

Theory and Practice of Data Citation

Gianmaria Silvello
Department of Information Engineering, University of Padua, Via Gradenigo 6/b, Padua, Italy
silvello@dei.unipd.it
tel. +39 049 827 7500

Abstract

Citations are the cornerstone of knowledge propagation and the primary means of assessing the quality of research, as well as directing investments in science. Science is increasingly becoming “data-intensive”, where large volumes of data are collected and analyzed to discover complex patterns through simulations and experiments, and most scientific reference works have been replaced by online curated datasets. Yet, given a dataset, there is no quantitative, consistent and established way of knowing how it has been used over time, who contributed to its curation, what results have been yielded or what value it has.

The development of a theory and practice of data citation is fundamental for considering data as first-class research objects with the same relevance and centrality of traditional scientific products. Many works in recent years have discussed data citation from different viewpoints: illustrating why data citation is needed, defining the principles and outlining recommendations for data citation systems, and providing computational methods for addressing specific issues of data citation.

The current panorama is many-faceted and an overall view that brings together diverse aspects of this topic is still missing. Therefore, this paper aims to describe the lay of the land for data citation, both from the theoretical (the why and what) and the practical (the how) angle.

Introduction

Citations are the cornerstone of knowledge propagation in science, the principal means of assessing the quality of research and directing investments in science as well as one of the pillars of the scholarship architecture.

Currently we are witnessing a crucial change in the way research is conducted and science progresses (Bechhofer *et al.*, 2013). We are rapidly transitioning towards *the fourth paradigm of science* (i.e. data-intensive scientific discovery), where data are as vital to scientific progress as traditional publications are. Data and scholarship are increasingly interwoven and new concepts such as “data scholarship” (Borgman, 2015) and “data-intensive research” (Hey & Trefethen, 2005) are becoming popular and central to the academic and scientific world.

Experimental and observational data and scientific models are born digital, and this move to digital content has to be taken into account by scholarly publications, credit attribution processes and scientific and economic impact measurement. Vast amounts of scientific data – molecular, geospatial, astronomical, pharmacological and more – are now collected and made available in structured, evolving and often distributed databases (Buneman, Davidson & Frew, 2016); relevant examples of widely used and accessible scientific databases are the pharmacological IUPHAR/BPS¹ database, the Eagle-i² biomedical dataset, the DrugBank³ bioinformatics and cheminformatics database, the Reactome⁴

¹ <http://www.guidetopharmacology.org/>

² <https://www.eagle-i.net/>

³ <https://www.drugbank.ca/>

⁴ <http://www.reactome.org/>

pathways database and the VADMC⁵ federated database of atomic and molecular data. Moreover, the scientific community is taking action to promote an *open research culture* (Nosek *et al.*, 2016) and effective means to share, discover and access scientific data that too often still remain “hidden” at their origin or are shared in sub-optimal ways (Ohno-machado *et al.*, 2015). Indeed, as a matter of good scientific practice, scientific databases are recommended to be *FAIR* (Findable, Accessible, Interoperable, and Reusable) (Wilkinson *et al.*, 2016) and accessible from the articles referring to or using it (Cousijn *et al.*, 2017).

Since scientific works and publications increasingly rely on curated databases, traditional references are starting to be placed alongside references to data; hence, scientific publishers (e.g. Elsevier, PLoS, Springer, Nature) have been defining data policies and author guidelines to include data citations in the reference lists (Cousijn *et al.*, 2017; Walton, 2010). At the institutional level, within the Horizon 2020 program, the European Commission has introduced the Open Research Data Pilot (ODP), which aims to improve and maximize the access to and re-use of research data and to increase the credit given to data creators; in the same vein, also the National Science Foundation (NSF) and the National Institutes of Health (NIH) in the United States and the Economic and Social Research Council in the United Kingdom ask researchers applying for grants to provide data management plans describing how the data they use and produce will be shared and referenced (MacKenzie, 2012; Spengler, 2012). There is also evidence that: (i) scientists respond to incentives and that increasing citation would drive software development and data sharing (Niemeyer, Smith & Katz, 2016); (ii) research activity increases when outputs are formally counted (McNaught, 2015); and (iii) direct citations to datasets stimulate more data curation and data sharing than indirect citations through publications (Belter, 2014). Moreover, fundamental aspects of scientific research, such as reproducibility of experiments, the availability and discovery of scientific data and the connection between scientific results with the data providing evidence, have been found to be closely connected with data citation (Honor, Haselgrove, Frazier & Kennedy, 2016).

As a consequence, there is a strong demand (CODATA, 2013; FORCE11, 2014; RDA, 2015) to give databases the same scholarly status of traditional references and to define a shared methodology to cite data.

Nevertheless, none of the largest citation-based systems – i.e. Elsevier Scopus, Thompson Reuters Web of Science, Microsoft Academia and Google Scholar – consistently take into account scientific datasets as targeted objects for use in academic work. The Thompson Reuters (now Clarivate Analytics) Data Citation Index (DCI) (Force, Robinson, Matthews, Auld & Boletta, 2016) is one notable exception that aggregates information about the usage of data. The DCI is an infrastructure that keeps track of data usage in the scientific domain, which indexes a number of scientific data sets, allows for discovering the data sets selected and validated by Thompson Reuters and provides the technical means to connect entire data sets or repositories to scientific papers. On the other hand, DCI is still in its infancy, “has yet to be proven cost-effective or efficient” (Mayernik, Hart, Mauli & Weber, 2016) and presents skewed citation patterns (Robinson-Garcia, Jiménez-Contreras & Torres-Salinas, 2016). Therefore, given a dataset, we still not have any quantitative, consistent and established way of knowing how it has been used over time, what analyses have been conducted on it, what results have been yielded, who contributed to its curation and which data subsets have been relevant or impactful for scientific and economic development (Honor *et al.*, 2016).

This is mainly due to the lack of a “deep and persistent mechanism for citing data” (Ingwersen & Chavan, 2011). Indeed, citation rules and practices have stratified over centuries and are well established for text but less so for data (Borgman, 2015). Data citations and traditional paper-based citations have a common intent, but present some epistemological differences that have not yet been fully analyzed and comprehended. Among other things, these differences impact the very use of data citations in bibliometrics (and more generally in infometrics) as pointed out by

⁵ http://portal.vamdc.org/vamdc_portal/home.seam

Borgman (2016): “[the] leap from citing publications to citing data is a vast one [and the transfer of] bibliographic citation principles to data must be done carefully and selectively”. Indeed, bibliometrics is based on publications in the sense of scholarly communications and data do not necessarily qualify as such (Mayernik, Callaghan, Leigh, Tedds & Worley, 2015). Bibliometrics relies on the distinction between reference and citation where a “reference is the acknowledgement that one document gives to another; a citation is the acknowledgement that one document receives from another” (Narin, 1976). Whereas, in data citation, this distinction is often blurred and reference and citation are treated as identical signs. Moreover, another distinctive feature of bibliometrics are reciprocal citations, which are hardly possible with data citations since data may receive citations but cannot cite something else.

Nevertheless, data citations are often referred to as “formal ways to ground the research findings in a manuscript, upon their supporting evidence, when that evidence consists of externally archived datasets” (Cousijn *et al.*, 2017) and it is argued that citations to data should be placed in the reference list section as published works are (Altman, 2012; Corti, Van den Eynden, Bishop & Woollard, 2014; Park & Wolfram, 2017). Hence, a data citation could be indicated with the triple {R, C, D}, where *R* is the reference paper, *C* is the citation and *D* is the data set – i.e. the object of the citation. In this context, the major practical issues for citing data, regard the definition and the creation of the text snippet to be included in the reference list of *R* and the identification and retrieval of *D*. In addition, we must consider that data are more complex and varied than documents (Castelli, Manghi & Thanos, 2013; Wynholds, Wallis, Borgman, Sands & Traweek, 2012) and they introduce new challenges with respect to traditional publications. Indeed, text publications have a fixed form, do not change over time, are interpretable as independent units, share a common format and representation model and are composed of predetermined, albeit domain-dependent, sets of citable units (e.g. authors, title, pages). By contrast, scientific datasets are structured according to diverse data models – e.g. relational, hierarchical (XML), graph-based (RDF) – and accessed with specific query languages – e.g. SQL, XPath/XQuery, SPARQL. Data citable units range from a single datum to data subsets or aggregations specified on the fly with great variability, and deciding *a priori* what can be cited and what cannot, is not (always) feasible. Data evolve over time, but data citations must be *diachronic*: they must be consistent over time and always lead back to the original cited data. A data reference (or citation snippet) is required to allow the data to be understood and correctly interpreted and it must be composed of the essential information for identifying the cited data as well as contextual information. Such contextual information must be extracted from the given dataset and/or from external sources automatically, because we cannot assume one knows how to access and select additional relevant data and to structure them appropriately (Hourclé, 2012). The role of contextual information is of critical importance if we consider the potential, albeit undesirable, situation where data references may survive to the data themselves.

Data citation is a complex problem having both theoretical/conceptual and practical/technical implications that have been investigated both by information science and computer science. The former discipline has been mainly concerned with the motivations – the “why” of data citation – as well as with the principles of this field and the main features of citation systems expressed in terms of recommendations – the “what” of data citation. The latter discipline has been mainly concerned with the computational problems arising from data citation and the design of systems for automating the citation of data – the “how” of data citation.

