Data citation and the citation graph

Buneman, Peter; Dosso, Dennis; Lissandrini, Matteo; Silvello, Gianmaria

Published in:
Quantitative Science Studies

DOI (link to publication from Publisher):
10.1162/qss_a_00166

Publication date:
2022

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

RESEARCH ARTICLE

Data citation and the citation graph

Peter Buneman[1], Dennis Dosso[2], Matteo Lissandrini[3], and Gianmaria Silvello[2]

[1]University of Edinburg
[2]University of Padua
[3]Aalborg University

Keywords: bibliometrics, citation graph, data citation

ABSTRACT

The *citation graph* is a computational artifact that is widely used to represent the domain of published literature. It represents connections between published works, such as citations and authorship. Among other things, the graph supports the computation of bibliometric measures such as *h*-indexes and impact factors. There is now an increasing demand that we should treat the publication of data in the same way that we treat conventional publications. In particular, we should cite data for the same reasons that we cite other publications. In this paper we discuss what is needed for the citation graph to represent data citation. We identify two challenges: to model the evolution of credit appropriately (through references) over time and to model data citation not only to a data set treated as a single object but also to parts of it. We describe an extension of the current citation graph model that addresses these challenges. It is built on two central concepts: citable units and reference subsumption. We discuss how this extension would enable data citation to be represented within the citation graph and how it allows for improvements in current practices for bibliometric computations, both for scientific publications and for data.

1. INTRODUCTION

1.1. Citations and the Citation Graph

Citation is essential to the creation and propagation of knowledge and is a well-understood part of scholarship and scientific publishing. Citations allow us to identify the cited material, retrieve it, give credit to its creator, date it, and provide partial knowledge of its subject and quality.

The *citation graph*, or citation network, is a model used to describe how citations link research entities, typically papers, journals, and books (Harzing & Van der Wal, 2008; Tang et al., 2008). It enables a number of important activities such as the following:

- *Exploration of the graph* to find publications of interest.
- *Tracking of authorship* of papers: Citing and following citations is one way to attribute *credit* to authors and to *keep up to date* with the work of others.
- *Dissemination* of research findings: The exploration of citations and cited authors enables the dispersed communities of researchers to share their findings and engage in discussions.
- *Computation of bibliometrics* for the analysis of one researcher, venue, or publication impact in particular fields. The citation graph is the basis for nearly all the currently used bibliometrics, such as *impact factor* and *h-index*.

an open access journal

Check for updates

Citation: Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021). Data citation and the citation graph. *Quantitative Science Studies*, *2*(4), 1399–1422. https://doi.org/10.1162/qss_a_00166

DOI: https://doi.org/10.1162/qss_a_00166

Corresponding Author: Dennis Dosso
dennis.dosso@unipd.it

The MIT Press

Throughout this paper, we refer to an idealized "citation graph" as though it were a real and unique digital artifact that represents papers and the citations between them. Of course, it is not unique: Various organizations have distinct implementations of it. Among these, we count: Google Scholar, the Microsoft Academic Graph (MAG)[1], the Open Academic Graph (OAG) (Tang et al., 2008), Semantic Scholar (SS)[2], AMiner (AM)[3], and PubMed[4] (this is more a linked collection of documents than a full-fledged citation graph), Scopus[5], and the Web of Science[6]. These graphs differ in many aspects, such as their coverage, their being open- or closed-access, and their schema; but in all of these, the basic structure is a *directed* graph, in which the vertices represent publications and the edges represent citations from one publication to another (Price, 1965).

Most of the information about papers is contained in annotations of the nodes. The edges are generally typed but not annotated (an exception is MAG, which carries *context*, as we discuss later). Although in early models, nodes only represented papers and the only edges were "cites" edges, recently, citation graphs have been extended with richer information (Peroni & Shotton, 2020). These extensions may carry author nodes with a "wrote" edge to papers, journal/conference nodes with a "part of" edge from papers, and subject nodes with the corresponding edges. Although representations differ, the purpose is similar: to provide the services described above.

1.2. The Need for Data Citation

Scientific publications increasingly rely on curated databases, which are numerous, "populated and updated with a great deal of human effort" (Buneman, Cheney et al., 2008), and at the core of current scientific research[7]. In this context, references to data are starting to be placed alongside traditional references. Hence, there has been a strong demand (FORCE-11, 2014; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013) to give databases the same scholarly status as traditional scientific works and to define a shared methodology to cite data. Scientific publishers (e.g., Elsevier, PLoS, Springer, Nature) have taken up data citation by instituting policies to include data citations in the reference lists.

The *open research culture* (Nosek, Alter et al., 2015) is based on methods and tools to share, discover, and access experimental data. Moreover papers, journals, and articles should provide access to all the data that they use (Cousijn, Feeney et al., 2019). Researchers and practitioners (e.g., journalists and data scientists) who make use of electronic data should be able to cite the relevant data as they would cite a document from which they had extracted information (Cousijn, Kenall et al., 2017; *Nature Physics* Editorial, 2016). As we shall see, the citation graphs can become a fundamental tool in the pursuit of the goal of accessibility and networking between papers and data.

We also observe that data occupy a crucial role today in research, emerging as a driving instrument in science (Candela, Castelli et al., 2015). Data citations should be given the same

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[2] https://www.semanticscholar.org/
[3] https://www.aminer.cn/
[4] https://www.ncbi.nlm.nih.gov/pubmed/
[5] https://www.scopus.com/home.uri
[6] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[7] See https://fairsharing.org/databases/ for a detailed list of curated scientific databases commonly used in research.

scholarly status as traditional citations and contribute to bibliometrics indicators (Belter, 2014; Peters, Kraker et al., 2016). Principles such as Findability, Accessibility, Interoperability, and Reusability (FAIR) (Wilkinson, Dumontier et al., 2016) require data to be easily findable and accessible, qualities that are more readily available once data can be appropriately cited. In this sense, we can say that the FAIR principles encourage the adoption of data citation.

The reasons given for data citation are the same those given for a conventional citation (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013): recognition of the source (e.g., a title); credit for the author, curator, or agent; establishment of its currency (when it was created); where it was located; and how it was extracted. The last three of these fall under the general heading of provenance and are important when one wants to reproduce some analysis on the data or establish the trustworthiness of a claim.

Data sets and databases are usually more complex and varied than textual documents, and they introduce significant challenges for citation (Silvello, 2018). Text publications have a fixed form, do not change over time, are interpretable as independent units, share a standard format and representation model, and are composed of predetermined, albeit domain-dependent, sets of elements that are considered as *citable* (e.g., the whole paper or book or a chapter). Scientific databases are structured according to diverse data models and accessed with a variety of query languages. What can be cited may range from a single datum to data subsets or aggregations specified by the person or agent that extracts the relevant data, and deciding *a priori* what can and cannot be cited is rarely feasible. Data citation introduces multiple citation types, besides the classical papers citing papers. These are papers citing data, data citing papers, and data citing data.

1.3. Data Citation in the Citation Graph

Our purpose in this paper is to discuss whether, in its current form, the current model of the citation graph can properly accommodate data citation. We claim that, despite all the features and modifications that have been added to various implementations of the citation graph, at least two significant features are generally missing or poorly represented. These shortcomings already limit what we can represent with existing implementations, and we argue that they make impossible the proper representation of data citations.

The first shortcoming concerns the *assignment of credit* when a referenced scientific work is corrected or augmented with another version. A typical example case is that of a preprint paper that gets cited before its peer-reviewed version is published. It is common for the authors to prefer that the preprint citations are merged with those of the peer-reviewed version. Something similar happens also when an updated version of a data set is published.

In the case of data, we need to consider that a database may be composed of multiple independently citable parts (e.g., a single record, a table, a view). Every single citable part can evolve and change over time and obtain citations (also views or downloads, when monitoring other scientometrics signals) at a different point in time. Therefore, it can be necessary to aggregate these statistics over all the versions of the same part to measure its impact and that of the database. The MAG and S2ORC databases have also an explicit notion of multiple versions of a paper, for example preprints and final published versions. It is however uncommon to "move" citations from one version to another, following some criteria or algorithm to correctly allocate citations. Yet, aggregating citation to a single version of a scientific work would have, among other things, the desirable effect of allowing proper evaluation of the impact of the work.

The second feature is the representation of *context* of a citation. Context is required for various reasons. It is typically used to describe the relevant part (e.g., page number) of a *cited* document. It may also carry, as in MAG, the surrounding text within the *citing* document helping to understand the reason for the citation; for example, a simple mention, a confutation, or a validation, such as those described in the OpenCitations ontology (Daquino, Peroni et al., 2020). In the case of data citations, the context can contain the query identifying the cited data, expressed in different format (e.g., a URL, a filename, a SQL or SPARQL query, etc.). Despite a great deal of attention dedicated to the citation context—see, for instance, the Citation Context Analysis (CCA) discussed as early as the 1980s (Freeman, Ding, & Milojevic, 2013)—there is no systematic approach to representing it within citation graphs.

In fact, none of the largest citation-based systems, such as Scopus, MAG, and Google Scholar, properly take into account scientific databases as objects for use in the research literature. Google Data Search[8] allows us to search for indexed data sets, but it does not keep track of the citations to data or other types of statistics, such as clicks or downloads. Web of Science is one notable exception because it models data citations, even though only at the database level, via the Data Citation Index (DCI), now maintained by Clarivate Analytics (Force, Robinson et al., 2016). Note that DCI is not publicly available and the data sets are indexed after a validation process.

