The Road Towards Reproducibility in Science: The Case of Data Citation

Nicola Ferro and Gianmaria Silvello(✉)

Department of Information Engineering, University of Padua, Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. Data citation has a profound impact on the reproducibility of science, a hot topic in many disciplines such as as astronomy, biology, physics, computer science 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 to provide new, reliable and usable means for sharing and referring to scientific data. In this paper, we present the state of the art of data citation and we discuss open issues and research directions with a specific focus on reproducibility. Furthermore, we investigate reproducibility issues by using experimental evaluation in *Information Retrieval (IR)* as a test case. (This paper is a revised and extended version of [33,35,57]).

1 Motivations

Data citation plays a central role for providing better transparency and reproducibility in science [16], a challenge taken up by several fields such as Biomedical Research [2], Public Health Research [27] and Biology [18]. Computer Science is also particularly active in reproducibility, as witnessed by the recent *Association for Computing Machinery (ACM)* policy on result and artifact review and badging[1]. For instance, the Database community started an effort called "SIGMOD reproducibility" [38] "to assist in building a culture where sharing results, code, and scripts of database research"[2]. Since 2015, the *European Conference in IR (ECIR)* [34,41], allocated a whole paper track on reproducibility and in 2015 the RIGOR workshop at SIGIR was dedicated to this topic [12]. Moreover, in 2016 the "Reproducibility of Data-Oriented Experiments in e-Science" seminar was held in Dagstuhl (Germany) [3] bringing together researchers from different fields of computer science with the goal "to come to a common ground across disciplines, leverage best-of-breed approaches, and provide a unifying vision on reproducibility" [33,35].

In recent years, the nature of research and scientific publishing has been rapidly evolving and progressively relying on data to sustain claims and provide

[1] https://www.acm.org/publications/policies/artifact-review-badging.
[2] http://db-reproducibility.seas.harvard.edu/.

C. Grana and L. Baraldi (Eds.): IRCDL 2017, CCIS 733, pp. 20–31, 2017.
DOI: 10.1007/978-3-319-68130-6_2

experimental evidence for scientific breakthroughs [44]. The preservation, management, access, discovery and retrieval of research data are topics of utmost importance as witnessed by the great deal of attention they are receiving from the scientific and publishing communities [21]. Along with the pervasiveness and availability of research data, we are witnessing the growing importance of citing these data. Indeed, data citation is required to make results of research fully available to others, provide suitable means to connect publications with the data they rely upon [59], give credit to data creators, curators and publishers [20], and enabling others to better build on previous results and to ask new questions about data [19].

In the traditional context of printed material, the practice of citation has been evolving and adapting across the centuries [21] reaching a stable and reliable state; nevertheless, traditional citation methods and practices cannot be easily applied for citing data. Indeed, citing data poses new significant challenges, such as:

1. the use of heterogeneous data models and formats – e.g., flat data, relational databases, *Comma Separated Value (CSV)*, *eXtensible Markup Language (XML)*, *Resource Description Framework (RDF)* – requiring different methods to manage, retrieve and access the data;
2. the transience of data calling for versioning and archiving methods and systems;
3. the necessity to cite data at different levels of coarseness – e.g., if we consider a relational database, then we may need to cite a specific attribute, a tuple, a tuple sets, a table or the database as a whole – requiring methods to individuate, select and reference specific subsets of data;
4. the necessity to automatically generate citations to data because a 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.

As a consequence, traditional practices need to evolve and adapt in order to provide effective and usable methods for citing data.

IR represents a challenging field for data citation as well as for reproducibility. In particular, experimental evaluation in IR represents an effective testbed for new ideas and methods for reproducing experiments and citing data. Indeed, reproducing IR experiments is extremely challenging and there are three main different areas that are of major concern for reproducibility: experiments (or system runs), experimental collections, and meta-evaluation studies. Experiments can be seen as the output of a retrieval system – e.g., a ranking list of documents – given a corpus of documents and an information need; to reproduce an experiment we need to get access to the corpus or sub-corpus and to the information needs used in the experiments as well as the software and the methods employed.