The field of data citation can be likened to a broken mirror where each single piece of glass reflects a different facet of the problem. Therefore, it is difficult to get the complete picture of the matter, which is nevertheless fundamental for providing a shared theory of data citation as well as a sound data citation system. This paper aims to describe the lay of the land for data citation both from the theoretical (the why and what) and the practical (the how) angle. Moreover, we provide an overview of the most pressing open problems in data citation as well as some related research directions for information and computer science.

A note on literature search

Literature search was conducted as a four-stage process. The first stage was a journal database search based on key term search in Google Scholar and Scopus. We searched for the exact key phrase "data citation" and we discarded most of the papers regarding citation data analysis which are out of scope for this review. We analyzed articles that cover data citation as a need/service/tool to be employed in a specific domain (e.g. biology or pharmacology) and articles discussing the principles and the technical aspects of data citation. We did not consider articles about data publishing, sharing and re-use if they did not explicitly mention data citation.

In the second stage, we analyzed articles published in discipline-specific venues (both journals and conferences), in particular we considered information science, digital libraries and computer science with a specific focus on information management and access systems. In the third stage, we conducted a general search for the "data citation" key-phrase by employing general-purpose Web search engines. This allowed us to look for relevant blog entries, white papers and publisher web pages. As a final stage, we scanned (recursively) the reference list of the papers selected in the previous stages for additional relevant items.

Why data citation: Motivations

The main motivation advocating the importance of data citation is that “*data in research are as valuable as papers and monographs*” (Ball & Duke, 2012); moreover, data citation is becoming central in science because of the great production of data, new research tools for managing, accessing, analyzing and sharing data and the shift of research policy that has been happening in recent years (Borgman, 2012a). The centrality of data citation is widely recognized across many scientific fields ranging from earth sciences, physics and biology to information studies and social sciences. The need for a standard citation mechanisms to give credit to data collectors and to locate and examine the data used in investigations is a long-standing requirement also in the computer science area, especially for database management (French, Jones & Pfaltz, 1990).

Data citation has many facets and it is needed to satisfy several requirements depending at times on the specific discipline in question. Despite this heterogeneity, we have identified six main motivations for data citations that are shared in many different scientific fields:

- i. *Data attribution*: Scientific data are collected and curated with a great deal of work and in many cases this is done directly by qualified scientists and researchers. Giving credit to the people involved in the creation and curation of data is important for scientific recognition (Tenopir *et al.*, 2011). Data attribution opens up the need to identify the author or the person responsible for data with variable granularity; indeed, in many fields it is not enough to give credit to the person who manages the dataset as a whole, as there may be different groups of authors for different subsets of data within the same dataset. Parsons (2012) and Borgman (2015) relate the concept of attribution with that of accountability for the data since if we can identify who is responsible for the data, we know who gets the merit for the data as well as who can be held accountable for them.
- ii. *Data connection*: Data citation methods are required to connect scientific papers with the underlying data on which they are based. Castelli, Manchi & Thanos (2013) indicate data citation as the “main mechanism enabling the alignment and integration between data and publications in the scientific communication process”. This is also the idea behind executable papers or enhanced publications (Vernooy-Gerritsen, 2009) which link data with textual publications allowing the data to be consulted or downloaded while reading a scientific paper in a digital form as well. Several studies (Aalbersberg, Heeman, Koers & Zudilova-Seinstra, 2012; Attwood *et al.*, 2010; Bardi & Manghi, 2015; Brammer, Crosby, Matthews & Williams, 2011; Jankowski, Scharnhorst, Tatum & Tatum,

2012) have been dedicated to the definition and realization of enhanced publications and they all to some extent require a methodology for citing data (and software/code) (Niemeyer, Smith & Katz, 2016). There is also the emerging idea of using data citation in conjunction with the Linked Data paradigm to create a “claim network evidence” spanning different documents (de Waard, 2012; Silvello, 2015). The idea is to backup claims in scientific papers with the data providing evidence for their validity.

- iii. *Data discovery*: Being able to cite data means being able to identify, reach, access and retrieve a dataset or a subset of a dataset, which are the fundamental operations required to make data discoverable. Furthermore, data citations may act as entry points to hidden data sources which are not indexed by search engines and thus are virtually unreachable to all those scientists who do not have the knowledge and the means to reach and interrogate these data sources.
- iv. *Data sharing*: The possibility of citing data and thus of giving credit to data scientists and institutions is considered a decisive factor for augmenting the willingness of scientists to share data. Indeed, since the technological factor is no longer an impediment for data sharing, the main barrier still standing is the fear scientists have of losing a competitive advantage while “receiving no credit and losing funding or publishing opportunities” (Mooney & Newton, 2012). Data sharing is a condition for enabling data discovery and data re-use as well as one of the main factors for the success of the Linked Data paradigm in the scientific context. To this end, Bechhofer *et al.* (2013) highlighted that Linked Data is a compelling approach for sharing and disseminating scientific data, but in order to be effective for data re-use and discovery it has to be connected to the research methodology and respect the rights and the reputation of the researchers.
- v. *Data impact*: Citing data allows us to define new usage metrics for determining the impact of data; data impact can be interpreted as a way to measure or discover which results have been obtained using the data, how many times they have been used, where they have been used and so on. Note that in the literature, data impact is, frequently, intended as the rate of data usage and not as a quality measure. This aspect is particularly important for public and private funding agencies that often invest in the creation of scientific dataset as well as research infrastructures; data citation is required to estimate the use of data going beyond alternative metrics such as altmetrics (Mayernik, 2016). de Waard (2016) pointed out that “making data citable” is one of the ten main characteristics for unlocking the potential of research data; in particular, being able to cite data is important for increasing research exposure. There are also voices calling for considering data citations for bibliometrics with the purpose of evaluating research performances of institutions and individuals (Costello, 2009; Parr, 2007). Moreover, funding agencies underline the need to cite data for measuring the impact of researchers working with data (Ahalt *et al.*, 2015) in order to “help people in career advancements and make their contribution clear” (Spengler, 2012). On the other hand, the use of citation data (e.g. citation count and citation indexes and databases) for evaluating scholarship and funding distribution also raises ethical and policy issues (Borgman, 2016; Garfield & Welljams-Dorof, 1992; Furner, 2014) when applied both to traditional and data citations. In particular, for data citations, given the current absence of “norms of citing behavior”, we have to be even more careful by considering that they cannot be straightforwardly analogized to paper citations (Stuart, 2017).
- vi. *Reproducibility*: Data citation has a profound impact on the reproducibility of science (Baggerly, 2010), a hot topic in many disciplines such as astronomy (Kurtz, 2012), biology (Bloom, Ganly & Winker, 2014), physics, computer science (Ferro, 2017; Freire, Fuhr & Rauber, 2016) and more. Lately, several authoritative journals have been requesting the sharing of data and the provision of validation methodologies for experiments (e.g. Nature Scientific Data and Nature Physics); these publications and the publishing industry in general see data citation as the means for providing new, reliable and usable means for sharing and referring to scientific data.

	Data Attribution	Data Connection	Data Discovery	Data Sharing	Data Impact	Reproducibility	Application Domain
(Altman & King, 2007)	✓					✓	Information Studies
(Altman, 2012)	✓		✓				Information Studies
(Arend et al., 2016)				✓	✓		Biology
(Ball & Duke, 2015)	✓	✓	✓	✓		✓	Information Studies
(Bardi & Manghi, 2014)	✓	✓	✓	✓		✓	Data Publishing
(Belter, 2014)	✓				✓	✓	Oceanography
(Borgman, 2012)	✓	✓	✓				Information Studies
(Borgman, 2015)	✓	✓	✓	✓			Information Studies
(Bravo et al., 2015)	✓		✓	✓	✓	✓	Biodiversity
(Callaghan et al., 2012)	✓			✓			Environmental Science
(Candela et al., 2015)			✓	✓			Data Publishing
(Chavan, 2012)			✓	✓			Biodiversity
(Cook et al., 2016)	✓	✓	✓	✓			Environmental Science
(Costello et al., 2013)				✓	✓		Biodiversity
(de Waard, Cousijn & Aalbersberg, 2015)	✓			✓	✓		Data Publishing
(Dodd, 1979)			✓	✓	✓		Social Sciences
(Duerr et al., 2011)	✓					✓	Earth Science
(Edmunds et al., 2012)	✓			✓		✓	Biology
(Fecher, Friesike & Hebing, 2015)	✓			✓	✓		Data Management
(Herterich et al., 2016)	✓			✓			Physics
(Honor et al., 2016)	✓		✓	✓	✓	✓	Neuroimaging
(Huang, 2015)	✓			✓			Crystallography
(Ingwersen & Chavan, 2011)					✓		Bioinformatics
(Kafkas, Kim, Pi & McEntyre, 2015)		✓					Biomedicine
(King, 1995)						✓	Information Studies
(Klump, Huber & Diepenbroek, 2016)						✓	Earth Science
(Koers, 2015)	✓					✓	Clinical Epidemiology
(Mayernik, 2012)	✓	✓			✓	✓	Information Studies
(Mayernik, 2016)					✓	✓	Information Studies
(Mooney & Newton, 2012)	✓		✓	✓			Data Publishing
(Nature Physics, 2016)	✓						Physics
(Parson, 2012)	✓				✓	✓	Earth Science
(Parsons & Fox, 2013)	✓					✓	Data Publishing
(Peters, 2016)					✓		Information Studies
(Pröll & Rauber, 2013)						✓	Computer Science
(Sieber & Trumbo, 1995)	✓			✓	✓	✓	Information Studies
(Silvello & Ferro, 2016)	✓	✓				✓	Digital Libraries
(Silvello, 2015)	✓	✓				✓	Information Retrieval
(Silvello, 2017)	✓	✓				✓	Computer Science
(Simons, Visser & Searle, 2013)	✓			✓		✓	Digital Libraries
(Starr et al., 2015)		✓			✓		Computer Science
(Thorisson, 2009)	✓	✓					Biotechnology
(White, 1982)	✓		✓		✓	✓	Social Sciences
(Wormack, 2015)			✓	✓			Information Studies
(Zwolf et al., 2016)	✓			✓	✓	✓	Spectroscopy

Table I Summary of the main motivations for data citation in different application domains

In Table I we present a summary of the main motivations for data citation as presented by several papers discussing this topic from different perspectives and application domains. We can see that data attribution and reproducibility are the core topics motivating data citation, followed by data sharing. Data connection has been gaining traction in recent years as it is closely related to reproducibility. Data discovery and data impact are growing and becoming central motivations for data citation, especially within the information studies, data publication and data curation fields.

