluation initiatives. The interactive track at the Cultural Heritage at CLEF (Conference and Labs of the Evaluation Forum) lab (iCHiC) revived this in 2013 and merged with the INEX Social Book Search track to form the Social Book Search (SBS) lab at CLEF, running an interactive track in 2014-2016.

This paper provides a chronological overview of the development and history of these two IIR initiatives and their outcomes. We focus on the lessons learned for future collaborative IIR evaluation exercises and for potential re-use scenarios. We start by chronicling the timeline of the different interactive labs that were organized in Sections 2-6. We then highlight the most important lessons learned for the configuration of IIR evaluation experiments. We conclude by discussing consequent activities and insights for the re-use of IIR resources.

Workshop on Barriers to Interactive IR Resources Re-use at the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2019), 14 March 2019, Glasgow, UK

2 CULTURAL HERITAGE IN CLEF @ CLEF 2011-2012

2.1 Setup

The EU-funded PROMISE[1] project (Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation) ran from 2010-2013 with the goal of providing a virtual and open laboratory for research and experimentation with complex multimodal and multilingual information systems [7]. In order to evaluate its concepts and prototypes, three use cases were defined to guide real-world requirements analysis and contextual testing: 'Unlocking Cultural Heritage' (information access to cultural heritage material), 'Searching for Innovation' (patent search) and 'Visual Clinical Decision Support' (radiology image retrieval).

For the 'Unlocking Cultural Heritage' (CH) use case, a workshop at the 2011 CLEF conference was organized in order to review existing information access use cases in the CH domain and then develop retrieval scenarios that could be used for evaluating CH information access systems [11]. In addition to qualitative usability tests of user interfaces, transaction log analyses and Cranfield-style text retrieval evaluation, other forms of user studies were also considered as viable evaluation approaches. The study and analysis of different interaction patterns with CH materials was the main interest of the workshop's participants[2].

At the 2012 CLEF conference, a pilot evaluation exercise was organized for the CH domain, progressing work from the workshop format to an evaluation lab [26]. It was based on a real-life collection of CH material: the complete index of the Europeana digital library[3], which encompassed ca. 23 million metadata records in 30 different languages at that time. The information needs were based on 50 queries (harvested from Europeana logfiles), translated into English, French and German. The tasks in this pilot exercise comprised both a conventional system-oriented scenario (i.e., ad-hoc retrieval) as well as more specialized retrieval scenarios for the CH domain– the semantic enrichment and variability tasks[4]. The evaluation

[1] http://www.promise-noe.eu
[2] http://www.promise-noe.eu/chic-2011/
[3] www.europeana.eu
[4] http://www.promise-noe.eu/chic-2012/home

followed the Cranfield paradigm by pooling the retrieval results and assessing their relevance using human assessors.

2.2 Lessons learned

Although the 2011 CHiC workshop had already emphasized that a focus on user interaction patterns was an important evaluation aspect for the CH domain, this first CHiC lab in 2012 had no interactive tasks. Instead, it utilized a document collection based on Europeana and used queries harvested from Europeana logs to construct information needs. The vision was to extend the ad-hoc style retrieval evaluation with interactive and other evaluation scenarios (particularly result presentations and alternative methods for relevance assessments) in the next phases.

The Europeana document collection, albeit a real-world collection, turned out to be very challenging. While an effort was made to normalize the provided metadata by wrapping it in a special XML format and removing certain metadata fields, the content in the metadata had very different descriptive qualities, depending on the original content provider. Both the data sparseness and multilinguality of the content posed serious challenges for the participants. Image data, such as thumbnails of graphical material in Europeana, could not be provided due to copyright reasons.

Some of the provided topics were not suitable for relevance assessment, because information needs could not always be unambiguously inferred from the provided queries. The topics mostly contained short queries of 1-3 words and only half of them had short descriptions added, which did not help much when the topic was vague. For the CH use case, IIR studies focusing on interaction patterns were needed, so an additional interactive task was proposed for the next round.

3 INTERACTIVE CHIC TRACK @ CLEF 2013

3.1 Setup

The Interactive Track[5] at the CHiC 2013 lab at CLEF (iCHiC) aimed at building a bridge for IIR and behavior researchers to work in a TREC-style evaluation environment. The idea was to develop a data collection of IIR evaluation data, which could be re-used and built upon. This task intentionally used a subset of the document collection used in the other CHiC ad-hoc retrieval experimental tasks to allow for later triangulation of results. Based on approximately 1 million metadata records from the English Europeana collection and representing a broad range of CH objects, a simple search interface was envisioned that would allow for browse and search interactions with the metadata records for the IIR experiments [25]. One non-goal oriented task (based on Borlund's simulated work tasks [4, 5]), which simulated "casual" use of the system ("*spend 20 minutes on the system and explore*") was provided to all experiment participants.