Meta-evaluation studies are even more complex since they often involve manipulation of the data used in the actual analysis; this, among other things, requires to keep track of the provenance of the data and to include provenance information also in the citations to data.

This paper is organized as follows: Sect. 2 briefly presents the state of the art of research in data citation and some open issues and research lines focusing also on provenance which is particularly important for reproducibility in IR. Section 3 describes the main issues concerning reproducibility in IR evaluation with a specific focus on the role of data citation in this context. Finally, Sect. 4 draws some final remarks.

2 Data Citation: Open Issues and Research Directions

Data citation is a complex problem that can be tackled from many perspectives and involves different areas of information and computer science. Overall, data citation has been studied from two main angles: the scholar publishing viewpoint and the infrastructural and computational one.

The former has been investigating the core principles for data citation and the conditions that any data citation solution should meet [1,37]; the need to connect scientific publications and the underlying data [17]; the role of data journals [26]; the definition of metrics based on data citations [45]; and the measurement of datasets impact [11,53].

The latter has been focusing on the infrastructures and systems required to handle the evolution of data such as archiving systems for XML [23], RDF [49] and databases [51]; the use of persistent identifiers [47,58]; the definition frameworks and ontologies to publish data [40]; and, the creation of repositories to store and provide access to data [4,25].

As described in [22], from the computational perspective the problem of data citation can be formulated as follows: "Given a dataset D and a query Q, generate an appropriate citation C". Several of the existing approaches to address this problem allow us to reference datasets as a single unit having textual data serving as metadata source, but as pointed out by [51] most data citations "can often not be generated automatically and they are often not machine interpretable". Furthermore, most data citation approaches do not provide ways to cite datasets with variable granularity.

Until now, the problem of how to cite a dataset at different levels of coarseness, to automatically generate citations and to create human- and machine-readable citations has been tackled only by a few working systems. In [51] an approach relying on persistent and timestamped queries to cite relational databases has been proposed; this method has been implemented to work with CSV files [52]. On the other hand, this system does not provide a suitable means to automatically generate human- and machine-readable citations. In [24] a rule-based citation system that creates machine- and human-readable citations by using only the information present in the data has been proposed for citing XML. This system has been extended into a methodology that works with database views provided that the data to be cited can be represented as a hierarchy [22]; this work has been further extended for general queries over relational databases in [28–30]. [55] proposed a methodology for citing XML data based on machine learning techniques, which allows us to create citations with variable granularity learning from examples and reducing the human effort to a minimum. In [54] a methodology based on named meta-graphs to cite RDF sub-graphs has been proposed; this solution for RDF graphs targets the variable granularity problem and proposes an approach to create human-readable and machine-actionable data citations even though the actual elements composing a citation are not automatically selected. In the context of RDF citation, [40] proposed the nano-publication model where a single statement RDF triple is made citable in its own right; the idea is to enrich a statement via annotations adding context information such as time, authority and provenance. The statement becomes a publication itself carrying all the information to be understood, validated and re-used. This solution is centered around a single statement and the possibility of enriching it.

A great deal of attention has been dedicated to the use of persistent identifiers [9,47,58] such as Digital Object Identifiers (DOI), Persistent Uniform Resource Locator (PURL) and the Archival Resource Key (ARK). Normally, these solutions propose to associate a persistent identifier with a citable dataset and to create a related set of metadata (e.g., author, version, URL) to be used to cite the dataset. Persistent identifiers are foundational for data citation, but they represent just one part of the solution since they do not allow us to create citations with variable granularity, unless we create a unique identifier for each single datum in a dataset, which in most of the cases may be unfeasible. As a consequence, the use of persistent identifiers as well as their study and evaluation is mainly related to the publication of research data in order to provide a handle for subsequent citation purposes rather than a data citation solution itself.

Data citation is a compound and complex problem and a "one size fits all" system to address it does not exist, yet. Indeed, as we have discussed above, flat data, relational databases, XML and RDF datasets are intrinsically different one from the other, present heterogeneous structures and functions and, as a consequence, require specific solutions for addressing data citation problems. Furthermore, different communities present specific peculiarities, practices and policies that must be considered when a citation to data has to be provided.