These motivations explain why data citation is needed for scholarly communications and they constitute a *prior analysis* of why data should/must be cited. From these motivations, though, we cannot infer the reasons why authors decide to cite data and their data citing behavior. To this end, we should reconsider the meaning of data citations, as it may differ from textual citations (Mayernik, 2012) and “we should ask whether the motivations for citing data are the same as for citing research papers” (Robinson-Garcia, Jiménez-Contreras & Torres-Salinas, 2016). In the context of traditional publications, there are different theories of citation (Leydesdorff, 1998) analyzing why we cite (Cozzens, 1981) and when we cite (Garfield, 1996). In particular, it has been pointed out that citations can be seen from a normative perspective as ways to acknowledge intellectual debt (Kaplan, 1965). Alternatively, citations are seen as a means for persuasion, since scientific papers have the role of convincing people of the validity of claims and providing support for them (Gilbert, 1977); or as symbols relating the author’s own private interpretation of the reference with the cited paper itself (Small, 1978). Garfield (1996) listed fifteen major reasons for citation and Cronin (1984) reported several alternative citation categories and typologies defined over the years, to conclude that there is reason to believe that “there are no absolute and exclusive categories which fully describe the relationship of a citing publication to the cited publication” (Lipetz, 1965). Nonetheless, textual analysis of the citing paper could suggest a plausible explanation about why an author cites as s/he does (Cronin, 1981); moreover, given that the citation process is “subjective and inhospitable to standardization” (Cronin, 1981), we have to rely on “ostensible reasons for citation or reasons which can be adduced from the context of the citing work” (Frost, 1979).

For data citation, we can find just a few works specifically considering the context of the citing work (Kafkas, Kim & McEntyre, 2013; Piwowar & Vision, 2013) in order to understand why data are cited. This is also due to the fact that data citations “are useful for the discovery of links, but less so for understanding why the link was made” (Mayernik Philips & Nienhouse, 2015). Mayernik *et al.* (2014) are among the few hypothesizing that data may be cited as “forms of reward to data providers”, as a “quid pro quo for the data provider support” or as a “persuasive statement illustrating how data of high visibility or sufficient quality underlie a scientific result”. Up to now, we can only find isolated and sporadic efforts to understand why data are cited. However, a shared effort for defining a theory of data citing is very much needed, especially before we could “safely” employ bibliometrics for data citations given that “with limited understanding of the norms of data publishing and data citation, too early an emphasis on metrics may do damage to the nascent data ecosystem” (Stuart, 2017).

To this end, we may recall that the development of a citation theory (and the study of citations as a serious academic subject) emerged as a consequence of the successful development of commercial citation indexes (Cronin, 1984). Hence, for data citation, the Thomson-Reuters DCI (and other indexes that may emerge in the next years) can play a role towards the study of a theory of data citation. In particular, the DCI has been used to investigate why data are cited in Genetics and Heredity (Park & Wolfram, 2017) and in the Humanities (Robinson-Garcia, Jiménez-Contreras & Torres-Salinas, 2016). A similar study was carried out, long before the DCI was established, by White (1982) who conducted an analysis of data citations in the context of social sciences. All these studies concluded that the acknowledgement of data usage is the prior motivations for citing a dataset. Other studies (Fear, 2013; Kafkas, Kim & McEntyre, 2013; Weber, Mayernik & Worley, 2014) corroborated this result by suggesting that the number of citations to a dataset correlates with its usage rate. Thus, from these initial and circumscribed analyses, it emerges that a strong

motivation for an author to cite a dataset is to acknowledge the fact that she used it in her research. Having said that, it is complex to define what is meant by “usage” and data citation requires “a more nuanced understanding of data “use” to be effective” (Wynholds, Wallis, Borgman, Sands & Traweek, 2012).

What data citation: Principles and system requirements

The principles of data citation have been extensively described in (CODATA, 2013) and then aggregated, summarized, revised and endorsed by the Joint Declaration of Data Citation Principles (JDDCP) (FORCE11, 2014). These principles can be classified into two main groups: the former states the role of data citation in scholarly and research activities and the latter defines the main guidelines a data citation methodology and system should respect.

In the first group, we find *importance, credit and attribution, evidence, verifiability* and *interoperability*. Importance coincides with the first motivation for data citation we presented above, which is that data should be considered as a first-class citizen in the science panorama and thus be treated as any other scholarly record and scientific paper. This principle is foundational for data citation, because it opens up to the use of data for assessing the work of individuals and institutions, for defining new impact measures and for directing investments in research. This calls for a change of policy in research that invests in public institutions as well as in publishing houses and individual researchers. Furthermore, the principle of importance of data opens up a plethora of theoretical issues that have been only partially discussed in the literature and call for new research; Borgman (2016) points out several theoretical problems of data citation mainly due to the lack of agreement of what constitutes data, amongst which the fact that data are different from traditional publications and require a new set of theoretical premises for bibliometrics considering how citation practices differ between genres of publications and data. Moreover, there is not a shared vision about which data could or should be cited also considering that different application domains call for different data citation practices accounting for references to primary data sources, raw data (e.g., sensor or streaming data), post-processing data, or formatted data as tables and plots.

The JDDCP principles of credit and attribution, evidence and verifiability correspond to some of the motivations for data citation identified above; indeed, citation to data should give credit to people and institutions involved in their creation and curation and scientific claims should be related by the specific data supporting them. Within traditional publications, the accepted protocol for credit and attribution is for a publication to include “in-text citations and a reference section the lists those cited publications” (Honor *et al.*, 2016). Also in the context of data citation, every citation should come with a citation text – i.e. citation snippet or reference text – which describes the object being cited and should contain enough information to give credit to data creators and curators, to understand the motive of the citation and the meaning of the cited object, to locate the object referred to as evidence, and to verify that the located object is semantically identical to the cited one (Altman & King, 2007; Altman, 2012). The citation snippet should be human- and machine-readable; human-readable snippets are required to give immediate sense of a citation even to non-experts without requiring access to the data being referred to by the citation (Kafkas, Kim, Pi & McEntyre, 2015; Silvello & Ferro, 2016; Van de Sompel, 2012). Machine-readable snippets are required for automatic processing of the citation information and they can contain more exhaustive information than human-readable ones; for instance, we may have hundreds of contributors for a given dataset that can be exhaustively reported only in a machine-readable snippet. To this end, a fundamental requirement for citation methods is to automatically create citation snippets in order to ensure consistency of references (Thorisson, 2009). Indeed, recent studies about data citation practices showed that data citation snippets are often underspecified with vital information missing (Mathiak & Boland, 2015) and may present several errors compromising the identification and retrieval of the data being cited (Henderson & Kotz, 2015). *Automatic methods for building citation snippets* are also required because humans cannot remember or know what the

necessary information is, when a snippet is complete, what the citation format is required in a given context and where to gather the relevant information, especially in a big data context where manual exploration of data is unfeasible (Buneman *et al.*, 2014; Crosas *et al.*, 2015; Jagadish, 2015; Minister, 2012; Peters, 2016). The problem of automatic citation snippet creation has been clearly defined by (Buneman *et al.*, 2016) from a database management system perspective: “Given a database *D* and a query *Q*, generate an appropriate citation”; in this context, the term database and query are used in a very general way to refer to any mechanism used to extract the data from a generic source. Borgman (2012; 2016), Buneman *et al.* (2000) and Groth (2012) also outlined the relationship between provenance and data citation. Provenance information should be included in a citation snippet in order to cite the correct version, manipulation and transformation of data. This aspect is of particular relevance to cite meta-analyses conducted on raw or curated data in order to cite and keep track of the original dataset as well as of the transformations and manipulations of the data. The consistency and completeness of citation snippets within a discipline is a concern also from the metadata viewpoint. Indeed, Uhler (2012) underlines the importance of reference consistency since, for instance, (Bolter, 2014) reported cases where there are several hundred variants of citations to the same dataset (i.e. the World Ocean Atlas and World Ocean Database). Borgman (2012b) and NSF (2015) stressed the importance of developing shared and extensible metadata formats for data citation. The last JDDCP principle in the first group is *interoperability*, which requires that data citation methods are flexible enough to operate through different communities and citation practices.

The second group of JDDCP principles defines some of the requisites for data citation methods and systems; it is composed of *unique identification*, *access*, *persistence* and *specificity*. Unique identification indicates that a citation system must provide a methodology for identification that is machine-actionable and widely used within the community of reference. This principle is connected to: (i) *persistence*, which states that unique identifiers should persist even after the lifespan of data they are associated with; and (ii) *specificity*, which states that we should be able to identify specific data (even a data subset) to support a claim and that such data should be connected to provenance information useful for reconstructing their context. Persistence or *fixity* is a key factor for data citation because a data citation system should guarantee that a cited object is always available in the cited form across time. Ensuring the persistence of a database may not be enough (Thorisson, 2009), because to guarantee fixity of data citation we need database versioning systems; indeed, maintaining only the latest version of a dataset is not sufficient given that cited data may change or become unavailable across time without the knowledge of those who are citing it (Borgman, 2012b; Pröll & Rauber, 2013; Rauber, Ari, van Uytvanck & Pröll, 2016).