The same experimental infrastructure, which hosted the web-based interfaces, documents and logged the interactions [19] was provided to all participating research groups. All groups had to recruit at least 30 participants: at least 10 of them had to be observed in the lab, while at least 20 could use the system remotely. Apart from the logged interactions on the systems, participants also filled out pre- and post-task questionnaires, assessed their experience on the User Engagement Scale [24] and evaluated the usefulness of found objects (relevance assessment) and the interface (usability).

[5] http://www.promise-noe.eu/chic-2013/tasks/interactive-task

3.2 Lessons learned

The iCHiC track ended up collecting data on 208 experiment participants and their interactions from four participating research groups. As a pilot experiment for collaborative data gathering, this first interactive task was successful overall.

The most important lesson learned from iCHiC and the reason why it was merged with the INEX Social Book Search lab (see Section 4) was that the provided metadata records were not "rich" enough in content to provide an interesting case study for casual browsing and search. The sparseness of the document collection had already been a problem for the ad-hoc retrieval tests, and real users did not like them any better. The actual purpose of iCHiC—to study users' interactions with the content—was hampered by the lack of interesting content.

The experimental set-up and questionnaire instruments represented a significant effort for the participants to complete. However, the collected data was deemed necessary for further analysis.

An original plan for the set-up of this task was to provide the metadata collection, simulated work tasks, and the experimental setup (questionnaires, logging protocol) to the participating research groups and have them provide their own infrastructure for data gathering. After discussions, the organizers concluded that having different groups each building infrastructures would add too much variability and also pose a large barrier to entry especially for groups that did not have software or GUI design specialists.

The data gathering at the University of Sheffield's servers had the additional advantage of having a central place where all the data was stored. This also posed a problem in later years, however, when researchers affiliated with the University of Sheffield moved to a different institutions and neither the preservation and maintenance of the infrastructure and data nor its legal ownership were established.

Four teams participated in the track, but not all of them were able to recruit the 30 required participants. The uneven contribution led to some discussion about the fairness of all groups then being able to use the same data in later analyses. Initial discussions on who would get to analyze the data with which research questions in which priority (important for later publications) were never successfully resolved as the organizers moved on to new tasks. Some of the organizers published follow-up analyses of the data [16], while other participating research groups did not.

The participating groups all adhered to research ethics requirements set forth by the University of Sheffield, which hosted the platform and the collected data. Different ethical requirements (e.g., based on national law) were not considered. The experimental participants were asked to consent to their responses being shared not just with the organizers, but with the wider research community, which allows for re-use of the data. However, processes for enabling the data sharing at a later time were not considered.

The proposal for the interactive task had planned for a two-year period, where the data gathering (user interaction logging) and preliminary data analysis would happen in the first year. In year

two, an aggregated data set of all logged interactions was to be released to the research community in order to inform an improved system design for data gathering, which would start again in year two. While the organizers provided an initial analysis of the data [32], a planned follow-up analysis of the data did not take place.

4 FIRST INEX iSBS TRACK @ CLEF 2014

Social Book Search (SBS)[6] started as a system-centered evaluation campaign at INEX in 2011 [21], focusing on retrieval and ranking of book metadata and associated user-generated metadata, such as user reviews, ratings and tags from Amazon and LibraryThing [1]. The main research question behind the track was how to exploit the different types of curated and user-generated metadata for realistic and complex book search requests as expressed on the book discussion forums of LibraryThing. After its third year, the organizers discussed changes to the SBS lab, specifically the nature of book search tasks and how they are evaluated. At the same time, the iCHiC organizers were looking for a different collection than the Europeana cultural heritage objects, because they struggled to come up with a meaningful task that engaged users, as the cultural heritage metadata descriptions got little interest from participating users. Initial discussions between the SBS and iCHiC organizers suggested books and associated social media data might be a more natural domain for participating users. By tying an interactive track to a system-centred track around the same collection and tasks, lessons learned in one track could feed into the other. Thus the interactive SBS (iSBS) track was launched.

Another important initiative was to study the different stages of the search process and how they could be supported by different interfaces [**?**]. We considered models of the information search process [10, 22, 33] in combination with models of how readers select books to read [15, 28–31]. The book selection models distinguish between book internal features (e.g., subject, treatment, characters, ending) and external features (e.g., author, title, cover, genre) [29], but all are based on interaction in physical libraries and book shops, so they had to be adapted to online environments, where the users have no access to the full-text, but to additional data in the form of user-generated content. Thus, selection is based only on external features.