Another effort is the Scholix framework (Burton, Koers et al., 2017), which can be regarded as a set of guidelines and lightweight models that can be quickly adopted and expanded to facilitate interoperability among link providers. Finally, an example of an initiative that includes data and databases among the entities of the graph is the OpenAIRE Research Graph Data Model (Manghi, Bardi et al., 2019), which leverages the OpenAIRE services to populate a *research graph* whose nodes include scientific results, organizations, funding agencies, communities, and data sources.

The conventional approach is to treat a data set as a single entity, in the same way, one would treat a scientific publication. However, this is far from ideal as typically only a small part of the data set or database is cited, and the authorship—the people who have contributed to the database—can vary widely with the part of the database being cited (Buneman, Davidson, & Frew, 2016).

In this paper, we discuss the extension of the current model to enable the proper inclusion of data citations in the citation graph; and we discuss the *evolution* of a database: What happens to citations when new versions of the database appear? For the versioning issue, we describe a relation between scientific works (either papers or data) called *subsumption*. Through different policies, this relationship models effectively how credit should be transferred through time when updated versions of data appear in the graph. Finally, we discuss how to introduce data in the citation graph, considering the most common data citation strategies currently used in the world of research. In particular, we take inspiration from one of the solutions proposed by the *Research Data Alliance* (*RDA*)[9]. The RDA is a community-driven initiative launched in 2013 by different commissions. One of its working groups, the "Working Group on Data Citation: Making Dynamic Data Citable" (WGDC), has as one of its goals the identification and citation of arbitrary views of data. As a potential solution, the WGDC recommends an identification method based on PIDs assigned to queries.

[8] https://datasetsearch.research.google.com/
[9] https://www.rd-alliance.org/

The focus of this work is on data citation; but to ease the comprehension of the paper, we first discuss the limitations of the citation graph and the possible extensions we propose by focusing on textual documents, and then we extend the reasoning to data citation.

The paper is organized as follows: Section 2 describes some preliminary concepts and the limits of the citation graphs; in Section 3 we discuss the proposed solutions for the first three issues; Section 4 presents the proposed solution for the introduction of data in the citation graph; Section 5 sums up our main proposals and discusses possible lines of research and development; Section 6 describes the related work; finally, Section 7 presents conclusions and future work.

2. THE CITATION GRAPH: CONCEPTS AND LIMITS

2.1. Core Concepts

2.1.1. Citable unit

By citable unit (CU), we mean a published entity—be it a paper, a chapter, or portion of data—which presents all the qualities necessary to be considered as a "citable work." The characterization of a CU that we use, given in Wilke (2015), requires that: it must be uniquely and unambiguously identifiable and citable; it must be *available* in perpetuity and in *unchanged* form; it must be *accessible*; and it must be *self-contained* and *complete*. Self-contained and complete means that whatever new contribution is contained inside the piece of work, that contribution needs to be fully and clearly explained. This is not always the case for certain publications. Consider the slides of a scientific presentation. As they are used merely as a support for the oral presentation, they often cannot be fully understood without the corresponding talk. Also, the combination slides/registration of the talk may be incomplete, as many presenters tend to skip technical details during their presentations, referring to the complete published work.

Although some of these requirements are subjective, and not straightforward in databases, they still provide a workable starting point. The requirement that is most problematic for databases is that the citable unit must be *unchanged*. Databases evolve rapidly, and creating a citable unit for each version may be counterproductive. This is something we address in Section 4.2. Generally, what constitutes a citable unit is decided by convention. We should also note that some citable units comprise other citable units. The proceedings of a conference may be cited as may be a book on a topic whose chapters are written by different people and may also be individually cited. There is thus a "part-of" relationship between CUs that we discuss later.

In (Daquino et al., 2020) a similar concept, *bibliographic resource*, is defined as a resource that cites and can be cited by other resources.

2.1.2. Reference

At the end of this paper, there is a list of references. Traditionally, a reference is a *pointer* to, and a brief description of, another publication in the literature. It is a short text composed of fields such as title, authors, year, venue, and others, that enables us to identify and find the entity (i.e., a paper, a book, or a survey) being referenced. Depending on aspects of the citing CU's nature, like its field of research, the publication venue, or even language, different attributes of the reference may vary such as the format or the fields composing the reference. In physics, for example, titles are often omitted.

The important point is that, apart from the stylistic rendition of the reference, its contents are determined by the cited CU; hence, to within stylistic variations, the reference to a CU will be

the same in any paper. In this paper, the reference determines the existence of a directed edge between two CUs: the citing and the cited one.

2.1.3. Citation

There is no universal agreement on the distinction between *reference* and *citation*, and the two terms are often used interchangeably (Altman & Crosas, 2014; Daquino et al., 2020; Osareh, 1996; Price & Richardson, 2008).

One distinction proposed in Gilbert and Woolgar (1974) is that "reference" refers to the works mentioned in the reference section or bibliography of a paper. A reference may be mentioned once or many times in an article. Each of these mentions is considered a citation.

The distinction is crucial to our understanding of the citation graph. If we look at what goes in the body of a paper, we may find, for example, "Austen, J. (2004). pp 101–104." We note that this textual artifact contains two parts. The first one is "Austen, J. (2004)," which we call a **reference pointer**. A reference pointer is, in general, a textual means that is used to denote a single bibliographic reference in the reference section when mentioned in the body of a paper. The second part of the citation is composed of some additional information, in this case "pp 101–104," which may help the reader locate specific information within the cited paper. Note that the same reference pointer can occur several times in a paper and may have differing additional information, such as "pp 10–25" and "pp 110–120."

Therefore, we can say that a **citation** is composed of the combination of the reference pointer with the (optional) information added to it in the paper's body. The optional information in the paper's body may be referred to as a form of *context* for the citation. This implies that there is a many-one relationship between citations and references, a fact that is supported by some discussions on the topic, for example "... the second necessary part of the citation or reference is the list of full references, which provides complete, formatted detail about the source, so that anyone reading the article can find it and verify it." (Wikipedia, 2021).

2.1.4. Reference annotation

We shall call this extra information, such as "pp 101–104," *reference annotation*. In this paper, the reference annotation consists of all the information added to a reference pointer to qualify how it is used. This information is not part of the reference and can change depending on how that particular resource is used.

The Citation Typing Ontology (Shotton, 2010) is replete with examples of other kinds of annotations such as "refutes," or "ridicules," which are clearly about the relationship between the citing and cited documents. In the Microsoft Academic Graph (Sinha, Shen et al., 2015), the *context*—the text surrounding a citation in the source document—may be recorded as another form of annotation. The OpenCitations ontology (Daquino et al., 2020) contains a class called *annotation*[10] attached to the in-text citation and to a reference which has a similar role. Here, we do not need to distinguish between the context of a reference pointer and its reference annotation: For our purposes these two concepts are the same, however it may be that certain applications will require some finer distinctions.

These definitions differ slightly from those in Daquino, Peroni, and Shotton (2018) and Daquino et al. (2020), where a reference (called a bibliographic reference) and a reference pointer are *manifestations* of a citation. Moreover, in our example, the part "pp. 101–104" is a reference annotation, whereas in Daquino et al. (2020) it is a *specialization* of the citation.

[10] https://www.w3.org/ns/oa#Annotation

We do not specifically model the concept of specialization, as it can be inferred from the content of the reference annotation. Also, in Daquino et al. (2020) the pointer may include additional information, but the citation does not.

Summing up, we consider a reference annotation as a "box" that can contain information derived from the context of a reference pointer.

Generally speaking, the Citation Context Analysis (CCA), whose basis was first developed in the early 1980s, is the syntactic and semantic analysis of citation content, used to analyze the context of research behavior (Freeman et al., 2013). CCA has been used as a promising addition to traditional quantitative citation analysis methods. One of the main aspects of CCA is that it incorporates qualitative factors, such as how one cites. In Daquino et al. (2020) this idea is captured by the concept of **citation function**, which is the function or purpose of the citation (e.g., to cite as background, extend, agree with the cited entity) to which each in-text reference pointer relates. In our proposal, this qualitative factor, or citation function, can be located in the reference annotation, and it could be inferred from the context of the reference pointer.

Even in a citation graph that represents conventional citations it is necessary to be able to attach information to a reference to create proper citations. Yet, in some citation graph implementations, this is impossible, because the reference relationship is represented as a directed but unannotated edge. As noted above, an exception is the Microsoft Academic Graph, which contains two kinds of edges between publications: unannotated edges and edges annotated with context. The reason for this omission may be the difficulty of collecting the relevant information; it may also be that it is not needed in the computation of most bibliometrics.

2.1.5. Part-of

The *part-of* relationship exists between two citable units in the graph; it describes the situation where one citation unit is somehow "contained" in the other. This is the case of papers published in an instance of a venue (e.g., the 2020 version of the ACM SIGMOD), and these issues being part of the venues themselves (e.g., ACM SIGMOD). This information is present for example in databases such as MAG and AMiner.

In the case of data, the part-of relationship is particularly important. Many databases and data sets have a hierarchical structure and may be cited at different levels of detail.

2.1.6. Database categories and citation

There is a broad spectrum of databases for which citation is appropriate. In discussing data citation it is helpful to divide them into three rough categories.