In a context where citation to big and evolving datasets has to be created and maintained over time, the requirements a data citation method should respect are becoming more demanding and complex than those defined by JDDCP. Unique identification is the most general and widely recognized requirement of a data citation method (Fenner *et al.*, 2016), but the use of persistent identifiers (PID) such as the Digital Object Identifier (DOI) has been highlighted as not being the “magic bullet” able to solve all the identification problems because we need to be able to uniquely identify subsets of data with *variable granularity* (Kafkas, Kim, Pi & McEntyre, 2015; Van de Sompel, 2012; Crosas *et al.*, 2015) since “different fields require different levels of detail to be able to reproduce data” (Mayernik, 2016); in other cases, we may need to identify sets of data aggregated from different sources. The granularity problem cannot be addressed by the use of data papers as proxies for the data to be cited (Candela, Castelli, Manghi & Tani, 2015) since they can only be used for citing a dataset as a whole; the same issue affects widely-used on-line resources, which associate a PID to a dataset and make it available and de-referenceable on-line, but do not provide variable granularity access and identification of subsets of the data.

This problem introduces both theoretical and practical (further discussed below) issues that exist also for textual

citations. Indeed, in the context of traditional publications there are cases where it is customary to cite entire documents (e.g. scientific practice) or to cite single pages or passages (e.g. in the humanities). Nevertheless, in bibliometrics the convention is to aggregate documents by common elements (e.g. author or journal) and the unit of analysis is often the cited document (Borgman, 1990). On the contrary, for data citation there is not an agreed and shared protocol to be followed and being able to identify and cite a subset of data is a central problem that requires further research.

How data citation: Data citation methods

From the analyses conducted so far, we can state that the ideal data citation system should uniquely identify a dataset and subsets of it with different levels of coarseness (*identification*), attribute the ownership and responsibility of the data with variable granularity to the right people/institutions (*attribution*), guarantee the persistence of the data being cited as well as the citations themselves (*fixity*), and automatically create complete and consistent citation snippets (*completeness and consistency*) according to community practices and shared *metadata standards*.

This citation system should be flexible enough to accommodate diverse requirements and citation practices across different disciplines as well as heterogeneous data models such as the relational model, the XML hierarchical model, or the RDF graph model.

From case to case the objects being cited within a given dataset may vary considerably; indeed, a data citation system needs to be able to cite:

- A *single resource* such as a complete dataset or a specific resource within a dataset; e.g. an RDF resource identified by a URI, an XML node identified by a single path from the tree root, a single tuple within a table in a relational database.
- A *subset/selection of resources* such as a selection of resources or portions of resources within a dataset; e.g. a bunch of RDF statements (e.g. a subgraph of an RDF dataset), an XML sub-tree or a set of tuples selected from a relational table.
- An *aggregation of resources* such as the union or join of resources coming from different parts of a dataset or even different datasets; e.g. an RDF graph created by joining together two RDF sub-graphs, a bunch of XML subtrees spanning one or more files, a set of relational tuples obtained by joining tables within the same database or from different databases.

In Table II, we report the main characteristics of the data citation systems and methods proposed in the literature according to the aspects outlined above. We highlight how these methods address the problem of identification, the automatic creation of citation snippets (i.e. automatic or manual), the coarseness of the produced citations (i.e. single resources, subsets or aggregations) and if these methods are general or defined for specific data models.

Table II Summary of the main characteristics of the state-of-the-art citation methods. We take into account the identification method, how the citation snippet is created, how fixity is handled, the level of coarseness of data citation and the data type. N/A indicates that the specific characteristic has not been taken into account by the specific solution.

Reference	Identification method	Citation snippet creation method	Coarseness of the citation	Data model/type
(Alawini <i>et al.</i> , 2017)	PID (URI)	Automatic, view based	Single resource	RDF
(Bandrowsky <i>et al.</i> , 2015)	PID (RRID)	Manual/Template based	Dataset	Any data type
(Buneman & Silvello, 2010)	PID + path to node	Automatic, rule-based	Single resource	XML
(Buneman, Davidson & Frew, 2016)	PID + path to node	Automatic, view-based	Multiple resources	XML

(Crosas, 2011) DataVerse	PID + UNF	Manual	Single resource	Any data type
(Davidson, Deutsch, Milo & Silvello, 2017)	N/A	Automatic, view-based	Multiple resources	Relational DB
(Groth <i>et al.</i> , 2010) Nanopublication	PID	Manual	Single statement (triple)	RDF
(Honor <i>et al.</i> , 2016)	PID	Manual/Landing page	Multiple resources	File system
(Rauber, Ari, van Uytvanck & Pröll, 2016)	Queries used as proxy for data	Manual/Metadata based	Multiple resources	Any data type. CSV and relational DB implementation.
(Silvello, 2015)	Named-graphs	Manual	Multiple resources	RDF
(Silvello, 2017)	PID + path to node	Automatic, machine learning-based	Single resource	XML
(Zwölf, Moreau & Dubernet, 2016)	Queries used as proxy for data	Manual/Landing page (DUI)	Multiple resources	Relational DB

Identification

Most of the solutions proposed in the literature use PID for addressing the identification problem. As we highlighted above, the use of PID is a viable solution for citing a dataset as a whole, but it is not straightforward to use them for citing subsets or aggregations of data. In several cases, identification is provided to a high level of coarseness and identifying data with variable granularity is not possible. For instance, Bandrowsky *et al.* (2015) promotes the use of Research Resource Identifiers (RRID), which are unique persistent identifiers assigned to scientific resources. In particular, they have been used with biological resources such as antibodies, model organisms, and tools (software, databases, services). RRID are meant to identify resources at a high level of granularity and “*for software and databases, we elected to identify just the root entity and not a granular citation of a particular software version or database*”.

Van de Sompel (2012) outlines two options to address the identification problem: to mint a new PID for each segment of data that need to be cited, or to mint a single PID for the dataset and to store the user query selecting or aggregating data.

The first solution has been chosen by DataVerse (Crosas, 2011,) which assigns a DOI to a dataset and generates a UNF (Universal Numerical Fingerprint) for each data segment to be cited within the dataset. A UNF is a PID assigned to specific digital objects or subsets within a dataset and it allows for identifying data with different granularities. This system needs to be extended/revised to handle the citations of big dynamic data in order to be able to cite a subset of data on (i) selected variables and observations for large quantitative data; (ii) time-stamp intervals; and, (iii) spatial dimensions (Crosas *et al.*, 2015) and also to be able to handle queries to the data. The use of DOI rose some concerns about the cost of minting identifiers for large data archives (Henderson & Kotz, 2015).

Buneman & Silvello (2010) and Silvello (2017) discussed the case of XML data, where every single node of an XML file can be a citable unit. In this case, a PID can be assigned to every XML file in a collection, but assigning a PID to every node would be unsustainable if not unfeasible. Buneman & Silvello (2010) proposed to identify a single node within an XML file by associating the PID identifying the XML file with the unique key identifying a specific node within the given XML file – e.g. the path of the node from the root of the file. Silvello (2015) discussed the case of RDF datasets, where we may need to cite a set of RDF statements (e.g. a subgraph of the whole RDF dataset at hand). In this case, every single node in a RDF graph is uniquely identified by a URI (i.e. a PID), but we cannot possibly assign a URI for every possible aggregation of nodes or statement in a dataset. The proposed solution is based on the

use of named graphs (Klyne, Carroll & McBride, 2014), which are a set of connected RDF statements identified by a global PID – i.e. an URI, which is the name of the graph.

Groth *et al.* (2010) proposed the nanopublication model where research data that may need to be cited (or for which provenance and attribution information have to be held for any other motive) are modeled as single statements in a subject-property-object form, as in an RDF triple. The RDF triple is uniquely identified by a URI associated to the triple as in named graphs. The nanopublication model defines the minimum citable unit to be a triple; even though, it is not straightforwardly extensible to identify an aggregation of statements, the nanopublication model is widely used by the bioinformatics research community (Clark, Ciccarese & Goble, 2014; Mina *et al.*, 2015) and there are some studies about its adoption in the humanities (Gradmann, 2014; Golden & Shaw, 2016).

Pröll & Rauber (2013) and Rauber, Ari, van Uytvanck & Pröll (2016) proposed an identification method based on assigning PID to queries, which are used as proxies for the data subset to be cited. The access to data subset is allowed by re-issuing the stored query and a citation is associated with the PID of the query identifying the data. This method is flexible and in principle works for any data format and model; it requires the development of an additional infrastructure for storing the queries and assigning a timestamp to each query on the basis of the last update of the whole database. Moreover, it requires the development of some methods to guarantee query uniqueness – the proposal is to re-write queries to a normalized form and then to compute a checksum to detect identical queries – and stable sorting to ensure that the sorting of the records in the returned dataset is unambiguous and reproducible. This method has been developed in the context of the RDA (RDA, 2015) and it has been implemented by several scientific datasets as detailed in (RDA, 2016).