This led to a three-stage model of *browsing*, *searching* and *selection*, each with separate interfaces that carry over user choices when switching between interfaces, based on Goodall [15]. These stages correspond to the three stages in Vakkari's model of *pre-focus*, *focus* and *post-focus* [33]. There was a lengthy discussion on what functionalities to include in each stage and how to label the different interfaces, to ensure that they made sense to users while retaining a close connection to the three search stages and selection stages from the literature. It took many iterations of UI choices to adapt the system to the data that was available and deemed most useful to the searcher based on book search studies [15, 28, 30]. Such extensive tailoring of the search UI to the data collection naturally makes reuse of UI components problematic.

We were interested in the difference between goal-oriented and non-goal oriented tasks, also to compare the non-goal oriented task in the book domain to the same non-goal task in CH as used in iCHiC [16]. In choosing a simulated work task, we considered tasks that could be connected to specific stages in the search process, similar to Pharo and Nordlie [27].

4.1 Setup

The 2014 iSBS Track did not run as a full evaluation campaign, because most of the year was used to prepare and set up the multi-stage search system, tasks and protocol [17]. However, each of these components improved on the iCHIC set-up: a more interesting collection, more focus on the user interfaces and more varied tasks. The track organizers recruited a small number of participants (41) but decided to open up the experiment to other groups only in the second year. The multi-stage system was compared against a baseline system that had mostly the same features but all in a single view. The experiment included a training task, a goal-oriented task and a non-goal oriented task. Pre- and post-experiment questionnaires asked for demographic and cultural information, and the overall experience and engagement with the interface. Post-task questionnaires asked about the usefulness of different interface features. Most of the questions were constructed specifically for this domain and system, but the engagement questions were reused from the iCHiC Track. The underlying experimental system of the iCHiC experiments was also reused, but had to be modified somewhat to fit the iSBS Track.

4.2 Lessons learned

Although the long preparation phase left little time for gathering data, it resulted in a consensus among the large group of organizers about the set of generic research questions that the experimental setup and search systems should be able to address.

The setup did not lead to enough complex interactions to identify stage transitions in the search process and to test the value of multi-stage interfaces. We considered multiple causes: (1) the tasks were relatively simple and did not require complex interactions; (2) the instructions and training task were not sufficient to get users familiar with such an interface; and (3) the interface was not self-explanatory enough for users to interact with meaningfully. The questionnaire data suggested the tasks could be completed with little effort. We subsequently discussed whether we should use more complex yet still realistic book search tasks.

There was a conflict between the goal of studying social book search with realistic tasks and the goal of studying the value of interfaces for different stages in the search process. The models of Kuhlthau [22] and Vakkari [33] are based on researchers and students searching information to write a report or essay and are perhaps less relevant to casual leisure search for books. Or perhaps the users lack a felt need with the simulated tasks, but would display more complex interactions if they really were searching for one or more books to buy.

5 SECOND iSBS TRACK @ CLEF 2015

5.1 Changes from previous edition

The second year of the iSBS track was open to other research groups and had a longer data gathering period with many more participants (192 in total) [14]. Most of the setup was kept the same to allow comparison with the results of the previous year. However, the

[6] http://marijnkoolen.com/Social-Book-Search/

goal-oriented task was redesigned to have five different sub-tasks, to make users interact more and for longer periods of time.

5.2 Lessons learned

We found that the fact that metadata in the book collection was exclusively available in English was a hurdle for several non-native English speaking users. As some participating groups contributed many more users than other groups, with more non-native English speakers, the balance was very different than the year before, which makes comparison of cohorts difficult.

Users also spent a lot of time on the goal-oriented task with sub tasks, causing some of them to abandon the experiment after the first of the two tasks. In their feedback, others indicated that the overall experiment took too long. This could mean that the gathered data is biased towards more persistent participants.

6 THIRD iSBS TRACK @ CLEF 2016

6.1 Changes from previous edition

In the third edition of the iSBS track we made more significant changes to the experimental setup. Some modifications were made to the experiment structure to avoid participants abandoning the experiment. The main change was that users only had one mandatory task, but could continue with other tasks as long as they were willing to continue. We added eight tasks based on book search requests from the LibraryThing discussion forums to provide as realistic tasks as possible [13]. Another big change was that we focused only on the multi-stage interface to have fewer variables in the gathered data. FInally, a third change was that each participating institution had their own instance of the experiment to ensure participant allocation was balanced for each institution, not only for the overall experiment. This was mainly because some institutions had specific cohorts, which they could not analyse across the variables when balancing was only done overall.