- *Static databases*, which are used to support claims in a publication. These are typically "one-off" results of a set of experiments. For these databases, systems such as Mendeley[11] store data alongside the publication, so that a citation to the publication also serves as a citation to the data. Data journals (Candela et al., 2015) (i.e., journals publishing papers describing data sets) are also employed as proxies to cite static data sets.
- *Evolving databases* of source data such as *weather data* (Philipp, Bartholy et al., 2010) or *satellite image data* (Shanableh, Al-Ruzouq et al., 2019) that are collected for a wide range of purposes. Zenodo[12], like Mendeley, stores data together with its representative publication. However, a publication about a data set and the data set itself can also have

[11] https://www.mendeley.com/
[12] https://zenodo.org/

separate and unrelated DOIs. In this case the citation to the publication and to the database are distinguished. Moreover, it allows multiple versions of the same database to be deposited, with new DOIs for each one, thus keeping track of usage statistics like the number of downloads and views on each version. A citation to the database, or even to a document that describes the whole database, is generally regarded as inadequate. Usually, only a portion is used; hence, one needs to know the part (the sensor, the location of the image, or the time range) from which the data was extracted.

- Finally, we have *curated databases*. These have largely replaced conventional biological reference works (Buneman et al., 2008), and like the works they replace, involve substantial human effort. One advantage is that they are readily accessible and easy to search. Moreover, there are few limits on their size and complexity, and they can evolve rapidly with the subject matter. For these, the citation is a complex issue but it is just as crucial for curated databases as it is for the reference works that they replace.

The distinction between these three categories is not sharp, and there are many examples that lie in the overlap. For example, most source data databases involve a degree of curation.

2.2. Existing Limitations of the Citation Graph

Although implementations of the citation graph differ, the basic model consists of a directed graph $\mathcal{G} = (V, E)$, where V is the set of papers and $E \subseteq V \times V$ is the set of directed edges corresponding to the citations among them: An edge $\langle p_1, p_2 \rangle$ connects the papers p_1 and p_2, if p_1 cites p_2. The following limitations of this simple model are obstacles to the representation of data citation, but can already be seen in conventional citations to papers.

2.2.1. Lack of context

Although in the basic model of the citation graph the nodes often contain information such as the *title*, the list of *authors*, or the *venue* of publication, it is lacking the information about the *context* of the citation, that is, all that kind of information that could be inferred from the context of the reference pointers, such as the specialization of the citation or the citation function. The only information provided by the edge $\langle p_1, p_2 \rangle$ is that p_1 cites p_2, but it does not specify the *why* or the *how* of this citation. In the literature, we find the *contextual citation graphs*, which make apparent the textual contexts of each citation (Bird, Dale et al., 2008; Daquino et al., 2020; Lo, Wang et al., 2019). These graphs contain information about reference annotations, which is what, in this work, we consider as the citation context.

Note that a lack of citation context is an issue that is related to not only data citation but the whole scientific citation infrastructure and ecosystem. How one document is cited in another, whether cited as a piece of evidence or a tool, could greatly influence how the scientific bibliographic universe is built and how credit should be assigned between researchers.

2.2.2. Versions

Ideally, the papers in the citation graph should only cite papers in the past (i.e., papers that already exist when the new paper is introduced in the graph [Lo et al., 2019]). If this is the case, the citation graph is a DAG (Directed Acyclic Graph).

However, this often is not true because some of the papers in V go through revisions and modifications. This happens for many reasons and with many variations. Among the possible cases: It may be that several copies of one work are to be found on the internet; that one version is an "abstract" and is published in some conference proceedings, and a "full version" is

later published in some journal; or that one version is published in some archive online and then a fully fledged paper is released in a conference or journal.

To receive credit, it is generally in the authors' interest to have these documents seen as one. What appears to happen in Google Scholar, for example, is that all versions are clustered together, and one of them, the "main" version, is selected to be the recipient of all references.

Consider the following situation: document A is published, and a document P citing A is subsequently published. Document B, a revision and possibly an extension of A, is then published, taking A's place in the graph. If this new version B contains new outgoing citations to P, then a cycle is created, and the graph is no longer a DAG ($P \rightarrow A \rightsquigarrow B \rightarrow P$). This problem may be solved by separating A and B.

Another source for cycles in citation graphs that cannot be avoided are papers by the same authors created at the same time (e.g., a full paper written together with a demo paper or extended abstract). In this case, the problem can be solved, for example, by conflating the papers.

Another problem arises when the system, for some reason, decides that B becomes the "main" representative of the publication. In this case, what happens with services such as Google Scholar is that the references first given to A are rerouted to B. This can be confusing as the reference annotation (e.g., the page number) may no longer be valid.

2.2.3. Citations to data

One of the primary roles of data citation is to give credit and attribution to the work of data creators and curators (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). If integrated into the citation graph data citations can be represented and analyzed as if they were conventional citations, with data CUs and corresponding authors receiving citations and thus credit for their work. However, services such as Google Scholar or Scopus do not allow databases into their citation graph.

Data journals (Candela et al., 2015) enable the publication of papers describing a database that works as a proxy for it and its authors and receives its citations. This is a possible solution, but it is not complete as it does not consider citations referring to *general* queries.

To give appropriate credit to the contributors to the various parts of a complex curated database, one approach to data citation (Buneman, Christie et al., 2020) is to automatically create short papers, *citation summaries*, for each citable part of the database and publish them in a dedicated online journal. This enables the contributors to receive proper bibliometric credit for their contributions to the database. In this approach, a new summary for a view is generated whenever that view changes substantially. This summary can then be included in the current implementations of citation graphs and receive citations.

To conclude, unless there is some form of representation of the cited database or the cited query in the form of a paper or journal, current citation graphs do not include databases as nodes and citations to data as edges.

3. EXTENDING THE CITATION GRAPH

We describe two key extensions to the citation graph needed to deal with both the structural complexity and evolution of databases. These extensions already exist in a limited form in some implementations of the citation graph. However, we need to specify them precisely and understand how they help with the limitations described above and with data citations.

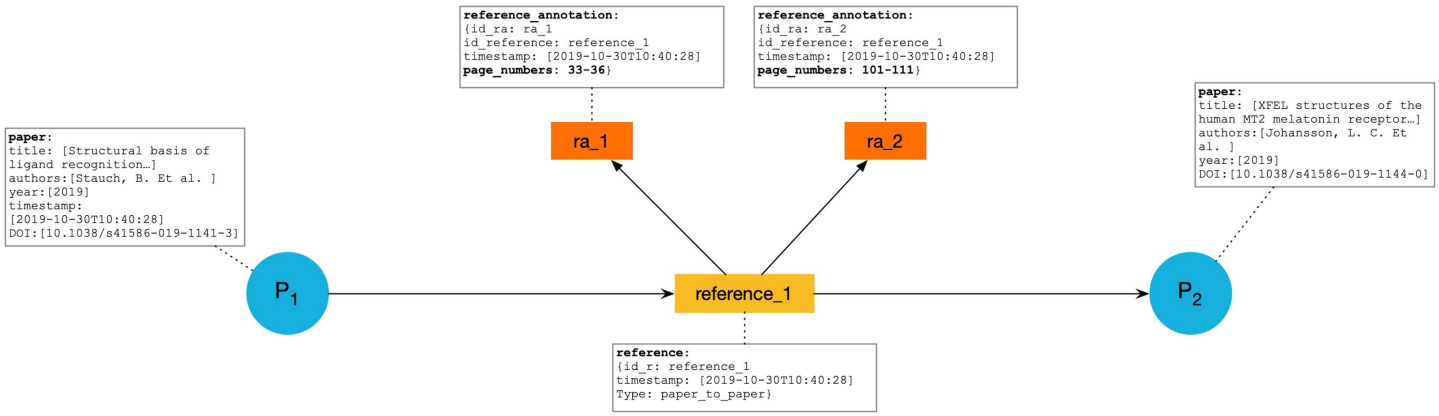

Figure 1. Use of references and reference annotations. Each reference is an edge connecting one citing unit to the cited one, and, if it exists, it is unique. One reference may have one or more reference annotations, each giving rise to a citation.

What we propose is independent of any specific implementation of the citation graph and, for the most part, it can be incorporated as extensions to those implementations rather than requiring a completely new implementation of the supporting database.

3.1. Reference Annotation

As discussed above, a reference is represented by an edge in the citation graph. However, to represent a citation accurately, we need to add *reference annotations*. That is, we need to annotate the edges. Unfortunately, most data models currently implemented do not support data on edges[13], so for consistency with these models, our diagrams include a new kind of node rather than a new kind of edge.

Consider Figure 1. Two papers, P_1 and P_2, are represented with circular nodes. We use these nodes to represent citable units. They are annotated with all the information that usually constitutes one reference, such as title, authors, year of publication, journal name, and DOI.

In this example P_1 references P_2. We can imagine the reference appearing in the "References" section of P_1 as something similar to "Johansson, L. C. et al. (2019). XFEL structures of the human MT 2 melatonin receptor reveal the basis of subtype selectivity. Nature, 569 (7755), 289–292. doi: 10.1038/s41586-019-1144-0." The use of this reference in the paper is reflected by the presence of the reference edge between P_1 and P_2 and the reference node `reference_1`. This is a different kind of node, which contains information such as the edge type (reference), the timestamp of when the citation was registered by the system and the type of reference (in this case from a paper to another paper). The actual information contained by the node can be modeled according to whatever model we decide to follow (e.g., the aforementioned Open Citation ontology).