Parsons (2012) analyzed the problem of identifying subsets of a given dataset in order to make them citable, stating that identifying the query with a PID and citing the query, or assigning a UNF to a data subset to be cited, as in the DataVerse model, are not viable solutions in the context of Earth Science where there are no queries to be used and there is no agreed canonical version for the data. The solution adopted in this field is to use spatial and temporal information (i.e. a structural index as they call it) to identify a specific subset, since in Earth Science these data are always available and are often sufficient to discriminate between two subsets of a given dataset. Hence, Klump, Huber & Diepenbroek (2016) discussed the use of PIDs to identify and cite time series and evolving datasets as in the case of Earth Science resources where “*time or space are dimensions of the data and a subset may be defined by a range or bounding parameters*”. The solution proposed is to enrich a PID with “fragment identifiers” in the form <PID>@<fragment identifier> where the “fragment identifier” can be the version of a resource or the time interval to be considered.

The Virtual Atomic and Molecular Data Centre (VAMDC) (Zwölf, Moreau & Dubernet, 2016) foresees a hybrid solution for citing resource aggregations generated by issuing SQL queries to a federated database putting together many heterogeneous data sources. For identification purposes VAMDC stores the user queries as proxies for the data, but instead of using the query PID for the citations, it creates a file containing the extracted data, information about the databases (e.g. version and contributors) and bibliographical information about the scientific papers related to the data (i.e. a form of provenance). This file is associated to a Digital Unique Identifier (DUI), which points to a landing page containing all the information in the file. The idea is to use the DUI associated to the file generated for the user query as a citation and to use the landing page to get the information usually provided by the citation snippet.

In the context of neuroimaging, (Honor *et al.*, 2016) addressed the problem of citing an aggregation of resources (i.e. images) created on the fly by the users. Here, there are three levels of identification: the image level where each resource is identified by an individual PID (i.e. DOI), the project level where pre-defined sets of images are assigned with a PID and the functional level, where aggregations of resources defined on the fly by users are identified by a

newly minted PID. This solution requires to provide a module for PID deduplication when two different PIDs are assigned to the same functional aggregation of resources.

Metadata schemas

One goal in automating data citations is to structure them in a standardized format so that they can be formatted according to predefined styles and they can be included in bibliography management tools.

There are several metadata formats for data citations (Altman & King, 2007; Green, 2010; Starr & Gastl, 2011) and they all share a common subset of elements (Ball & Duke, 2015): author, publication date, title, edition, version, URI, resource type, publisher, unique number fingerprint (a form of hash of the data), a persistent URL and location. DataCite (DataCite, 2016), a widely-recognized metadata format proposal for citing data, expands this common set of fields by adding others such as subject, contributor, format, size, description, language, rights and funding reference. DataCite is the citation format adopted by the Thompson Reuters DCI. (Parsons, 2012) stated that there are some differences between the DataCite schema and the metadata fields required for citing Earth Science Information Partners (ESIP) data: Authors, Release Date, Title, Version, Editors, Archive and/or Distributor, Locator, Date and Time accessed, Subset used. The main difference regards the specification of the subset of data being used. Starr *et al.* (2015) reports a different minimum set of metadata fields to be used to cite a dataset: Dataset Identifier, Title, Description, Creator, Publisher/Contact, PublicationDate/Year/ReleaseDate, Version, Creator Identifier(s) (optional), License (optional).

The Repository Early Adopters Experts Group (an initiative of FORCE11 and the NIH BioCADDIE⁶ program) (Fenner *et al.*, 2016) stated that the minimum set of required metadata is Dataset Identifier, Title, Creator, Data Repository, Publication Date, Version and Type. (Bravo *et al.*, 2015) proposed an even smaller group of mandatory fields for citing biological resources: Identification (ID and/or DOI), Bioresource name, Acronym (if available); (if applicable) Organization; Number of accesses/Date of last access. By contrast, in the field of biodiversity Chavan (2012) underlined the need to develop two different metadata schemas, one related to published dataset citations and another for query-based citations; he also proposed six alternative data citation styles.

It is evident that one size does not fit all when it comes to metadata formats for data citation. Metadata formats for “traditional” resources in the digital library domain such as the well-known Dublin Core adopted a dynamic solution that could be a viable possibility also in the data citation context. Indeed, the Dublin Core metadata standard is composed of 15 base elements – i.e. the Simple Dublin Core – that can be extended by using the so-called “application profiles” that adapt the standard to the requirements of specific application domains (although an application profile is not only a set of extra metadata fields, it also comprises policies and guidelines defined for a particular application, implementation or object type). In a similar vein, DataVerse (Crosas, 2011) proposed to use the (Altman and King, 2007) metadata schema based on a minimal set of elements which can be extended from case to case.

Metadata schemas should also be employed to enable the machine readability of data citations. The FORCE11 Data Citation Implementation Group in 2014 focused on this aspect by proposing to extend the NISO Z39.96-2012 JATS XML standard for exchanging journal article content in a machine-readable fashion (Mietchen, McEntyre, Beck & Maloney, 2015). This effort led to the definition of the JATS 1.1d2 schema comprising several tags specifically defined for data citation. (see <http://jats4r.org/data-citations>).

⁶ <https://biocaddie.org/>

Completeness and consistency

Completeness and consistency of data citations can be achieved by providing tools able to automatically create citation snippets for the cited data. The creation of citation snippets is often demanded of users; for instance, describing the research identification initiative, Bandrowsky *et al.* (2015) explained that in a federated search context the data creators are asked to insert the correct citation for the data that will be then used as a template provided to the users. Users are asked to build the citation snippet manually using the given template as a guide. Similar methods are employed by many other relevant scientific databases such as the Eagle-i open RDF dataset (Torniai, Bourges-Waldegg & Hoffmann, 2015), a “resource discovery” tool built to facilitate translational science research. All the resources in Eagle-i are citable and, on the one hand, the system provides a “cite me” functionality, which returns a text description about how to cite a given resource and where to retrieve all the relevant data; on the other hand, the citation snippet has to be composed manually by the user. Reactome, a free, open-source, curated and peer reviewed pathway database for bioinformatics funded by the US NIH (Croft, 2013) uses the same manual method to instruct users about how to cite a specific resource. DataVerse assigns metadata to each dataset in a static way; citations to subsets are created starting from the citation of the main dataset and by adding some specific information such as UNF, but there is not a dynamic creation of the citation snippet. In the nanopublication model (Groth *et al.*, 2010) each citable statement is enriched with provenance and attribution annotations that can be used to manually create a citation snippet for the statement.

It has been highlighted that the manual creation of citation snippets is a barrier towards an effective and pervasive data citation practice as well as a source of inconsistencies and fragmentation in the citations (White, 1982; Thorisson, 2009). Indeed, especially for big, complex and evolving datasets, users may not have the necessary knowledge (both of the domain and of the technical aspects) to create complete and consistent snippets.

Indeed, one of the RDA recommendations for citation systems (Raubert, Ari, van Uytvanck & Pröll, 2016) regards “automated citation text generation”, which is required to lower the barrier for citing and sharing the data. The solution sketched out in this context is similar to the one proposed by Bandrowsky *et al.* (2015); during the ingest phase of a dataset, data creators are required to provide descriptive metadata that will be used to create citation snippets.

Recently, the problem of automatically creating citation snippets has been defined as a computational problem that requires new ideas from computer science to be addressed (Buneman, Davidson & Frew, 2016). Existing techniques to automatically build citation snippets consider a single data model at a time and/or a narrow subset of queries.

In this context, the XML model has been investigated thoroughly, principally because it is widely used for data sharing and exchange. Buneman & Silvello (2010) proposed a rule-based system creating citations by using only the information present in the data themselves. Given an XML file, this system requires that the nodes corresponding to citable units be identified (in advance) and tagged with a rule that is then used to generate a citation; the form of the rule is $C \leftarrow P$ where C provides a concrete syntax of a human or machine-readable citation and P is a path augmented with some decorated variables. The purpose of P is to bind the decorated variables in order to use them in C . Once the given XML file has been prepared to be cited (i.e. the rules are in place), the citation of a citable unit within this file is generated by a conjunction of the rules retrieved from the node corresponding to the citable unit up to the root of the XML file. Basically, the system gathers all the rules in the path from the citable unit to the root and each rule contains a specification of the elements to be comprised in the citation that has to be generated. This system allows the automatic generation of both human- and machine-readable citations of single XML nodes.

Buneman, Davidson & Frew (2016) build on and extend this system by defining a view-based citation method for hierarchical data. The idea is to define logical views over an XML dataset, where each view is associated to a citation rule, which if evaluated generates the required citation snippet according to a predefined style. This method allows us to

create citations for general queries and thus for single resources as well as subsets and aggregations of resources. Basically, given a dataset D , a query Q , a set of views V defined over D and a set of citation rules (one for each view), the first step executed by the system is to rewrite Q by using the views in V and thus obtaining a valid rewriting Q' of Q . As an example, we may find that Q' uses two views, say $\{V1, V2\}$; the second step is to take the views used in Q' and to evaluate the rules associated to them. In our example, we would have two rules, one associated to $V1$ and one to $V2$, where each rule generates a valid citation, say $C1$ and $C2$. The third step is to employ a predefined function to combine $C1$ and $C2$ into a single citation C . Davidson, Deutsch, Milo & Silvello (2017) and Davidson, Buneman, Deutsch, Milo & Silvello (2017) further formalized and extended this work for it to work with general queries for relational databases; they also highlight the technical relationships between data citation and provenance. In the same vein by exploiting database views, Alawini, Chen, Davidson, Portilho da Silva & Silvello (2017) proposed a system for citing single RDF resources. They developed a working system tested for the Eagle-i RDF dataset; which creates automatic citation snippets for any RDF graph node, formats them in JSON and in a given human-readable format and maintains fixity thanks to an ad-hoc RDF versioned data store. These approaches are well-defined and could be implemented to work rather efficiently, but their principal drawback is that the rules as well as the views have to be defined by hand and they require the active involvement of experts (data creators and data curators) who know both the dataset and specific query languages and data models.