As a consequence, within the context of data citation, there are several open issues and research directions we can take into account:

Automatic Generation of Citations. Most of the solutions addressing this problem work for XML data because they exploit its hierarchical structure to gather the relevant (meta)data to be used in a citation. On the other hand, there is no ready to use solution for non-hierarchical datasets as it may be a relational database or a RDF dataset. A further problem is to automatically create citations for data with no structure at all.

Citation Identity. This problem refers to the necessity of uniquely identifying a citation to data and of being able 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.

Citation Containment. We need to define some methods to check if a citation refers to a superset or a subset of the data cited by another citation; somehow, we may need to define hierarchies of citations in order to identify the relationships they have one with the other.

Citation identity and containment have a direct impact on the definition of data citation indexes that can be used to assess the overall impact of a dataset and to quantify the impact and the contribution of a data creator/curator as we now do with bibliometrical indicators based on traditional citations.

Versioning. One of the main differences between traditional citations and data citations is that data may not be fixed, but it may evolve through time; indeed, new data may be added to a dataset, some changes may occur, some mistakes may be fixed or new information may be added. All these changes in a dataset reflect on the citations to data that have been produced. Indeed, a citation needs to ensure that the data a citation uses is identical to that cited [8]. Several archiving and versioning systems have been proposed especially for relational databases and XML data, but they have not been incorporated with data citation solutions, yet.

Provenance. Provenance information plays a central role because we may need to reconstruct the chain of ownership of a data object or the chain of modifications that occurred to it in order to produce a reliable citation. New solutions have to be provided to integrate data citation with currently employed systems controlling and managing the data workflow.

A further challenge is represented by streaming data which may not been always available or which keep constantly changing through time.

Groups of Citations and the Empty Set. Most of the solutions we discussed above are oriented to the citation of a single datum such as a single node, a set of connected nodes in a hierarchy or a set of connected statements in a RDF dataset. On the other hand, we may need to provide a suitable citation for hundreds or thousands of independent data; let us imagine a query to a relational database returning a hundred of possibly unrelated tuples, how do we provide a single citation for this result set?

Vice versa, a related problem is how to define a suitable citation for the empty set. In other terms, how do we create a citation for a query that returns no results?

Supporting Scientific Claims. Scientific claims are often based on evidence gathered from data. They could be related to a single datum or to multiple data coming from the same source or from different sources. Data citation can be used to support such claims and to provide a means to verify their reliability. Actionable papers aim at connecting the presented results with the data from which they have been derived; in this case, we are foreseeing an evolution of such papers, where every single component of a scientific statement can be related to a piece of evidence (data) supporting it and some sort of automatic inference can be carried out.

3 Reproducibility Open Issues: The Example of IR Evaluation

Performances of IR systems are determined not only by their efficiency but also and most importantly by their *effectiveness*, i.e. their ability to retrieve and better rank relevant information resources while at the same time suppressing the retrieval of not relevant ones. Due to the many sources of uncertainty, as for example vague user information needs, unstructured information sources, or subjective notion of relevance, *experimental evaluation* is the only mean to assess the performances of IR systems from the effectiveness point of view. Experimental evaluation relies on the Cranfield paradigm which makes use of *experimental collections*, consisting of documents, sampled from a real domain of interest; topics, representing real user information needs in that domain; and, relevance judgements, determining which documents are relevant to which topics [43].

Reproducing IR experiments is extremely challenging, even when they are very well-documented [14,15,36]. There are three main different areas that are of major concern for reproducibility: system runs, experimental collections, and meta-evaluation studies.

The most common concern for reproducibility are *system runs*, i.e. the outputs of the execution of an IR system, since they are what typically researchers and developers want to compare their new ideas against. Even if you use the same datasets and even if you rely on shared open source software, there are often many hidden parameters and tunings which hamper the reproducibility of algorithms and techniques. The situation is even more challenging when you also rely on user-interaction data. Approaches like Evaluation-as-a-Service [42], based on open interfaces and virtual machines as in *The Incredible Research Assistant (TIRA)*[3] [50], or Open Runs [61], i.e. system runs backed by a software repository that captures the code to recreate the run, are now starting to explore how to face these issues.