6.2 Lessons learned

A comparison of the 2015 and 2016 cohorts showed very few differences in terms of time spent on goal-oriented and non-goal tasks (the 2015 cohort showed no ordering effect between doing goal-oriented first and doing non-goal-oriented first), giving a strong indication that the experiment structure and tasks are producing reliable results. This also suggests that the two cohorts could be combined to reduce the impact of individual differences. One of the hardest struggles in IIR evaluation campaigns is getting a large and diverse enough set of users. Running such campaigns for long periods requires continuity. The same experimental systems need to remain available with at most small changes.

The additional tasks based on requests from the LibraryThing discussion forums resulted in different search behaviour from the simulated goal-oriented and non-goal oriented tasks, but also showed large differences between the LibraryThing tasks themselves, with more subjective, fiction-oriented tasks leading to less interaction than concrete, non-fiction-oriented tasks. This suggests that IIR findings may be very sensitive to the specifics of the simulated work tasks used. It may also signal that in order to study information search for reading for one's own enjoyment, it is important that users have 'skin in the game' and feel a personal connection to leisure-focus work tasks.

A problem encountered since running the 2016 iSBS Track is that organizers move between institutions, which causes problems for maintaining experimental systems, websites and repositories when they loose institutional access to servers where the infrastructure is hosted on. This in turn endangers the continuous availability of research data and experiments. A natural solution to this recurring problem could be an independent or inter-institutional platform and repository for these systems and materials.

7 OUTCOMES: WHAT DID WE LEARN?

7.1 Document Collections

One important lesson learned from the iCHiC and iSBS tracks is the importance of a suitable document collection that is realistic in both size and content variety. The document collection used for iCHiC was based on metadata from Europeana. Even though it represented a broad range of different topics, the individual items in the dataset were often sparse in their information content. In the iSBS tracks, the document collection based on Amazon and LibraryThing data offered richer information that is more suitable for an interesting task for users, but over the course of the different iSBS editions the collection grew increasingly out-of-date. We found this negatively affected search behavior as well as user engagement, especially during the open search task. Users were looking for recent book titles and got frustrated that they could only find books that were at least six years old.

While re-use of IIR resources is important for replicability and reproducibility, oftentimes older document collections are simply not interesting anymore for participants—something system-based evaluation suffers from to a lesser degree. How to obtain realistic, engaging, and up-to-date document collections, while at the same time maintaining comparability across evaluation iterations, remains an open question.

Using a live document collection from a production system would not allow for the same number of interactions to be studied and poses difficulty for logging. It is not a simple alternative. Arguably, what matters is not the stability of the set of documents that are searchable, but the extent to which that set is up-to-date. Book search interactions gathered in 2014 can be compared with those gathered in 2019 if in both cases users could search books published in the last five years, despite there being no overlap between the two collections, as long as the type and amount of information about books remains the same. To improve re-usability, it may be more valuable to investigate and describe relevant aspects of document collections, so that IIR studies with different document collections can be compared based on their overlapping relevance aspects, e.g., recency, structure, type, and amount of information.

Unfortunately, realistic document collections tend to exhibit a larger degree of variety and complexity. This may make them more engaging and interesting to participants, but it also increases the complexity of the analysis of their behavior. One could argue that to achieve a more detailed and thorough analysis, perhaps simpler document collections would be more suitable, thereby setting up a trade-off between complexity at the experimental and the analysis stages.

7.2 Information Needs

In order to have meaningful impact, IIR studies should be representative of the real-life variety in domains, system designs, and user types and needs. One way in which iCHiC and iSBS attempted to do this was by using a varied and realistic set of simulated work tasks [6] and cover stories that include extra context about the background task to support the search behavior of participants. How best to generate such realistic information needs is an open question. One potentially fruitful approach in the 2016 iSBS track involved taking real-world examples of complex information needs from the LibraryThing forums and using them as optional additional work tasks. These tasks were judged as being rich in variety and detail by our participants, so this could be an interesting avenue for future work. However, as the difference between fiction and non-fiction tasks showed, personal interest does play an important role in user engagement, so using real-world requests as simulated work tasks is not a catch-all solution.

Despite the proven usefulness of simulated work tasks, they are still not the same as a user's own information needs. We therefore also included work tasks in iCHiC and iSBS that focused on the participants' own information needs. Non-restrictive tasks, in which users can search whatever and however they want for as long or short as they want, offer more realistic aspects of information behavior, but they make comparison more difficult. Differences between users can be due to them having wildly different 'tasks' in mind. Although we experimented with different types of tasks, we feel that we have only scratched the surface here. True information needs can be multilingual and multicultural, making assessment even more challenging.