Silvello (2017) proposed the learning to cite framework, which builds on the approaches described above, but overcomes their main drawback by automating the creation of rules/views. The basic idea is to learn a citation model directly from a given data collection by using a sample set of citation snippets for training purposes and then exploit such a model to build citations on the fly for any citable unit within that collection. This system produces citations for single resources (i.e. single XML nodes) which are not formally exact, but as close as possible to what is considered a “correct citation”. Silvello (2017)’s method needs to be revised in order to face the challenging computational problems emerging when working with more general queries and with relational and graph-based models.

Fixity

Guaranteeing the persistency of the data being cited as well as of the citations themselves is a core aspect of data citation. The main technique addressing the fixity problem has been put forward by the RDA Working Group on Data Citation (Rauber, Ari, van Uytvanck & Pröll, 2016), which proposed a versioning system for relational databases; this versioning system works in concert with the query store and guarantees that the cited data are always retrievable over time even though the dataset has been modified.

However, in order to address fixity for all the different kinds of scientific datasets, we need to address versioning for all the relevant data models adopted to manage, share and access scientific datasets; i.e. at least relational databases, XML and RDF datasets. Versioning systems must work with time queries because data referred by a citation needs to be retrieved at the time the citation was issued. There are some viable solutions for relational databases (Bohlen, Gamper, Jansen & Snodgrass, 2009) that may require improvements from the efficiency viewpoint and an effective solution for XML (Buneman, Khanna, Tajima & Tan, 2004) that needs to be extended to work with time queries, whereas RDF-based versioning requires a major methodological advancement to be usable with time queries (Geerts, Unger, Karvounarakis, Fundulaki & Christophides, 2016) in the data citation context. A first RDF versioning system for RDF datasets is proposed in (Alawini *et al.*, 2017) even though time queries are not explicitly handled.

Citation infrastructures

Data infrastructures storing, managing, providing access and preserving datasets are essential to the development of data citation. Only a little more than five years ago, De Waard (2012) pointed out that: “Overall, commercial publishers are not interested in owning or charging for research data or running repositories. There might be exceptions, but in general this is the case.” King (2011) also underlined the need to cooperate across scholarly fields: “[...] unless we are content to let data sharing work only within disciplinary silos we need to develop solutions that operate, or at least interoperate, across scholarly fields.” Lately the data repository situation has been evolving considerably and there are several viable alternatives providing services to store, manage and access open research data and cite them both at the private and public level and across scholarly fields. Indeed, we see the emergence of numerous general-purpose data repositories, at scales ranging from institutional, to open globally-scoped repositories such as Dataverse, FigShare, Dryad, Mendeley Data, Zenodo, DataHub, DANS, and EUDat (Wilkinson *et al.*, 2016). These infrastructures mint persistent identifiers (DOI), allow for versioning (in some cases) and accept a wide range of file formats (Amorim, Castro, Rocha da Silva & Ribeiro, 2016). Moreover, in most cases, they provide a ready-to-use text snippet (created from manually inserted metadata) for the datasets and links to the papers using the datasets. There are several other public and private services which mint and assign DOI to datasets, but do not store them and thus handle versioning “in-house”; relevant examples are the *DataCite initiative*⁷ and the *Dataverse network*⁸.

All the above-mentioned data infrastructures worked in the direction of allowing the citation of data at fixed levels of coarseness – i.e. dataset, database, single file – and do not allow variable granularity data citations. From this, it follows that citation snippets are statically assigned to the datasets and that there are no mechanisms in place to automatically create citation snippets for data subsets defined on the fly by users.

It emerges from the analysis we have conducted that data citation is not a by-product of data storage and access and cannot be handled just by adopting persistent identifiers or by requiring users to provide describing metadata to be used to create the citation snippets. Data citation requires the implementation of complex and comprehensive solutions that will have quite an impact on current data infrastructures.

Conclusions

Profound changes have been taking place in the way research is done and science progresses. Data has become fundamental for building and interpreting new scientific results and a major player in scientific advancement. Experimental and observational data and scientific models are now “born digital”, and the progressive shift towards the *data-intensive science discovery* paradigm impacts all scientific disciplines. Nevertheless, scientific data are still not considered first-class players in the system of science.

In this article, we investigated the main motivations why data citation is central for the development of science and we analyzed several studies from different scientific fields to understand their motivations and check for common elements and viewpoints. We have concluded that data citation is required to give credit to data creators and data curators; it is a common belief that credit and attribution serve as an incentive to scientists for sharing more and better data leading to the reproducibility of scientific experiments and results. To this end, data citation is central also for the development of executable papers that allow data to be connected with the results presented in a scientific paper; a further extension of this view is the possibility of sustaining and validating scientific claims in a dynamic and automatic way. Data citation has a major impact also for easing the discovery of hidden data sources by providing new access points to them. Public and private institutions investing in dataset creation and curation see data citation as an important means for measuring the impact of data; this is important also for directing investments and research funding. In fact, the diffusion of

⁷ <http://www.datacite.org/>

⁸ <http://dataverse.org/>

pervasive and consistent data citation practices could lead to the definition of new impact measures and methodologies for assessing the scientific production of individuals and research institutions. This aspect has risen several concerned voices asking for a more profound comprehension of the norms regulating the citation of data and analyzing author citation behavior before using data citations for bibliometrics purposes.

Despite its relevance and the attention dedicated to the topic, data citation poses a number of questions that have only been partially addressed by the information and computer science communities:

- (*Identification problem*) What data can be cited? How do we define a data citable unit? How do we identify a single resource, a subset of resources and an aggregation of resources?
- (*Completeness problem*) When data is extracted from a large, complex, evolving database, how do we create an appropriate and informative citation for it? How do we guarantee the consistency of citation texts?
- (*Fixity problem*) How do we guarantee that cited data will be accessible in their cited form?

We analysed these questions in detail and provided an overview of the existing approaches to address these issues. We have seen that significant effort has been made to pursue the identification of data, mostly relying on PIDs. Other approaches use queries as proxies for identifying data, but general agreement about what is the best solution for this issue has not yet been reached.

The completeness problem is even more open for discussion, since most solutions still rely on different forms of manual (or computer-aided) creation of citation snippets, thus often leading to incomplete and inconsistent citations. Automatic methods are required to improve consistency and ease the citation process; the existing automatic methods are defined for specific data models or query types and need more study and work to be generalized in order to be used for any kind of scientific datasets and general queries.

Guaranteeing the persistency of both the cited data and citations is a fundamental task any citation system has to accomplish; up until now, few solutions provided a fully-fledged solution able to deal with fixity, and more research on versioning systems and time queries is required to deal efficiently with this issue.

There are other issues connected to data citation to which not enough attention has been dedicated in the past and that could open up new research directions in the field. One of these is *citation identity* or the ability to discriminate between two citations referring to different data or different versions of the same data and between two different citations referring to the same data. Another challenge is *citation containment* or the possibility to determine if a citation refers to a superset or a subset of the data cited by another citation. We have also seen that only initial efforts have been conducted in the direction of defining a *theory of data citing* and that the understanding of the motivations for citing data and why data is cited, is a challenge of fundamental importance for the definition of reliable data citation metrics and data citation indexes.

Furthermore, easy to use citation tools have to be developed both from the data creators/curators/administrators and final users' viewpoint. The former requires tools for specifying what data can be cited (e.g. which are the citable units) and how data have to be cited (e.g. for defining rules, views, training sets, etc.). The latter require tools for actually citing the data and using the produced citations in their work. To this end, we may need to develop tools able to create citation for data selected through a graphical user interface and not necessarily through well-formed and formal queries.

References

Aalbersberg, I. J., Heeman, F., Koers, H., & Zudilova-Seinstra, E. (2012). Elsevier's article of the Future enhancing the user experience and integrating data through applications. *Insights*, 25(1), 33–43.