Experimental collections are the core of evaluation and they are used for many years, often for purposes different from those that led to their creation. Nevertheless, they are not yet a primary focus for reproducibility, even if they should be, given their central role in experimentation. Indeed, it is important to understand their limitations and their generalizability as well as to reproduce the process that led to their creation. This is not always trivial since, for example, documents may be ephemeral data such as tweets [10], topics may be

[3] http://www.tira.io/.

sampled from real system logs, relevance judgments are made by (disagreeing) humans [60] and, more and more often, using crowdsourcing [7].

Even if IR has a long tradition in ensuring that the due scientific rigor is guaranteed in producing experimental data, it has not a similar tradition in managing and taking care of such valuable data [5,32]. This represents a serious obstacle to facing the above mentioned challenges. For example, there is a lack of commonly agreed formats for modeling and describing the experimental data as well as almost no metadata (descriptive, administrative, copyright, etc.) for annotating and enriching them. The semantics of the data themselves is often not explicit and it is demanded to the scripts typically used for processing them, which are often not well documented, rely on rigid assumptions on the data format or even on side effects in processing the data. Finally, IR lacks a commonly agreed mechanism for citing and linking data to the papers describing them [57].

As there are many different terms relating to various kinds of reproducibility [31], the *Platform, Research goal, Implementation, Method, Actor, and Data (PRIMAD)* (pronounce "primed") model, proposed by [3,35], can act as a framework to distinguish the major components describing an experiment in computer science (and related fields):

Research Goal characterizes the purpose of a study;
Method is the specific approach proposed or considered by the researcher;
Implementation refers to the actual implementation of the method, usually in some programming language;
Platform describes the underlying hard- and software like the operating system and the computer used;
Data consists of two parts, namely the input data and the specific parameters chosen to carry out the method;
Actor refers to the experimenter.

As an example, consider a student performing a retrieval experiment. The research goal is to achieve a high retrieval quality, and as method chosen is the BM25 formula. Experiments use the LEMUR system as implementation, under the operating system Ubuntu 15.10 on a Dell xyz server. The GOV2 collection serves as input data, and a specific setting of the BM25 parameters is chosen. The actor is the student performing the runs.

When another researcher now tries to reproduce the experiment described above, she will change one or more of the components. In case she tries to rerun the experiment without changing anything else[4], then we have another actor, that is, A is changed to A′, the actor is "primed". If successful, this experiment would demonstrate that the original researcher has supplied enough information to ensure reproducibility. In case the results of the experiment are the same, then the original findings have been successfully reproduced and thus confirmed.

Now let us look at changes of the other components:

$\mathbf{R} \rightarrow \mathbf{R}'$: When the research goal is changed, then we *re-purpose* some of the components of the experiment for another research question (for example,

[4] Actually, this would be difficult to achieve.

performing interactive retrieval experiments). So method and implementation usually are also changed, and other components as well.

M → **M′** : Most of the research in the field of IR deals with the investigation of alternative methods (retrieval models, formulas). This implies also a new implementation I′, which often runs on a different platform. However, for performing comparisons, the (input) data should be the same.

I → **I′** : Here a researcher uses a different implementation, say Terrier instead of Lemur, or does their own reimplementation.

P → **P′** : In most cases, independent researchers do not have access to the platform used in the original experiment. Even different versions of system libraries, or external resources such as dictionaries, might have subtle effects on the outcome of experiments.

D → **D′** : Rerunning an experiment with different parameters might be useful for testing the robustness of a method. Applying the implementation to different input data (for example, test collections) aims at investigating the generality of the method.

For ensuring reproducibility, there is the need to be able to share as many PRIMAD components as possible. Research goal and method are what we currently share via publications in conference proceedings or journals (although details of the method are often missing). Sharing implementations are possible via making it open source and uploading it on Web sites focusing on this task (for example, Github). Platforms can be shared by means of virtual machines or dockers, or by "evaluation as a service". For the input data, there are a number of standard test collections which are generally available. When researchers use their own test collection, however, reproducibility can only be ensured in case this collection is shared with the community, ideally via a trustworthy repository.