Suppose now that P_1 cites P_2 twice. Each time, it does not merely refer to the whole paper P_2, but specific parts of it. The node `reference_1` has two other neighbor nodes, called *reference annotation nodes*, `ra_1` and `ra_2`. These two nodes contain the information describing the reference annotations found in P_1 used to cite P_2, such as the context, references to particular tables or images, comment on the nature of the citation (e.g., that the authors of P_1 agree or disagree with P_2). In the example, these annotations carry page numbers. Hence, the combination of `reference_1` with `ra_1` makes one citation.

[13] Property graphs are an exception because they allow data to be assigned to edges.

Reference and reference annotation nodes are the addition that we make to the citation graph to face the first problem.

3.2. Subsumption

Often new documents take the place of older versions, becoming also the recipients of both new and old citations. This behavior is handled behind the scenes by some existing implementations of the citation graph (notably Google Scholar). To deal with this phenomenon transparently, we propose the introduction in the citation graph of a new relation, called *subsumes*.

In Figure 2 we see a situation similar to the one of Figure 1, where P_1 is citing P_2 at time 1. Now, imagine that a new version of the same paper, P'_2, is published and inserted in the citation graph at time 2. The reference for P'_2 should also have a version number or something that distinguishes it from P_2. The relation *subsumes* between P'_2 and P_2 indicates that the former is a new version of the latter, and is, from now on, *the* paper to consult and reference.

In some scientific areas, a journal "paper" P'_2 may be treated as a version of an earlier conference "abstract" P_2, even though the two differ substantially. Because of this we do not want to destroy the original link from P_1 to P_2; to do so would be to "rewrite history" and remove

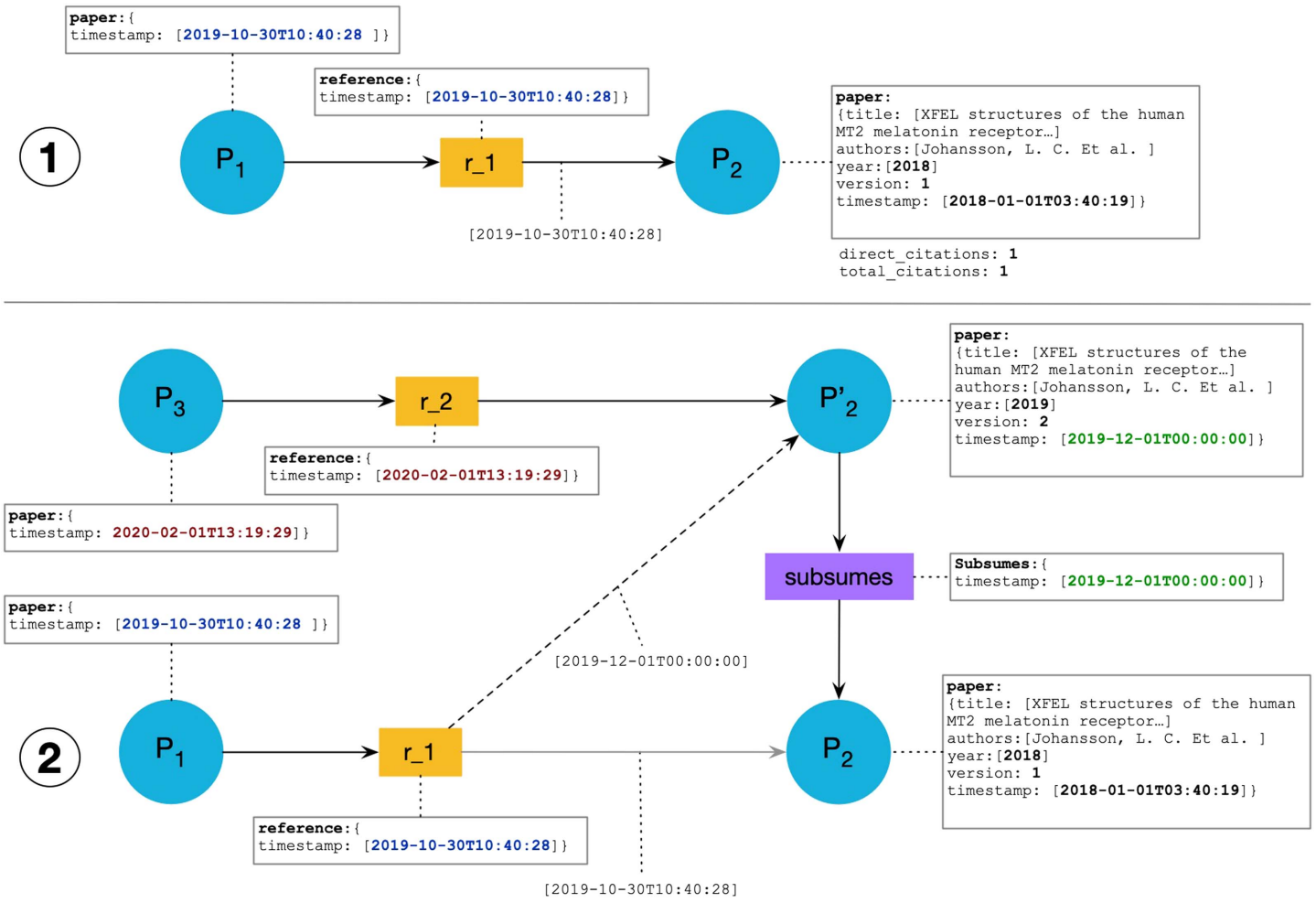

Figure 2. The "subsumes" relation between two CUs.

information from the graph, and we strongly feel this should not be the case with the citation graph. The subsumes relation is present to indicate that one paper is a version of another and, crucially, that author credit can be transferred from the subsuming paper to the subsumed paper. On the one hand, the transfer of credit enables a more comprehensive measure of author contributions (e.g., increasing the number of citations on the latest version of the publication). On the other hand, credit transfer also transparently reflects the impact that the publication, seen as the aggregation of its different versions, has on the research community. Different types of subsumption can be defined, such as the kind of subsumption that propagates the citations to the single papers to their journal, thus computing its impact factor.

It would be wrong to transfer the credit for writing a paper to more than one other paper, so the subsumption relation is many-one. It is necessarily acyclic, thus it is a *forest* with the roots of the trees in that forest being the papers that are designated to receive the credit. It may be useful to have a term for a root node on the subsumption graph, perhaps *primary citable units* (PCU). It is interesting to note a similar approach in the MAG[14], which lists the CU P_2 under the PCU P_2', keeps the citation count for P_2 and P_2' distinct, and reports, for example, "124 citations" for P_2, "325 citations" for P_2' but adds, to P_2', "449 citations for all."

4. DATA IN THE CITATION GRAPH

Here we discuss how we place databases in the citation graph. We shall find that the two extensions we have discussed—edge annotation and subsumption—are essential to accommodating databases. In particular, they allow us to deal with databases, which tend to be updated and thus change much more frequently than papers. We could treat each version or instance of the database as a distinct document, but—at least for author credit—this would be a limitation, if not counterproductive.

First of all, we use the term "database" in the most general sense to refer to a conventional relational database, an ontology, some form of graph database, or a database that is a collection of files (Buneman et al., 2016). One might then say that anything one has termed a database is a citable unit. The problem is that *parts* of the database may also be citable units. The reason we need to discuss parts of the database is twofold: First, where in the database one finds something is, like page numbers, a form of location or partial provenance; the second authorship may vary with what part of the database is being cited (Buneman et al., 2016).

With "part" of a database we intend a *view* (Buneman et al., 2016). A view is a query which we again generalize to being anything from a relational query for a relational database, a directory path or URI for a collection of files, or some query in one of the several languages that have been developed for ontologies and graph databases. It is assumed that the database administrators will define these views and hence the citable units. MODIS (Justice, Vermote et al., 1998) is an example of a large evolving database of Earth images for which various subcollections have different authorship; and GtoPdb (Southan, Sharman et al., 2015) is a complex curated relational database in which authorship is represented within the database and can be assigned to views determined by the curators.

4.1. Part-of and Reference Annotation

Consider, for simplicity, the case in which the database is static, or that we are only interested in representing citations for one version of the database (we address the more complex case of dynamic databases in the next section). The first observation is that by defining the CUs as

[14] https://tinyurl.com/y9clyx8d, retrieved March 16, 2020

views, we immediately obtain a part-of relationship: view V_1 is a part of view V_2 if V_1 can be answered from the result of V_2. Formally, V_1 is part of V_2 if there is a query Q such that for all possible instances of the database, $V_1(D) = Q(V_2(D))$.

We have already discussed reference annotations and the information they carry. Among other things, they contain information about *where* in the cited document the relevant information being cited is to be found. If we look at data citation, this notion of location has much greater importance. For example, the DataCite schema (DataCite Metadata Working Group, 2016; Starr & Gastl, 2011) contains the support for the depiction of geospatial data, with properties such as `GeoLocation` and in particular the subproperty `GeoLocationBox`, which specifies a *bounding box*, that is the spatial limits of a box. Most generally we can describe the "location" in the database as a *query* that extracted the relevant information. This is the approach taken in systems that provide accurate provenance (Pröll & Rauber, 2013). It meshes perfectly with what we are suggesting: The query used to extract the data is a fundamental part of the data citation itself, and the query—or possibly a URL which contains that query—is an essential part of the reference annotation in the citation.

Many approaches can now be defined to decide how to introduce the CU corresponding to data in the citation graph. Here we explore two possibilities, stemming from two of the most used strategies in the research world today. We exemplified these two possible strategies in Figure 3.