In addition, by focusing only on single information needs, we believe that we are ignoring valuable aspects of the entire information seeking process, both individual and collaborative [20]. Information search is only one aspect of information behavior and is commonly combined with exploration, browsing, or interaction with a recommender system. Moreover, information behavior often takes place across and between different devices (desktop vs. smartphone), information systems (e.g. Amazon, LibraryThing, Google but also social media channels like Facebook and Twitter [9]) and modalities (digital vs. paper). On the other hand, a large number of varied information needs and task contexts leads to a wide distribution of experimental data points, which—if not enough users can be persuaded to participate—may result in insufficiently significant analyses.

7.3 Study Participants

Ideally, an IIR evaluation campaign recruits participants that are a realistic representation of the general target population to avoid the introduction of biases [8, p. 241]. However, in most IIR tracks—including our own—researchers have often relied on recruiting students from participating universities or research groups as participants. Due to the short-term preparations and research cycles, this is often the only way to include enough participants in an IIR experiment. However, students are only one of several user groups that need to be taken into account when dealing with complex search tasks. It needs to be assured that users are selected based on the specific system, feature or task to be tested as ignoring these relationships and dependencies is likely to lead to invalid results. Longer preparation time or access to user databases with potential participants could help overcoming such biases in participant recruitment.

One of our findings in iSBS was that the cultural background makes a significant difference. This is something that is rarely reported in studies, but that appears to be an important aspect to include. This also challenges the assumption that by providing the same infrastructure and tasks but using different user group distributions over the years or across national boundaries, measured user interactions can be aggregated across these groups. There were some analyses that clustered users based on certain aspects, but the question remains which users can be viewed in aggregation. Since academic IIR studies often rely on students, perhaps studies can explicitly describe criteria of representativeness of the target user group and add questions to the questionnaire that capture aspects of users that allows mapping them to these aspects of representativeness.

7.4 Search User Interface

The search user interface is perhaps the most important aspect to get right for the IIR system used in the experiments as our experience with iCHiC and iSBS tracks has taught us. The ubiquity and popularity of modern-day search engines means that any search user interface has certain minimum expectations to meet in terms of layout and/or functionality. Not meeting these expectations means risking distracting users and has a deleterious effect on their search behavior. It would be beneficial if the IIR system offered the flexibility of choosing different search interfaces to study the effects of the GUI on information seeking behavior. This was used to great effect in the iSBS tracks to examine how different interfaces can support the different search stages.

This flexibility came at a price, however, as the software components needed for the infrastructure became increasingly complex. Both iCHiC and iSBS used a customized infrastructure developed by one of the organizers, which made this possible [18]. Maintaining customized software for future experiments is a hard problem. Making infrastructure publicly available with appropriate documentation is one way to alleviate this.

Another difficulty is that the design of interfaces can be informed by different theoretical models of information interaction. In setting up the iSBS track and designing the multistage interface, we discussed the appropriateness of numerous information seeking/search models as well as book selection models and strategies, how they are related to each other and how they correspond to or are supported by aspects of the interface. A further complication is that our choices were also steered by the research questions we wanted to address. These issues add another set of variables to take into account when considering comparison and reuse, and should be described in studies.

7.5 Experimental Setup

IIR research usually includes several complex components that can affect the quality and success of each experiment. While the importance of some elements such as task development have been extensively discussed, other aspects remain less considered. Only a

few studies report on or discuss measures used to analyze or interpret results from IIR experiments. So far, IIR measures are highly contextual varying from experiment to experiment. Measures used span from data on interactions, such as session duration or clicks, to qualitative data derived from questionnaires or interviews. Often several data points are complemented or correlated.

A collaborative IIR study requires that participating research groups pool their gathered data and aggregating this data generates substantial overhead. If institutions gather their own data, aggregation may involve harmonizing inconsistencies. In the iCHiC and iSBS tracks, a single system was used to gather all experimental data, but this system had to be developed and adapted with each iteration. A comprehensive documentation and accurate descriptions of the data gathering tools is crucial for the evaluation and re-use of these aspects in future studies.