- Ahalt, S., Carsey, T., Couch, A., Hooper, R., Ibanez, L., Idaszak, R., ... & Robinson, E. (2015). NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution. *National Science Foundation Workshop Report on Software and Data Citation*, National Science Foundation and Sloan Foundation.
- Alawini, A., Chen, L., Davidson, S., Portilho Da Silva, N. & Silvello, G. (2017). Automating data citation: The Eagle-i experience. In *Proc. of the 17th ACM/IEEE Joint Conference on Digital Libraries*. ACM Press, New York, USA.
- Altman, M. (2012). Data Citation in the DataVerse Network. In *National Academy of Sciences Board on Research Data and Information, Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*, pages 99–106. National Academies Press: Washington DC.
- Altman, M. & Crosas, M. (2013). The Evolution of Data Citation: From Principles to Implementation. *IAssist Quarterly*, 37(1–4):62–70.
- Altman, M. & King, G. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, 13(3/4).
- Altman, M., Borgman, C. L., Crosas, M., & Martone, M. (2015). An Introduction to the Joint Principles for Data Citation. *Bulletin of the Association for Information Science and Technology*, 41(3):43–45.
- Amorim, R. C., Castro, J. A., Rocha da Silva, J., and Ribeiro, C. (2016). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, pages 1–12.
- Arend, D., Junker, A., Scholz, U., Schuler, U., Wylie, J., & Lange, M. (2016). PGP Repository: A Plant Phenomics and Genomics Data Publication Infrastructure. *The Journal of Biological Databases and Curation*, pages 1–10.
- Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., and Thorne, D. (2010). Utopia Documents: Linking Scholarly Literature with Research Data. *Bioinformatics* 26.18 (2010): i568–i574.
- Baggerly, K. (2010). Disclose all Data in Publications. *Nature*, (467):401.
- Ball, A. & Duke, M. (2015). How to Cite Datasets and Link to Publications. Technical Report, Edinburgh: Digital Curation Centre.
- Bandrowski, A., Brush, M., Grethe, J. S., Haendel, M. A., Kennedy, D. N., Hill, S., ... & Vasilevsky, N. (2015). The Resource Identification Initiative: A Cultural Shift in Publishing. *F1000Research*, 4(134).
- Bardi, A. & Manghi, P. (2014). Enhanced Publications: Data Models and Information Systems. *LIBER Quarterly*, 23(4):240–273.
- Bechhofer, S., Buchan, I. E., Roure, D. D., Missier, P., Ainsworth, J. D., Bhagat, J., ... , & Goble, C. A. (2013). Why linked data is not enough for scientists. *Future Generation Comp. Syst.*, 29(2):599–611.
- Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLoS ONE*, 9(3):e92590.
- Bloom, T., Ganly, E., & Winker, M. (2014). Data Access for the Open Access Literature: PLOS's Data Policy. *PLoS Biol*, 12(2).
- Bohlen, M. H., Gamper, J., Jensen, C. S., & Snodgrass, R. T. (2009). SQL-Based Temporal Query Languages. In *Encyclopedia of Database Systems*, pages 2762–2768.
- Borgman, C. L. (1990). Editor's introduction. In *Scholarly Communication and Bibliometrics*, pages 10–27. Newbury Park, CA: Sage.

- Borgman, C. L. (2012a). The Conundrum of Sharing Research Data. *JASIST*, 63(6):1059–1078.
- Borgman, C. L. (2012b). Why are the Attribution and Citation of Scientific Data Important? In *National Academy of Sciences Board on Research Data and Information, Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*, pages 1–8. National Academies Press: Washington DC.
- Borgman, C. L. (2015). *Big Data, Little Data, No Data*. MIT Press.
- Borgman, C. L. (2016). Data Citation as a Bibliometric Oxymoron. In *Theories of Informetrics and Scholarly Communication*, Sugimoto, C. eds. pages 93–116. De Gruyter Mouton.
- Brammer, G. R., Crosby, R. W., Matthews, S. J., & Williams, T. L. (2011). Paper Mâché: Creating Dynamic Reproducible Science. *Procedia Computer Science*, 4:658–667.
- Bravo, E., Calzolari, A., De Castro, P. M. L., Napolitani, F., Rossi, A. M., & Cambon-Thomsen, A. (2015). Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). *BMC Medicine*, 13(33):1–12.
- Buneman, P. & Silvello, G. (2010). A Rule-Based Citation System for Structured and Evolving Datasets. *IEEE Data Eng. Bull.*, 33(3):33–41.
- Buneman, P., Davidson, S. B., & Frew, J. (2016). Why Data Citation is a Computational Problem. *Comm. of the ACM (CACM)*, 59(9):50–57.
- Buneman, P., Khanna, S., Tajima, K., & Tan, W.-C. (2004). Archiving Scientific Data. *ACM Trans. on Database Systems (TODS)*, 29(1):2–42.
- Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., ... & Wright, D. (2012). Making Data a First-Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, 7(1):107–113.
- Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data Journals: A Survey. *JASIST*, 66(9):1747–1762.
- Castelli, D., Manghi, P., & Thanos, C. (2013). A vision towards scientific communication infrastructures. *International Journal on Digital Libraries*, 13(3):155–169.
- Chavan, V. (2012). Data Citation Mechanism and Service for Scientific Data: Defining a Framework for Biodiversity Data Publishers. In *National Academy of Sciences Board on Research Data and Information, Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*, pages 113–115. National Academies Press: Washington DC.
- Clark, T., Ciccarese, P., & Goble, C. (2014). Micropublications: A Semantic Model for Claims, Evidence, Arguments and Annotations in Biomedical Communications. *Journal of Biomedical Semantics*, 5(1).
- 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*, 12:1–67.
- Cook, R. B., Vannan, S. K. S., McMurry, B. F., Wright, D. M., Wei, Y., ... & Kidder, J. (2016). Implementation of Data Citations and Persistent Identifiers at the ORNL DAAC. *Ecological Informatics*, 33:10–16.
- Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice*. SAGE, Los Angeles, CA.
- Costello, M. J. (2009). Motivating Online Publication of Data. *BioScience*, 59(5):418–427.
- Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q., & Bourne, P. E. (2013). Biodiversity Data Should be Published, Cited, and Peer Reviewed. *Trends in Ecology & Evolution*, 28(8):454–461.

- Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Murphy, ... & Clark, T. (2017). A data citation roadmap for scientific publishers. *bioRxiv*. doi: <https://doi.org/10.1101/100784>
- Cozzens, S. (1981). Taking the Measure of Science: A Review of Citation Theories. *International Society for the Sociology of Knowledge Newsletter*, 7(1/2):16–21.
- Croft, D. (2013) Building Models Using Reactome Pathways as Templates. *Methods Mol. Biol.* 1021:273-83.
- Cronin, B. (1981). The need for a Theory of Citing. *Journal of Documentation*, 37(1):16–24.
- Cronin, B. (1984). The citation process. The role and significance of citations in scientific communication. London: Taylor Graham.
- Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA:FORCE11, 2014
- Data Citation Working Group Meeting, September 2016, Denver, USA. <https://www.rd-alliance.org/group/data-citation-wg/post/rda-wgdc-session-p8-denver>
- DataCite (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data, Version 4.0. Technical Report, *DataCite Metadata Working Group*.
- Davidson, S. B., Buneman, P., Deutsch, D., Milo, T., & Silvello, G. (2017). Data Citation: a Computational Challenge. In Proc. of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2017), pages 1-4.
- Davidson, S. B., Deutsch, D., Milo, T., & Silvello, G. (2017). A Model for Fine-Grained Data Citation. In Proc. of the 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017), 7 pages.
- de Waard, A. (2012). Linking Data to Publications: Towards the Execution of Papers. In *National Academy of Sciences Board on Research Data and Information, Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*, pages 157–159. National Academies Press: Washington DC.
- de Waard, A. (2016). Research Data Management at Elsevier: Supporting Networks of Data and Workflows. *Information Services & Use*, 36:49–55.
- de Waard, A., Cousijn, H. & Aalbersberg, I. J. (2015). 10 aspects of highly effective research data. Elsevier. <https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data>
- Dodd, S. A. (1979). Bibliographic References for Numeric Social Science Data Files: Suggested Guidelines. *JASIST*, 30(2):77–82.
- Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B. R., Lenhardt, W. C., Glassy, ... & Slaughter, P. (2011). On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations. *Earth Science Informatics*, 4(3):139–160.
- Edmunds, S. C., Pollard, T. J., Hole, B., & Basford, A. T. (2012). Adventures in Data Citation: Sorghum Genome Data Exemplifies the New Gold Standard. *BMC Research Notes*, 5:223.
- Fear, K. (2013). Measuring and anticipating the impact of data reuse. Ann Arbor: University of Michigan.
- Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *PLoS ONE*, 10(2).
- Fenner, M., Crosas, M., Grethe, J., Kennedy, D., Hermjakob, H., Rocca-Serra, P., ... & Clark, T. (2016). A data citation roadmap for scholarly data repositories. *bioRxiv*. DOI: 10.1101/097196
- Ferro, N. (2017). Reproducibility Challenges in Information Retrieval Evaluation. *ACM JDIQ*, 8(2):1-4 (2017).

- 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.
- Freire, J., Fuhr, N. & Rauber, A. (Eds.). 2016. Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. *Dagstuhl Reports*, 6, 1.
- French, J. C., Jones, A. K., & Pfaltz, J. L. (1990). Summary of the Final Report of the NSF Workshop on Scientific Database Management. *SIGMOD Record*, 19(4):32–40.
- Frost, C. (1979). The use of citations in literary research: preliminary classification of citation functions. *Library Quarterly*, 49(4):399–414.
- Furner, J. (2014). The Ethics of Evaluative Bibliometrics. In Cronin, B. and Sugimoto C. R. eds. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, pages 85–107. MIT Press, Cambridge, MA.
- Garfield, E. (1996). When to Cite. *Library Quarterly*, 66(4):449– 458.
- Garfield, E. & Welljams-Dorof, A. (1992). Citation Data: Their Use as Quantitative Indicators for Science and Technology Evaluation and Policy-Making. *Science and Public Policy*, 19(5):321–327.
- Geerts, F., Unger, T., Karvounarakis, G., Fundulaki, I., & Christophides, V. (2016). Algebraic Structures for Capturing the Provenance of SPARQL Queries. *Journal of the ACM*, 63(1):1-63.
- Gilbert, G. N. (1977). Referencing as Persuasion. *Social Studies of Science*, 7:113–122.
- Golden, P. & Shaw, R. (2016). Nanopublication Beyond the Sciences: The Period Gazetteer. *PeerJ Computer Science*, 2:e44.
- Gradmann, S. (2014). From Containers to Content to Context: The Changing Role of Libraries in eScience and eScholarship. *Journal of Documentation*, 70(2):241–260.
- Green, T. (2010). We Need Publishing Standards for Datasets and Data Tables. Technical report, OECD Publishing.
- Helbig, K., Hausstein, B., & Toepfer, R. (2015). Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences. *Journal of Librarianship and Scholarly Communication*, 3(2), eP1220.
- Henderson, T. & Kotz, D. (2015). Data Citation Practices in the CRAWDAD Wireless Network Data Archive. *D-Lib Magazine*, 21(1/2).
- Herterich, P. & Dallmeier-Tiessen, S. (2016). Data Citation Services in the High-Energy Physics Community. *D-Lib Magazine*, 22(1/2).
- Hey, T. & Trefethen, A. E. (2005). Cyberinfrastructure for eScience. *Science*, 308(5723):817–821.
- Honor, L. B., Haselgrove, C., Frazier, J. A., & Kennedy, D. N. (2016). Data Citation in Neuroimaging: Proposed Best Practices for Data Identification and Attribution. *Frontiers in Neuroinformatics*, 10(34):1– 12.
- Hourclé, J. A. (2012). Advancing the practice of data citation: A to-do list. *Bulletin of the American Society for Information Science and Technology*, 38(5):20–22.
- Ingwersen, P. & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): An Incentive for Publishing Primary Biodiversity Data Through Global Information Infrastructure. *BMC Bioinformatics*, 12(S-15):S3.
- Jagadish, H. V. (2015). Big Data and Science: Myths and Reality. *Big Data Research*, 2(2):49–52. Visions on Big Data.
- Jankowski, N. W., Scharnhorst, A., Tatum, C. & Tatum, Z. (2012). Enhancing Scholarly Publications: Developing Hybrid Monographs in the Humanities and Social Sciences. *Scholarly and Research Communication*, 4(1): 010138.