In Figure 3A we see that a database is represented with a node, DB_1. A whole database is a citable unit, and every time a paper wants to cite data in that database, it cites the entire database. The reference annotations contain the queries to get the cited data. The paper P_1 presents two citations to DB_1. Therefore, it has one reference and two reference annotations containing the two different queries being used. P_2 is citing DB_1 only once. The total count of citations to DB_1 is two in this case.

With this solution DB_1 is the only recipient of citations. This means that its number of citations can become very high. On the other hand, it may happen that the rightful authors and curators of the parts of the database being actually cited do not receive any credit for their work.

In Figure 3B we see the strategy adopted by the RDA. Every time a paper cites a data subset extracted through a new query, a citable unit is created in the citation graph; we represent this CU as a view, corresponding to that query. In this case, P_1 is citing DB_1 twice by using two different queries, thus there are two distinct references, corresponding to the two views being cited, V_1 and V_2. P_2 citing DB_1 with the query Q_3, generates another view (i.e., V_3).

With this solution, new views are created every time it is necessary. This can produce an explosion in the number of nodes in the citation graph, many of which receive only one citation. However, in this way it is possible to cite the exact set of data extracted by the query and to give the credit for the citation to the rightful authors of that data. Moreover, the three views of the example are connected to DB_1 by a "part-of" relationship. This means that DB_1 may inherit all their citations when needed.

We note that to assign only a single CU for the whole database and dispense with the part-of relationship fails when, for example, authorship varies with the part of the database being cited. This is the case with both MODIS and GtoPdb. In this case, we reiterate that it is up to the curators or database administrators to determine the views that define the CUs.

In the case of GtoPdb, both the curators and the contributors agree that the PCU should be the data summary (Buneman et al., 2020) for the most recent view of the database. Unfortunately, the PCU is not determined by the curators but by the system that scans the dedicated

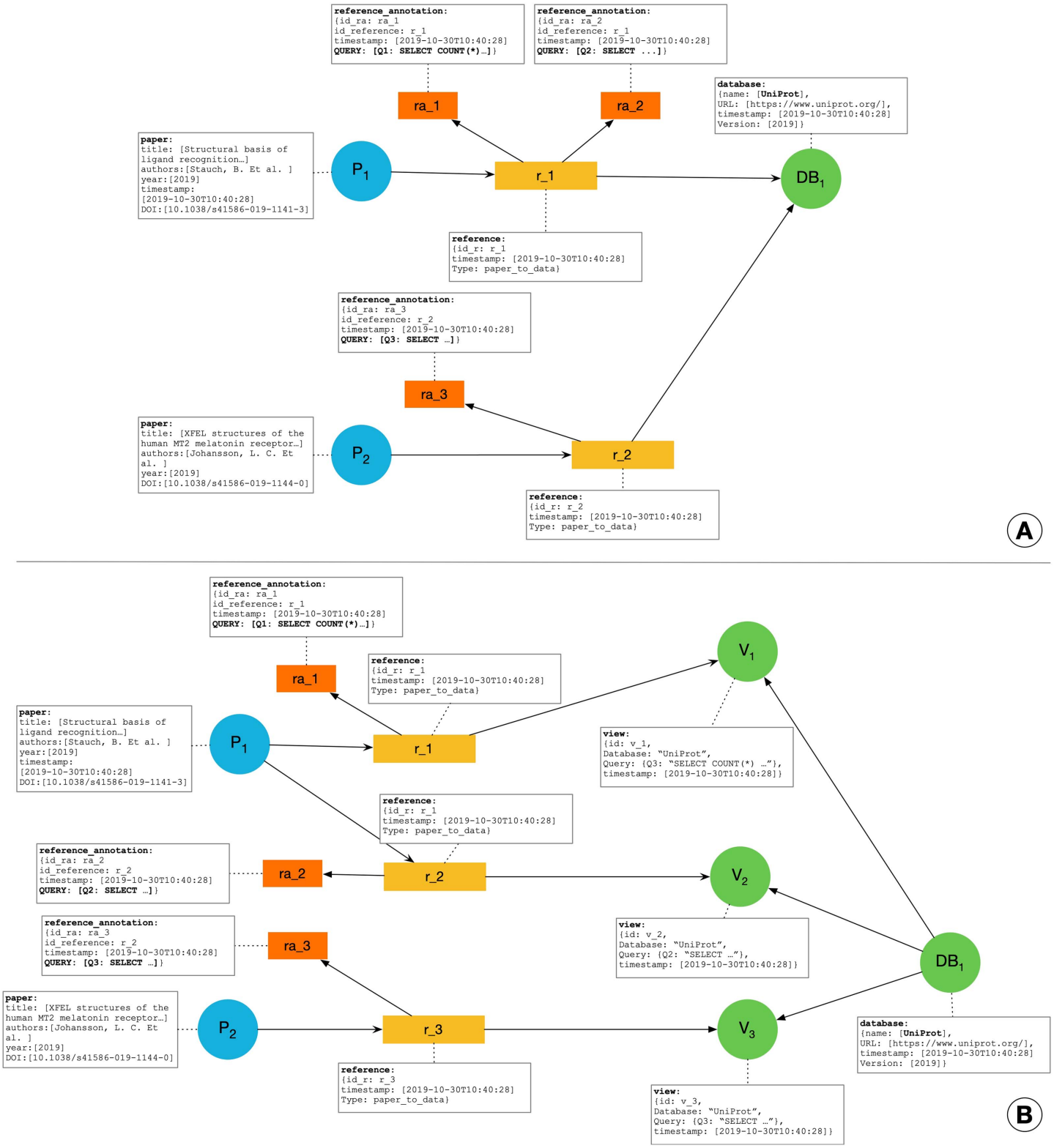

Figure 3. Two examples of possible strategies when citing data. A: Always cite the database. B: Create a view for every new query issued, if it does not already exist, and cite that view.

journal and creates citation graphs. For a given database, it is the responsibility of the curators or administrators to determine the subsumption relation. Even for conventional publications, we believe that the subsumption relation should be determined by the authors and publishers.

4.2. Dealing with Dynamic Data: Subsumption for Data

Most databases are not static. Unlike documents, they are expected to evolve over time. If versions of a database are released, say, every year, it might be appropriate to treat each version as a new CU. On the other hand, as we discussed, a database in the Citation Graph can present a hierarchy of CUs connected among them through the part-of property. Even though a database may change rapidly, the result of a view, part-of a database, may remain unchanged. The lower a CU is in the part-of hierarchy, the less frequently it will change. Also, even if a part-of CU does change, we may want to treat it as a new *version* of the previous CU rather than an entirely unrelated new CU, just as we treat an extended or improved version of a paper.

Given these observations about the introduction of dynamic data, it is necessary to answer these questions:

- When is it necessary to introduce a new CU representing a view?
- If a new CU has been introduced, when can it be considered a new version of the previous CU or an entirely new entity?
- If a new CU has been introduced, how do we connect older CUs with the new ones and still keep track of their citation counts?

The answer to the first two questions can only be given by the database administrators. Every time a new version of the database is released, the administrators go through the different CUs that compose the part-of hierarchy of the database and decide which ones need a new version. Recall that subsumption was needed to transfer credit, in the case of papers, from one CU to another: the primary CU (PCU) (i.e., the root of the part-of hierarchy). The same can be done with data.

As we have defined CUs by views when the database changes, we only need to consider creating a new CU if the view changes. More precisely if D and D' are successive versions of the database and V is a view, if $V(D) = V(D')$ the reference for $V(D)$ needs no change, and no new CU is necessary. However if $V(D) \neq V(D')$, we may want to create a new *CU*.

Once it has been decided that a new CU needs to be created, it is necessary to determine whether the CU associated with $V(D')$ is a new version of the CU for $V(D)$, or whether it is, instead, an entirely new CU. The model we propose accommodates both the possibilities; again, this is something that the database administrators or curators can decide. If the content is different in the sense that there is some kind of structural change, then an entirely different CU may be appropriate. Moreover, if the authorship changes, then a different CU may be desirable, as the two versions of the same CU are typically expected to have the same authorship. These are only two examples of reasons why the DBAs may decide to consider the new CU a new, independent, entity.

On the other hand, normally the change will be such that we want the CUs associated with $V(D)$ and $V(D')$ to be versions of each other, and the PCU can now become the later version $V(D')$. This preserves the accuracy of the references and allows credit to accumulate on the latest version of the view.

In this second case, it is possible to connect the CU representing $V(D)$ to the one representing $V(D')$ through the subsumption relationship. This new relationship has precedence over the part-of relationship, and thus new citations to the older version will be propagated to the new CU, and not upward to the older hierarchy.

5. DISCUSSION

Because citation graphs are currently unsuited for representing databases as first-class citizens, we have proposed how to instead extend them to represent data citation in the citation graph. Among other things, this allows us to capture the many citations given to databases and to give credit to the relevant authors or contributors. The new model that we propose is based on a few adjustments, and builds on emerging practices in the world of data citation. Above all, it has the goal of enabling easy adoption, as it is proposed as an extensions of existing models without requiring drastic changes. We argue that, with these extensions of the current model for citation graphs, we can fully achieve the goal of enabling data citation without jeopardizing the existing infrastructures.