Different research groups and individuals often want to study slightly different aspects of the problem domain or setup, requiring different questions in the questionnaire, different tasks or users, or different search system components. With every change, new users need to be recruited, and comparisons to previously collected data becomes harder. The long preparatory discussions among the iSBS organizers regarding research questions, theoretical frameworks and research designs suggests that it is possible to some extent to incorporate a broad set of research questions in the overall research design to allow a range of studies with the same setup. But often research questions change or new questions are prompted during and following the experiments, calling for an iterative development of the research design. We are not aware of any guidelines on how to best update designs to allow some backwards comparability. While there is large variability in research questions and research designs, the group would have benefited from re-using other researchers' research design components, as was done with the User Engagement Scale [24] in both iCHiC and iSBS. Apart from documenting the broad aspects of the experimental set-up in the track overview papers, a thorough documentation and subsequent publication of questionnaire items, scales and other measures would not only help other researchers in not having to re-invent standard items (e.g., demographic questions), but also support the standardization of IIR research.

7.6 Data Storage, Infrastructure Maintenance & Intellectual Property Rights

From 2011 until 2016, the various interactive tracks generated a wealth of data, but also went through numerous organizational changes, both in terms of the individuals involved and the institutions that provided infrastructure. iSBS started as part of INEX with some data stored on servers dedicated to INEX activities, other data stored on servers maintained by one of the organizers' institutions and the search indexes on another set of servers of another organizing institution.

Recurring questions are (1) what happens if organizers leave and own crucial pieces of the data or infrastructure, and (2) what happens when organizers move between institutions, thereby losing access to data or infrastructure? For research data management purposes, it is important that organizers of IIR studies make explicit who is responsible for which part of the data and systems, who owns the data or infrastructure, and what happens when organizers move to other institutions or leave the project, or when new organizers join.

While always intended, the organizers of iCHiC and iSBS could find hardly any re-use of the gathered data for IIR studies or triangulation studies with the related ad-hoc retrieval experiments in CHiC or SBS. One reason may have been the insufficient availability of the research data along with a proper rights clearance.

There are generic platforms for storing and sharing scientific data, such as the Open Science Framework[7] and several Dataverse[8] instances. These options solve some of the institutional issues, but they lack the flexibility to run experimental systems or to add domain-specific search and access features to datasets that make a repository like RepAST useful to the IIR community. Publicly available repositories for software and software infrastructures also exist (e.g., GitHub), but present similar problems to the research data repositories.

Next to problems of storage and access of IIR research data, there are issues of copyright, privacy and ethics. The questionnaire informs users, which institutions are involved, but how should organizers deal with new researchers and institutions joining? One option is for organizers to agree on ethical guidelines for data gathering, informed consent and data representation. For further data re-use, it is crucial that users also give their informed consent for additional analyses of their data. To create a trustworthy environment, IIR researcher must provide concrete statements on who and for what future purposes the data will be used. This should be available additionally to the research data as part of an archived and documented research design (see Section 7.5).

7.7 Coordinating Collaborative Research

IIR research is a highly interdisciplinary field bridging areas of information seeking, interactive and system-centered (ranking, evaluation) IR and user interface design. Accordingly, researchers from different disciplines need to collaborate on complex questions and experimental setups. Entering the field of IIR research is still a challenge due to inconsistent or incompatible practices. Even for those that work on IIR problems, no collaboration on systems, tasks, data, participants or research questions can be observed. This might be the case due to time and resource constraints caused by traditional one-year research cycles as well as unawareness of other projects.

In assessing the interest in an interactive track in the SBS Lab during a joint iCHiC and SBS discussion session at CLEF 2013, everyone who stated their interest was involved in the initial discussions in setting up the track, to get an overview of what aspects they wanted to investigate, thereby shaping the track around a broad set of interests. This community input is valuable both in attracting groups to actively participate and in creating a setup with potential for long term community support and interest. A challenge of the desired community input and larger organizer numbers is the required additional overhead for the decision processes. Once again, good documentation and communication is vital as are well-understood guidelines or practices about the consequences of researchers joining or leaving the initiative. Collaborative research

[7] https://osf.io/
[8] https://dataverse.org/

also entails a joint understanding of how research results will be presented (e.g. rules of authorship and priority). This is especially important in large collaborations.

Collaborative research, by its very nature, tries to study aspects which require a large-scale infrastructure, a large number of users or other aspects that need a strong community input. This will necessarily prolong the design and implementation phases of any study, which is a detriment in a fast-paced scholarly context as IIR research, especially within the large evaluation campaigns or research conferences, which run on annual cycles. This type of work would be best supported by a multi-year project or by moving to a slower research output model.

8 OUTCOMES: WHERE TO GO FROM HERE?

Based on previous experiences from the CLEF/INEX Interactive Social Book Search tracks, the two Supporting Complex Search Tasks (SCST) community workshops (2015 and 2017) [2, 12] were organized to discuss IIR challenges and future directions in the area of complex search scenarios since cooperation between the different tracks was rarely seen. The invited researchers from various fields concluded that collaborative IIR campaigns have great potential, but lack standardization and sustainability. Since previous efforts such as the Systematic Review of Assigned Search Tasks (RepAST) [34] have only been partly noticed or used, it remains an open question how to secure the persistence of IIR research designs and results.