- Kafkas, S., Kim, J.-H., & McEntyre, J. R. (2013). Database Citation in Full Text Biomedical Articles. *PLOS ONE*, 8(5):1–9.
- Kafkas, S., Kim, J.-H., Pi, X., & McEntyre, J. R. (2015). Database Citation in Supplementary Data Linked to Europe PubMed Central Full Text Biomedical Articles. *Journal of Biomedical Semantics*, 6(1).
- King, G. (1995). Replication, Replication. *PS: Political Science and Politics*, 28:444–452.
- King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, 331(6018):719–721
- Klump, J., Huber, R., & Diepenbroek, M. (2016). DOI for Geoscience Data – How Early Practices Shape Present Perceptions. *Earth Science Informatics*, 9:123-136.
- Klyne, G., Carroll, J. J. & McBride, B. (2014). “RDF 1.1 Concepts and Abstract Syntax”. *W3C Recommendation*, 25-Feb-2014. <http://www.w3.org/TR/rdf11-concepts/>
- Koers, H. (2015). How Do We Make it Easy and Rewarding for Researchers to Share Their Data? A Publisher’s Perspective. *Journal of Clinical Epidemiology*, 70:261–263.
- Kurtz, M. J. (2012). Linking, Finding, and Citing Data in Astronomy. In National Research Council. In *For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press, 2012. doi:10.17226/13564.
- Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, 43(1):5–25.
- Lipetz, B.-A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16(2):81–90.
- MacKenzie, S. (2012). Institutional Perspective on Credit Systems for Research Data. In National Research Council. In *For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press, 2012. doi:10.17226/13564.
- Mathiak, B. & Boland, K. (2015). Challenges in Matching Dataset Citation Strings to Datasets in Social Science. *D-Lib Magazine*, 21(1/2).
- Mayernik, M. S. (2012). Data Citation Initiatives and Issues. *Bulletin of the Association for Information Science and Technology*, 38(5):23–28.
- Mayernik, M. S., Callaghan, S., Leigh, R., Tedds, J., and Worley, S. (2015). Peer Review of Datasets: When, Why, and How. *Bulletin of the American Meteorological Society*, 96(2):191–201.
- Mayernik, M. S., Hart, D. L., Mauli, D. L., & Weber, N. M. (2016). Assessing and Tracing the Outcomes and Impact of Research Infrastructures. *JASIST*, in press:1–19.
- Mayernik, M. S., Philips, J., & Nienhouse, E. (2015). Linking Publications and Data: Challenges, Trends, and Opportunities. *D-Lib Magazine*, 22(5/6).
- McNaught, K. (2015). The Changing Publication Practices in Academia: Inherent Uses and Issues in Open Access and Online Publishing and the Rise of Fraudulent Publications. *The Journal of Electronic Publishing*, 18(3).
- Mietchen, D., McEntyre, J., Beck, J. & Maloney, C. (2015). Adapting JATS to support data citation. In *Proc. of the Tag Suite Conference (JATS-Con) 2015* [Internet]. National Center for Biotechnology Information (USA).
- Mina, E., Thompson, M., Kaliyaperumal, R., Zhao, J., van der Horst, E., Tatum, Z., ... & Roos, M. (2015). Nanopublications for Exposing Experimental Data in the Life-Sciences: a Huntington’s Disease Case Study. *J. Biomedical Semantics*, 6:5.

- Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B., den Dunnen, J. T., van Ommen, G., ... & Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4):281–283.
- Mooney, H. & Newton, M. P. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication*, 1(1).
- Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington, DC: National Science Foundation.
- Nature Physics Editorial (2016). A Statement About Data. *Nature Physics*, 12(10):889.
- Niemeyer, K. E., Smith, A. M., & Katz, D. S. (2016). The Challenge and Promise of Software Citation for Credit, Identification, Discovery, and Reuse. *Journal Data and Information Quality*, 7(4):16:1–16:5.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425.
- Ohno-machado, L., Alter, G., Fore, I., Martone, M., Sansone, S.A. & Xu, H. (2015). bioCADDIE white paper - Data Discovery Index. Figshare. <http://dx.doi.org/10.6084/m9.figshare.1362572>
- Park, H. & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, 111(1):443–461.
- Parr, C. S. (2007). Open Sourcing Ecological Data. *BioScience*, 57(4):309–310.
- Parsons, M. (2012). How to Cite an Earth Science Dataset? In *National Academy of Sciences Board on Research Data and Information, Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*, pages 117–124. National Academies Press: Washington DC.
- 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.
- Piwowar, H. A. & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1:e175.
- Pröll, S. & Rauber, A. (2013). Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In *Proc. of the 2013 IEEE International Conference on Big Data*, pages 307–312. IEEE Computer Society.
- Rauber, A., Ari, A., van Uytvanck, D., & Pröll, S. (2016). Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bul. of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, 12(1):6–15.
- Research Data Alliance Working Group on Data Citation. Making Data Citable: Case Statement. <https://rd-alliance.org/group/data-citation-wg/case-statement/wg-data-citation-making-data-citable-case-statement.html> (October 2016)
- Robinson-Garcia, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing Data Citation Practices According to the Data Citation Index. *JASIST*, 67(12): 2964-2975.
- Sieber, J. E. & Trumbo, B. E. (1995). (Not) Giving Credit Where Credit is Due: Citation of Data Sets. *Science and Engineering Ethics*, 1(1):11–20.
- Silvello, G. (2015). A Methodology for Citing Linked Open Data Subsets. *D-Lib Magazine*, 21(1/2).
- Silvello, G. (2017). Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. *JASIST*, in print:1–28. DOI: 10.1002/asi.23774

- Silvello, G. & Ferro, N. (2016). "Data Citation is Coming". Introduction to the Special Issue on Data Citation. *Bulletin of IEEE Technical Committee on Digital Libraries*, Special Issue on Data Citation, 12(1):1–5.
- Simons, N., Visser, K., & Searle, S. (2013). Growing Institutional Support for Data Citation: Results of a Partnership Between Griffith University and the Australian National Data Service. *D-Lib Magazine*, 19(11/12).
- Small, H. G. (1978). Cited Documents as Concept Symbols. *Social Studies of Science*, 8:327–340.
- Starr, J. & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, 17(1/2).
- Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R., Duerr, R., ... & Clark, T. (2015). Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications. *PeerJ Computer Science*, 1.
- Stuart, D. (2017). Data bibliometrics: metrics before norms. *Online Information Review*, 41(3).
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011) Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE* 6(6): e21101.
- Thorisson, G. A. (2009). Accreditation and Attribution in Data Sharing. *Nature Biotechnology*, 27:984–985.
- Torniai, C., Bourges-Waldegg, D., & Hoffmann, S. (2015). Eagle-i: Biomedical Research Resource Datasets. *Semantic Web*, 6(2):139–146.
- Vernooy-Gerritsen, M. (2009). Enhanced Publications: Linking Publications and Research Data in Digital Repositories. Amsterdam University Press.
- Walton, D. (2010). Data Citation - Moving to New Norms. *Antarctic Science*, 22(4), 333.
- Weber, N., Mayernik, M. & Worley, S. (2014). A citation analysis of "data publications" in Earth systems science. In *Proc. of the 9th International Digital Curation Conference*. Digital Curation Centre.
- White, H. D. (1982). Citation analysis of data file use. *Library Trends*, 31(3):467–477.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018 EP.
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., ... & Woolsey, J. (2006). DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration. *Nucleic Acids Res. (Database Issue)*, 34:D668–D672.
- Wormack, R. P. (2015). Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics. *PLoS ONE*, 10(12).
- Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices. In *Proc. 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012)*, pages 19–22. ACM Press, New York, USA.
- Zwölf, C. M., Moreau, N., & Dubernet, M.-L. (2016). New Model for Datasets Citation and Extraction Reproducibility in VADMC. *Journal of Molecular Spectroscopy*, 327:122–137.