The main limitation of existing models on which we have focused are the lack of context on citations between citable units; the inability to deal with different versions of the same CU; and consequently the inability of introduce data, data evolution, and data citation (down to citable portion of a data set) in the citation graph. We showed how, by solving the first two problems through the introduction of reference annotations and the subsumption property, we are also able to model data citation appropriately in the citation graph.

Unlike traditional scholarly publications, databases present a greater range of granularities and are subject to more frequent change. Concerning the granularity of data, although it is possible to consider various scenarios, we work with two main cases: either only the whole database is treated as a node, or each time a new query is issued, a new node is added to the graph, connected to the whole database through a part-of relationship.

The first solution is similar to what already happens with papers in data journals. In this case, the whole database is represented through a single CU (i.e., one node in the graph). Every time a paper cites data in the database, the citation goes to the database. Information such as the query and the rightful authors of the citation may be inserted in the reference annotation of the citation. This solution is simple, but gives all the citations to the whole database, thus without explicit recognition for the rightful curators of the cited data. Therefore, more computations are necessary to obtain the citation counts of the single queries and the corresponding authors.

With the second solution, which follows the RDA specifications (Rauber, Asmi et al., 2015), every time a new query is issued to the database, a new CU (hence, a new node) is created. In this case, the graph represents explicitly what is cited, and thus the rightful owners receive their citations without further computations. However, this solution may result in an explosion of nodes. To mitigate this problem different techniques could be deployed. For example, it could be possible to use algorithms of query containment to decide when a query behind a citation can be answered from a CU already deployed. In this case, that CU could receive that citation, instead of creating a new node. Of course, query containment is, in general, an NP-hard problem, and it could become computationally prohibitive to exploit this solution, in particular in situations where many nodes have already been created. Alternatively, the system could present to the interested user a series of precomputed queries, corresponding to already instantiated CUs, which may suit their citing needs. In this way, the system already knows to which node to assign the citation.

We also observe that it could be possible to extend the proposed data model where, instead of nodes presenting the metadata of the papers, the CUs are represented using or including the annotated full text of a paper. In this way, annotations on the paper can be used to keep track of different types of information, such as references and reference annotations. Although this solution has greater expressive power, it also increases the size and complexity of the model. As already discussed, the model proposed in this paper has the advantage of being easy to implement on top of already existing systems. A new model, considering the whole annotated text of a paper presents new implementation challenges, and thus requires the creation of a new application from scratch.

It is important also to note how, as of today, there are many challenges to the implementation and proper operation of data citation in general. Often the RDA guidelines for dynamic data citation are not implemented by many databases; it is often difficult to automatically produce context and thus reference annotations that are machine readable; and there are also many bad practices among researchers, such as that of depositing PDFs, images, and tables of their papers in data repositories, calling them research data. Although there are still many hindrances to the correct implementation of data citation, the research community has still showed a great desire for the implementation of common techniques and best practices for the correct application of these guidelines. Databases such as Eagle-i[15] already provide data citation snippets, whereas others, like GtoPdb, automatically produce PDFs of their pages to allow an easier citation of their data in form of CUs. Thus, it is our conviction that as data citation gets more traction and is implemented appropriately, it would be crucial to account for it and integrate such information in the common citation graph. In particular, a model such as the one we propose in this paper will allow a better and fairer implementation of data citation to be achieved, and will also benefit all researchers and become more and more needed as we transition toward the fourth paradigm of science. The more we learn about the current limits of data citation and how to address them, the faster we will come to the final goal of a correct system for citing data.

Considering new possible research problems, we note that the citation graph in fact is, among other things, a *historical record*, that is, a record of how researchers interacted with information and other works to build their expertise and new knowledge. Given this interpretation, then the graph should not be rewritable, that is, it should *not* be possible to *rewrite history*. Therefore, the graph should be a timestamped "append-only" structure in a way similar to the distributed ledgers. Thus, it should only be possible to insert data in it without the possibility to overwrite or modify already existing information.

Among others, these requirements are necessary for the computation of impact factors (Garfield, 2006) where it is necessary to know the number of citations received by a journal in the past 2 years. It is therefore mandatory that this information is not modified over time. This is true also for other types of statistics that researchers may be interested in.

In our examples, we have taken care to timestamp every element of information to make this possible. The timestamps in particular indicate the moments the events "occurred" (e.g., when a citation happened), not when they were inserted in the graph. However, there are several issues concerning the semantics and representation of temporal information in the citation graph that require further investigation.

If this property is correctly implemented, it should enable one to perform different types of query on the graph. That is, past versions of the database should be accessible for accurate

[15] https://www.eagle-i.net/

provenance. Ideally, given the state of the graph in the present, it should be possible to rebuild a previous state at any given time in the past. We call this property *history preservation*.

Several databases have this property. Weather data and geospatial data are generally accumulative (Justice et al., 1998). Blockchains are also based on the idea that once added, a block cannot be removed or modified, to guarantee the preservation of the history of the transactions.

On the other hand, curated databases are not, in general, history preserving, in the sense that they are updated and change with time. This is particularly problematic for data citation because one of its desiderata is that a citation should always allow retrieving or at least knowing what was cited (Buneman, 2006). Therefore, we see the correct extension and implementation of history preservation as an important future challenge to be tackled in the implementation of a data-aware citation graph.

6. RELATED WORK

6.1. Databases in Relation to Data Citation

As we mentioned above, there are three main categories of databases that can be cited: static databases; evolving databases; and curated databases. As a reasonable generalization, the problem of data citation is easily solved for the first category, as many systems and practices have been developed for static databases. In this case, databases are treated as they were traditional publications because they are never updated, the list of authors does not change, and even though only a portion of the database is cited, the citation goes to the whole database. In this case, when we consider the citation graph, we have one single node representing a database receiving all the citations from papers and data.

For the other two cases, data citation remains problematic. One relevant open issue is the citation of data subsets generated on the fly by issuing general queries to the database. In this case, the main problems are how to guarantee the persistence and accessibility of the data in the cited form and automatically provide a complete and correct textual reference for the cited data.

The first problem is tackled by the RDA[16]. The RDA is a community-driven initiative launched in 2013 by different commissions, including the European Commission and the US government's National Science Foundation. Its goal is to build the social and technical infrastructures to enable open sharing and reuse of data. The RDA "Working Group on Data Citation: Making Dynamic Data Citable" (WGDC)[17] (Rauber, Ari et al., 2016) has been working in recent years on large, dynamic, and changing data sets. Although the WGDC first focused on RDBMs as the first forms of pilot solutions, many other types of databases followed (XML, CSV, files, Git repositories, distributed databases such as VAMDC (Zwölf, Moreau et al., 2019), and multidimensional data cubes such as NetCDF/CCCA (Schubert, 2017)). The working group has finished the development of its guidelines, and has now moved on into an adoption phase.

In particular, among the goals of the RDA WGDC (Rauber et al., 2015), there is the identification and citation of arbitrary views of data. As potential solution, WGDC recommends an identification method based on assigning PIDs to queries, that are then used as proxies for the data subset to be cited. The access to a data subset is enabled by reissuing the stored query and

[16] https://www.rd-alliance.org/
[17] https://www.rd-alliance.org/groups/data-citation-wg.html

a citation is associated with the PID of the query identifying the data (Rauber et al., 2016). A PID is an identifier meant to uniquely and persistently (i.e., continually during the course of time) identify an object such as a publication, data set, or person, usually in the context of digital objects that are accessible over the internet. Considering the citation graph, this method based on PID adds a new citable unit every time a new query is cited and requires to check query equivalence (and/or containment) to avoid the creation of a new citable unit for an already cited query.

The second aspect is characterized as a computational problem (Buneman et al., 2016) and some solutions based on "query rewriting using views" (Davidson, Buneman et al., 2017) have been proposed, targeting general queries citations for relational databases (Alawini, Davidson et al., 2017b; Wu, Alawini et al., 2018; Wu, Alawini et al., 2019) and graph databases (Alawini, Chen et al., 2017a).

Overall, most approaches do not consider the evolution of data and the fact that databases are not monolithic objects. When those features are considered, some of the existing models propose the trivial solution of treating databases and views as standalone objects. In our model, instead, we explicitly model citable units and their subsumption relationships, which allow the appropriate distribution of credit.

6.2. Available Citation Graphs

The citation graph, or citation network, as a model of a graph where the vertices represent academic papers, has long been in use in the literature (Price, 1965) and has evolved considerably. There are different implementations of citation graphs, which favor certain aspects of the information regarding publications, citations, and authors, depending on the considered task. Some of them are provided explicitly for navigational purposes (e.g., the Open Academic Graph (OAG)). Others, instead, are the backbone of services allowing search and exploration of scholarly works; these are the Microsoft Academic Graph (MAG), Google Scholar, PubMed, Web of Science, Scopus, and Semantic Scholar.

The Microsoft Academic Graph (MAG) (Färber, 2019; Wang, Shen et al., 2019) is the backbone of the Microsoft Academic Service (MAS), and its nodes represent five different entities: field of study, author, institution, paper, venue, and event. An RDF version of MAG, called Microsoft Academic Knowledge Graph[18] (MAKG) is also available and connected to the Linked Open Data cloud.

The Open Academic Graph (OAG)[19] is an open-source citation graph generated from the linking of two other large academic graphs: MAG and ArnetMiner (or AMiner) (Wan, Zhang et al., 2019) (a free online service used to index, search, and mine big scientific data), designed to search and perform data mining against academic publications available on the Internet. This graph contains entities similar to those of MAG, and it can be used as a unified sizable academic graph for the study of citation networks, paper content, and the integration of multiple academic graphs with different fields and information.