The 2018 workshop on Barriers to IIR Resources Re-use (BIIRRR) switched the focus to the analysis and preparation of requirements for effective re-use of IIR resources or experiments [3]. The development of quality standards for the curation and re-use of research designs has been identified as one of the main tasks in this initiative, along with the appropriate documentation and publication of research data and the requisite software. Research designs were named as a priority, because they appear to have the highest potential for standardization and re-use in other IIR studies. This requires a proper analysis of previously used research design elements as well as motivation for or against potential re-use of these elements.

One idea is to develop a platform that would allow researchers from interdisciplinary fields to search for IIR research designs once they have been identified as re-usable and are stored and documented. Building such a repository requires an analysis and implementation of user requirements both for accessing and contributing research designs, the development and agreement on a standardized data infrastructure as well as a maintenance plan coordinated by a stable team of researchers.

Apart from a proper documentation and archiving strategy, this retrospective also pointed towards pre-study aspects, which are instrumental for re-using experimental research data and designs. This includes the establishment of guidelines for cross-national and cross-institutional data collection, informed consent and data distribution. As was declared several times in this paper, the reusability of research designs and other IIR study components strongly depends on the community's willingness to develop and maintain proper documentation, curation and publication guidelines. While this may not be as rewarding as creating new research data by implementing more IIR studies (and we need more of these as well), it is crucial for the community to standardize in order to move forward as a research discipline.

REFERENCES

[1] Thomas Beckers, Norbert Fuhr, Nils Pharo, Ragnar Nordlie, and Khairun Nisa Fachry. 2010. Overview and Results of the INEX 2009 Interactive Track. In *ECDL (Lecture Notes in Computer Science)*, Mounia Lalmas, Joemon M. Jose, Andreas Rauber, Fabrizio Sebastiani, and Ingo Frommholz (Eds.), Vol. 6273. Springer, 409–412.

[2] Nicholas Belkin, Toine Bogers, Jaap Kamps, Diane Kelly, Marijn Koolen, and Emine Yilmaz. 2017. Second Workshop on Supporting Complex Search Tasks. In *Proc CHIIR 2017*. ACM, New York, NY, 433–435.

[3] Toine Bogers, Maria Gäde, Mark Hall, Luanne Freund, Marijn Koolen, Vivien Petras, and Mette Skov. 2018. Report on the Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2018). *SIGIR Forum* 52, 1 (Aug. 2018), 119–128.

[4] Pia Borlund. 2003. The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research* 8, 3 (2003).

[5] Pia Borlund. 2016. Interactive Information Retrieval: An Evaluation Perspective. In *CHIIR '16: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 151–151.

[6] Pia Borlund and Peter Ingwersen. 1997. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation* 53, 3 (1997), 225–250.

[7] Martin Braschler, Khalid Choukri, Nicola Ferro, Allan Hanbury, Jussi Karlgren, Henning Müller, Vivien Petras, Emanuele Pianta, Maarten de Rijke, and Giuseppe Santucci. 2010. A PROMISE for Experimental Evaluation. In *Multilingual and Multimodal Information Access Evaluation*, Maristella Agosti, Nicola Ferro, Carol Peters, Maarten de Rijke, and Alan Smeaton (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 140–144.

[8] Donald O. Case and Lisa M. Given. 2016. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (4th ed.). Emerald Group Publishing, Bingley, UK.

[9] Otis Chandler. 2012. How Consumers Discover Books Online. In *Tools of Change for Publishing*. O'Reilly.

[10] David Ellis. 1989. A behavioural model for information retrieval system design. *Journal of information science* 15, 4-5 (1989), 237–247.

[11] Maria Gäde, Nicola Ferro, and Monica Lestari Paramita. 2011. CHiC 2011 - Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. In *CLEF Notebook Papers/Labs/Workshop*.

[12] Maria Gäde, Mark M. Hall, Hugo Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skove, Elaine Toms, and David Walsh. 2015. Report on the First Workshop on Supporting Complex Search Tasks. *SIGIR Forum* 49, 1 (June 2015), 50–56.

[13] Maria Gäde, Mark Michael Hall, Hugo C. Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skov, Toine Bogers, and David Walsh. 2016. Overview of the SBS 2016 Interactive Track. In *Working Notes of the CLEF 2016 Conference (CEUR Workshop Proceedings)*, Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald (Eds.), Vol. 1609. CEUR-WS.org, 1024–1038.