The OpenAIRE Research Graph (Manghi et al., 2019) is the implementation of a fully fledged Open Science Graph. It is a collection of metadata and links connecting research entities, including articles, data sets, software, etc., together with other elements such as organizations, funders, funding streams, projects, research communities, and data sources[20]. The

[18] https://ma-graph.org
[19] https://aminer.org/open-academic-graph
[20] https://graph.openaire.eu

graph today contains around 110 million publications, 10 million data sets, 180,000 software research products, and 7 million other products with 480 million links between them. The aim of the OpenAIRE RG is to bring discovery, monitoring, and assessment of science into the hands of the scientific community (Fava, 2020).

The PID Graph (Fenner, 2020; Fenner & Aryani, 2019) is another example of implementation of a citation graph based around the concept of *PID* (Persistent IDentifier). The PID Graph targets citations aggregation: for all versions of a data set or software source code; for all data sets hosted in a particular repository, funded by a particular funder, or aggregated by a particular researcher; and for a research object, such as a publication or the data used in the paper, together with the software and samples used to create the data set. The PID graph adopts the outputs of the RDA WGDC. One peculiarity of the PID graph is that it includes not only metadata about connections but also metadata about the resources and implicit relations about resources identified by the PIDs. This enables queries based on these metadata, making them more expressive.

Google Scholar, PubMed, Web of Science, and Scopus are all relevant services providing citation graphs, but their data is not directly accessible as a graph.

Google Scholar is an open general-purpose graph focusing on traditional publications and covering multiple languages and publication venues. PubMed, instead, focuses on medicine and biomedical sciences (Roberts, 2001). It covers medical bibliography from 1949 until today, with abstracts, review articles, and free full-text articles. Web of Science (WoS) provides subscription-based access to multiple databases with comprehensive citation data for many different academic disciplines (Falagas, Pitsouni et al., 2008). Finally, Scopus is Elsevier's abstract and indexing (closed) database featuring open access titles, indexes of web pages and patents, and links to both citing and cited documents (Burnham, 2006). Although PubMed is an important resource for clinicians and researchers, Scopus covers a wider journal range, offering also the capability for citation analysis, although limited with respect to WoS, which covers articles published before 1995. Google Scholar, on the other hand, presents all the pros and cons of a web search engine: It can help in the retrieval also of oblique information, but it may present inadequate and less often updated citation information (Falagas et al., 2008).

Semantic Scholar is a project developed at the Allen Institute for Artificial Intelligence and is an AI-backed search engine for scientific journal articles. It uses a combination of machine learning, natural language processing, and machine vision to produce a semantic analysis of the papers of the network and to extract figures, entities, and venues from the documents. It is designed to highlight the most important and influential articles and to identify the connections between them (Fricke, 2018).

As we can see, many of these graphs and systems could work as good starting points for the implementation of the proposed model. MAG and OAG already present the context, which can be used as reference annotation, but lack the ability to accurately cite data and manage their versions. On the other hand, the OpenAIRE graph is able to deal with granularity and different versions, but it still lacks the possibility to cite its data with reference annotations; thus *de facto* it is still unable to deal with data citations. Nonetheless, many of the systems implemented are close to the proposed model, and usually they lack one aspect (like the versioning of the data or the presence of context). Therefore, we believe that a viable way forward would be to implement the approach we propose on top of the already existing infrastructures.

Applications of the citation graphs are disparate. Some examples include prediction of user queries over the graph; recommendation systems for the generation of suggestions leveraging

the relationships across the different types of entities; exploration of papers, researchers, affiliations, and other entities; data integration; data analysis; and knowledge discovery of scholarly data through expert finding, geographic search, trend analysis, review recommendation, association search, course search, academic performance evaluation, and topic modeling (Wan et al., 2019).

Given the vital role of citation graphs and data citation, we argue that it is of crucial importance that existing citation graphs be extended with the appropriate tools to model data citation in various forms. Most of the existing citation graph do not expose their internal data model. Nonetheless, we can see they focus on the same core assumption that citable objects are atomic elements with no citable portions and where evolution through time is not considered. Hence, none of the models above tackle explicitly and directly the issues linked to the task of modeling databases and subsets of databases, as well as the evolution of citable elements through time, which is instead the goal of this work.

7. CONCLUSIONS AND FUTURE WORK

Starting from a basic model of the citation graph in which the nodes are the papers, and the edges are the citations between them, we highlighted three limitations of this model. They are the lack of context for citations, that is, information about the *how* and *why* the citation is used along with which part of the referenced object is used; the absence of a unified strategy of management of the versions of the papers in the graph; and the difficulty of representing citations to databases and data generated by queries in the graph.

To deal with these limitations, we proposed an implementation-agnostic model that includes reference annotations. These annotations contain the context of a citation (e.g., the page numbers of the citation, the query issued to obtain the data, or the considered bounding box).

We also discussed the subsumption property, which is used when a new version of a paper or a database is introduced in the graph. This property indicates that the new version "takes the place" of the previous one for the purpose of assigning credit. The old citations can be inherited from the new version or, depending on the context, such as situations where authors have changed, different policies can be put in place.

Using these extensions to the basic model, we discussed how to represent data citation in the graph. Although important, this work is preliminary, and further work is needed if we are fully to incorporate citations or other kinds of cross-reference between databases into the citation graph.

- Although we have used subsumption partly to deal with the evolution of citable units within databases, we believe there is much more to be said about evolution in databases and in the citation graph itself. We believe that all scientific databases should support "time travel": it should be possible to ask queries on some previous state of the database as easily as one asks queries on the current state. For many databases, especially "source data," it is important to support longitudinal queries, and this is true of the citation graph itself.
- We have dealt with citations *to* databases, but what about citations *from* databases? If, as happens in many curated databases, conventional citations are included within the database, then there should be few problems, but what happens when a part of one database is created by a query from another database? How is the citation represented; and how is it included in the citation graph?

- Finally, once we have properly supported databases within the citation graph, what kinds of bibliometric measures are possible? We have, for example, *h*-indexes and impact factors for conventional publications. How can we appropriately measure the impact of databases?

We note that there is currently marginal interest to cite software and code, even though interesting initiatives, such as the FORCE 11 working group[21], have been taking place and research groups are working on the topic (Alliez, Di Cosmo et al., 2020; Katz, Niemeyer et al., 2016; Katz, Bouquin et al., 2019). This task presents a new set of problems, in particular regarding authorship, because code is often copied or adapted from other repositories, passing from hand to hand, undergoing modifications. The characteristics of the life cycle of software open a whole new set of problems and research questions about who is the righteous author of that piece of cited code and who should receive credit from the citation.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their detailed comments and suggestions.

AUTHOR CONTRIBUTIONS

Following the CRediT guidelines[22], all authors contributed equally to the conceptualization, investigation, methodology, and writing of the paper.

COMPETING INTERESTS

The authors declare that they do not have any competing interest.

FUNDING INFORMATION

The work was partially supported by the ExaMode project, as part of the European Union H2020 program under Grant Agreement No. 825292. Matteo Lissandrini is supported by the European Union H2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 838216.

DATA AVAILABILITY

Not applicable.

REFERENCES

Alawini, A., Chen, L., Davidson, S. B., Portilho Da Silva, N., & Silvello, G. (2017a). Automating data citation: The eagle-i experience. In *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017* (pp. 169–178). IEEE Computer Society. https://doi.org/10.1109/JCDL.2017.7991571, PubMed: 29599662

Alawini, A., Davidson, S. B., Hu, W., & Wu, Y. (2017b). Automating data citation in CiteDB. *Proceedings of the VLDB Endowment, 10*(12), 1881–1884. https://doi.org/10.14778/3137765.3137799

Alliez, P., Di Cosmo, R., Guedj, B., Girault, A., Hacid, M. S., ... Rougier, N. P. (2020). Attributing and referencing (research) software: Best practices and outlook from Inria. *Computing in Science Engineering, 22*(1), 39–52. https://doi.org/10.1109/MCSE.2019.2949413

Altman, M., & Crosas, M. (2014). The evolution of data citation: From principles to implementation. *IAssist Quarterly, 37*(1–4), 62. https://doi.org/10.29173/iq504

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLOS ONE, 9*(3), e92590. https://doi.org/10.1371/journal.pone.0092590, PubMed: 24671177

Bird, S., Dale, R., Dorr, B. J., Gibson, B. R., Joseph, M. T., ... Tan, Y. F. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*. European Language Resources Association. https://www.lrec-conf.org/proceedings/lrec2008/summaries/445.html

[21] https://www.force11.org/group/software-citation-working-group

[22] https://casrai.org/credit/

Buneman, P. (2006). How to cite curated databases and how to make them citable. In *18th International Conference on Scientific and Statistical Database Management* (pp. 195–203). https://doi.org/10.1109/SSDBM.2006.28

Buneman, P., Cheney, J., Tan, W.-C., & Vansummeren, S. (2008). Curated databases. In *Proceedings of the 27th ACM-SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 1–12). https://doi.org/10.1145/1376916.1376918

Buneman, P., Christie, G., Davies, J. A., Dimitrellou, R., Harding, S. D., … Wu, Y. (2020). Why data citation isn't working, and what to do about it. *Database, Volume 2020*, 2020, baaa022. https://doi.org/10.1093/databa/baaa022, PubMed: 32367113

Buneman, P., Davidson, S., & Frew, J. (2016). Why data citation is a computational problem. *Communications of the ACM*, *59*(9), 50–57. https://doi.org/10.1145/2893181, PubMed: 29151602

Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, *3*(1), 1. https://doi.org/10.1186/1742-5581-3-1, PubMed: 16522216

Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., … Schindler, U. (2017). *Scholix metadata schema for exchange of scholarly communication links*. Geneva, Switzerland: CERN. https://doi.org/10.5281/zenodo.1120275

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762. https://doi.org/10.1002/asi.23358

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, *12*, CIDCR1–CIDCR7. https://doi.org/10.2481/dsj.OSOM13-043

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, *18*(1), 9. https://doi.org/10.5334/dsj-2019-009

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., … Clark, T. (2017). A data citation roadmap for scientific publishers. *bioRxiv*. https://doi.org/10.1101/100784

Daquino, M., Peroni, S., & Shotton, D. (2018). The OpenCitations data model. *Figshare*. https://doi.org/10.6084/m9.figshare.3443876.v7

Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., … Zumstein, P. (2020). The OpenCitations data model. In *International Semantic Web Conference* (pp. 447–463). https://doi.org/10.1007/978-3-030-62466-8_28

DataCite Metadata Working Group. (2016). *DataCite metadata schema for the publication and citation of research data*. Version 4.0. https://doi.org/10.5438/0012

Davidson, S. B., Buneman, P., Deutch, D., Milo, T., & Silvello, G. (2017). Data citation: A computational challenge. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (pp. 1–4). New York: ACM Press. https://doi.org/10.1145/3034786.3056123, PubMed: 29051698

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB Journal*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF, PubMed: 17884971

Färber, M. (2019). The Microsoft Academic Knowledge Graph: A linked data source with 8 billion triples of scholarly data. In *ISWC (2)* (Vol. 11779, pp. 113–129). Springer. https://doi.org/10.1007/978-3-030-30796-7_8

Fava, I. (2020). *OpenAIRE research graph: Connecting open science – consultation phase*. https://www.openaire.eu/openaire-research-graph-open-for-comments, retrieved September 2020.

Fenner, M. (2020). *Powering the PID graph: Announcing the DataCite GraphQL API*. https://doi.org/10.5438/yfck-mv39, retrieved September 2020.

Fenner, M., & Aryani, A. (2019). *Introducing the PID graph*. FREYA Blog. https://doi.org/10.5438/jwvf-8a66, retrieved September 2020.

Force, M., Robinson, N., Matthews, M., Auld, D., & Boletta, M. (2016). Research data in journals and repositories in the Web of Science: Developments and recommendations. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, *12*(1), 27–30.

FORCE-11. (2014). *Data Citation Synthesis Group: Joint declaration of data citation principles* (M. Martone, Ed.). FORCE11, San Diego, CA, USA.

Freeman, G., Ding, Y., & Milojevic, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503. https://doi.org/10.1002/asi.22850

Fricke, S. (2018). Semantic Scholar. *Journal of the Medical Library Association: JMLA*, *106*(1), 145. https://doi.org/10.5195/JMLA.2018.280

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90, PubMed: 16391221

Gilbert, G. N., & Woolgar, S. (1974). Essay review: The quantitative study of science: An examination of the literature. *Science Studies*, *4*(3), 279–294. https://doi.org/10.1177/030631277400400305

Harzing, A.-W. K., & Van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(1), 61–73. https://doi.org/10.3354/esep00076

Justice, C. O., Vermote, E., Townshend, J. R., Defries, R., Roy, D. P., … Barnsley, M. J. (1998). The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, *36*(4), 1228–1249. https://doi.org/10.1109/36.701075

Katz, D. S., Bouquin, D., Hong, N. P. C, Hausman, J., Jones, C., … Zhang, Q. (2019). Software citation implementation challenges. *arXiv, arXiv:1905.08674*. https://arxiv.org/abs/1905.08674

Katz, D. S., Niemeyer, K. E., Smith, A. M., Anderson, W. L., Boettiger, C., … Rios, F. (2016). Software vs. data in the context of citation. *PeerJ Preprints* (4), e2630v1. https://doi.org/10.7287/peerj.preprints.2630v1

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2019). GORC: A large contextual citation graph of academic papers. *arXiv, arXiv:1911.02782*. https://arxiv.org/abs/1911.02782v1

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., … Principe, P. (2019). The OpenAIRE research graph data model (version 1.3). *Zenodo*. https://doi.org/10.5281/zenodo.2643199

Nature Physics Editorial. (2016). A statement about data. *Nature Physics*, *12*(10), 889. https://doi.org/10.1038/nphys3923

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374, PubMed: 26113702

Open Polytechnic. (2020). What's the difference between references and in-text citations? https://www.openpolytechnic.ac.nz/current-students/study-tips-and-techniques/apa-referencing-and-avoiding-plagiarism/the-difference-between-references-and-citations/ (retrieved March 2020).

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri*, *46*(3), 149–158. https://doi.org/10.1515/libr.1996.46.3.149

Peroni, S., & Shotton, D. (2020). Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428–444. https://doi.org/10.1162/qss_a_00023

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4, PubMed: 27122647

Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., ... Tymvios, F. S. (2010). Cost733cat–a database of weather and circulation type classifications. *Physics and Chemistry of the Earth, Parts A/B/C*, *35*(9–12), 360–373. https://doi.org/10.1016/j.pce.2009.12.010

Price, D. J. de Solla. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510, PubMed: 14325149

Price, G., & Richardson, B. (2008). *MHRA style guide: A handbook for authors, editors, and writers of theses*. MHRA.

Pröll, S., & Rauber, A. (2013). Scalable data citation in dynamic, large databases: Model and reference implementation. In *Proceedings of the 2013 IEEE International Conference on Big Data* (pp. 307–312). https://doi.org/10.1109/BigData.2013.6691588

Rauber, A., Ari, A., van Uytvanck, D., & Pröll, S. (2016). Identification of reproducible subsets for data citation, sharing and re-use. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, *12*(1), 6–15.

Rauber, A., Asmi, A., van Uytvanck, D., & Proell, S. (2015). Data citation of evolving data: Recommendations of the Working Group on Data Citation (WGDC). *Result of the RDA Data Citation WG*, *20*.

Roberts, R. J. (2001). PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences*, *98*(2), 381–382. https://doi.org/10.1073/pnas.98.2.381, PubMed: 11209037

Schubert, C. (2017). *Implementing the RDA data citation recommendations by the Climate Change Centre Austria (CCCA) for a repository of NetCDF files webinar*. https://www.rd-alliance.org/implementing-rda-data-citation-recommendations-climate-change-centre-austria-ccca-repository-netcdf, retrieved December 2020.

Shanableh, A., Al-Ruzouq, R., Gibril, M. B. A, Flesia, C., & Al-Mansoori, S. (2019). Spatiotemporal mapping and monitoring of whiting in the semi-enclosed gulf using moderate resolution imaging spectroradiometer (MODIS) time series images and a generic ensemble tree-based model. *Remote Sensing*, *11*(10), 1193. https://doi.org/10.3390/rs11101193

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics*, *1*(1), S6. https://doi.org/10.1186/2041-1480-1-S1-S6, PubMed: 20626926

Silvello, G. (2018). Theory and practice of data citation. *Journal of the American Society for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917

Sinha, S., Shen, Z., Song, Y., Ma, H., Eide, D., ... Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246). https://doi.org/10.1145/2740908.2742839

Southan, C., Sharman, J. L., Benson, H. E., Faccenda, E., Pawson, A. J., ... Davies, J. A. (2015). The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: Towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Research*, *44*(D1), D1054–D1068. https://doi.org/10.1093/nar/gkv1037, PubMed: 26464438

Starr, J., & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, *17*(1/2). https://doi.org/10.1045/january2011-starr

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998). https://doi.org/10.1145/1401890.1402008

Wan, H., Zhang, Y., Zhang, J., & Tang, J. (2019). AMiner: Search and mining of academic social networks. *Data Intelligence*, *1*(1), 58–76. https://doi.org/10.1162/dint_a_00006

Wang, K., Shen, Z., Huang, C., Wu, C., Eide, D., ... Rogahn, R. (2019). A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data*, *2*, 45. https://doi.org/10.3389/fdata.2019.00045, PubMed: 33693368

Wikipedia. (2021). *Citing Sources*. https://en.wikipedia.org/wiki/Wikipedia:Citing_sources. (retrieved December 2021).

Wilke, C. (2015). *What constitutes a citable scientific work?* https://serialmentor.com/blog/2015/1/2/what-constitutes-a-citable-scientific-work

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18, PubMed: 26978244

Wu, Y., Alawini, A., Davidson, S. B., & Silvello, G. (2018). Data citation: Giving credit where credit is due. *Proceedings of the 2018 International Conference on Management of Data* (pp. 99–114). ACM Press. https://doi.org/10.1145/3183713.3196910

Wu, Y., Alawini, A., Deutch, D., Milo, T., & Davidson, S. (2019). ProvCite: Provenance-based data citation. *Proceedings of the VLDB Endowment*, *12*(7), 738–751. https://doi.org/10.14778/3317315.3317317

Zwölf, C. M., Moreau, N., Ba, Y. A., & Dubernet, M. L. (2019). Implementing in the VAMDC the new paradigms for data citation from the Research Data Alliance. *Data Science Journal*, *18*(1), 4. https://doi.org/10.5334/dsj-2019-004