[14] Maria Gäde, Mark Michael Hall, Hugo C. Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skov, Elaine Toms, and David Walsh. 2015. Overview of the SBS 2015 Interactive Track. In *Working Notes of the CLEF 2015 Conference (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, Gareth J. F. Jones, and Eric SanJuan (Eds.), Vol. 1391. CEUR-WS.org.

[15] Deborah Goodall. 1989. *Browsing in public libraries*. Library and Information Statistics Unit LISU.

[16] Mark Hall, Robert Villa, Sophie Rutter, Daniel Bell, Paul Clough, and Elaine Toms. 2013. Sheffield Submission to the CHiC Ineractive Task: Exploring Digital Cultural Heritage. CLEF Working Notes.

[17] Mark Michael Hall, Hugo C. Huurdeman, Marijn Koolen, Mette Skov, and David Walsh. 2014. Overview of the INEX 2014 Interactive Social Book Search Track. In *Working Notes of the CLEF 2014 Conference (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij (Eds.), Vol. 1180. CEUR-WS.org, 480–493.

[18] Mark M Hall, Spyros Katsaris, and Elaine Toms. 2013. A Pluggable Interactive IR Evaluation Work-bench. In *European Workshop on Human-Computer Interaction and Information Retrieval*. 35–38. http://ceur-ws.org/Vol-1033/paper4.pdf

[19] Mark Michael Hall and Elaine Toms. 2013. Building a Common Framework for IIR Evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 17–28.

[20] Preben Hansen, Chirag Shah, and Claus-Peter Klas. 2015. *Collaborative Information Seeking*. Springer.

[21] Marijn Koolen, Gabriella Kazai, Jaap Kamps, Antoine Doucet, and Monica Landoni. 2012. Overview of the INEX 2011 Books and Social Search Track. In *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011) (LNCS)*, Shlomo Geva, Jaap Kamps,

and Ralf Schenkel (Eds.), Vol. 7424. Springer.

[22] Carol C. Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science* 42, 5 (1991), 361–371.

[23] Ragnar Nordlie and Nils Pharo. 2012. Seven Years of INEX Interactive Retrieval Experiments – Lessons and Challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, Tiziana Catarci, Pamela Forner, Djoerd Hiemstra, Anselmo Peñas, and Giuseppe Santucci (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 13–23.

[24] Heather L. O'Brien and Elaine G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69. DOI: http://dx.doi.org/10.1002/asi.21229

[25] Vivien Petras, Toine Bogers, Elaine Toms, Mark Hall, Jacques Savoy, Piotr Malak, Adam Pawłowski, Nicola Ferro, and Ivano Masiero. 2013. Cultural Heritage in CLEF (CHiC) 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein (Eds.). Springer Berlin Heidelberg, 192–211.

[26] Vivien Petras, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio, and Juliane Stiller. 2012. Cultural Heritage in CLEF (CHiC) Overview 2012. In *CLEF 2012 Labs and Workshops*.

[27] Nils Pharo and Ragnar Nordlie. 2012. Examining the effect of task stage and topic knowledge on searcher interaction with a digital bookstore. In *Proceedings of the 4th Information Interaction in Context Symposium*. ACM, 4–11.

[28] Kara Reuter. 2007. Assessing aesthetic relevance: Children's book selection in a digital library. *Journal of the American Society for Information Science and Technology* 58, 12 (2007), 1745–1763.

[29] Catherine Sheldrick Ross. 1999. Finding without seeking: the information encounter in the context of reading for pleasure. *Information Processing & Management* 35, 6 (1999), 783 – 799. DOI: http://dx.doi.org/10.1016/S0306-4573(99)00026-6

[30] Catherine Sheldrick Ross. 2000. Making choices: What readers say about choosing books to read for pleasure. *The Acquisitions Librarian* 13, 25 (2000), 5–21.

[31] Katariina Saarinen and Pertti Vakkari. 2013. A sign of a good book: readersâĂŹ methods of accessing fiction in the public library. *Journal of Documentation* 69, 5 (2013), 736–754.

[32] Elaine Toms and Mark Hall. 2013. The CHIC interactive task (CHICi) at Clef2013. CLEF Working Notes.

[33] Pertti Vakkari. 2001. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of documentation* 57, 1 (2001), 44–60.

[34] Barbara M. Wildemuth and Luanne Freund. 2012. Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval (HCIR '12)*. ACM, New York, NY, USA, Article 4, 10 pages. DOI: http://dx.doi.org/10.1145/2391224.2391228