However, also in this case, we note a lack of attention to data citation. Indeed, the PRIMAD model allows us to have a common framework to describe what has changed from one experiment to another and to clearly define the kind of reproduciblity we are achieving (or not). Nevertheless, all these changes modeled by the framework should be backed by a proper data citation mechanism that allows us to track them and to reference back to them.

There have been early examples of systems to manage IR experimental data, such as EvaluatIR [13] and *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*[5] [4,6], but they have not been designed with reproducibility and/or data citation as goals. More recently, steps forward more fine grained models and systems have been proposed, as for example LOD-DIRECT[6] [56] which uses semantic Web and *Linked Open Data (LOD)* technologies to model IR evaluation data and make them linkable, or nanopublications for IR evaluation [48].

All these examples provide bit and pieces which may exploited or further developed to support reproducibility and data citation in IR evaluation but a more comprehensive and holistic approach would be needed. Indeed, a full

[5] http://direct.dei.unipd.it/.

[6] http://lod-direct.dei.unipd.it/.

fledged abstract conceptual framework for describing IR experiments with reproducibility and data citation in mind, e.g. an evolution of PRIMAD, should be paired with semantic models clearly formalizing it, e.g. a further development of LOD-DIRECT, and proper systems should be developed to implement and operationalize it, e.g. starting from DIRECT and TIRA.

4 Final Remarks

In this paper we discussed reproducibility in science by highlighting why it is important and the main issues that need to be addressed. Data citation plays a central role for enabling reproducibility, but despite its importance and the attention dedicated by the information and computer science communities, there still are several open issues that need to be tackled in order to have a general and usable data citation system. Hence, we outlined the main open issues and research direction in data citation. Moreover, we presented the concrete use case of IR experimental evaluation highlighting the state of the art, the open problems and where data citation can play a central role for enabling effective reproducibility in IR.

References

1. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data, vol. 12. CODATA-ICSTI Task Group on Data Citation Standards and Practices, September 2013
2. Reproducibility and reliability of biomedical research: improving research practice. Technical report, The Academy of Medical Science (2015)
3. Freire, J., Fuhr, N., Rauber, A. (eds.): Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. Dagstuhl Reports, vol. 6, no. 1. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Germany (2016)
4. Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., Silvello, G.: DIRECTions: design and specification of an IR evaluation infrastructure. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) CLEF 2012. LNCS, vol. 7488, pp. 88–99. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33247-0_11
5. Agosti, M., Di Nunzio, G.M., Ferro, N.: The importance of scientific data curation for evaluation campaigns. In: Thanos, C., Borri, F., Candela, L. (eds.) DELOS 2007. LNCS, vol. 4877, pp. 157–166. Springer, Heidelberg (2007). doi:10.1007/978-3-540-77088-6_15
6. Agosti, M., Ferro, N.: Towards an evaluation infrastructure for DL performance evaluation. In: Tsakonas, G., Papatheodorou, C. (eds.) Evaluation of Digital Libraries: An Insight into Useful Applications and Methods, pp. 93–120. Chandos Publishing, Oxford (2009)
7. Alonso, O., Mizzaro, S.: Using crowdsourcing for TREC relevance assessment. Inf. Process. Manage. **48**(6), 1053–1066 (2012)
8. Altman, M., Crosas, M.: The evolution of data citation: from principles to implementation. IAssist Q. **37**(1–4), 62–70 (2013)
9. Altman, M., King, G.: A proposed standard for the scholarly citation of quantitative data. IASSIST (2006). http://www.iassistdata.org/conferences/archive/2006

10. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M.: Overview of RepLab 2012: evaluating online reputation management systems. In: Forner, P., Karlgren, J., Womser-Hacker, C., Ferro, N. (eds.) CLEF 2012 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613–0073 (2012). http://ceur-ws.org/Vol-1178/
11. Angelini, M., Ferro, N., Larsen, B., Müller, H., Santucci, G., Silvello, G., Tsikrika, T.: Measuring and analyzing the scholarly impact of experimental evaluation initiatives. Procedia Comput. Sci. **38**, 133–137 (2014)
12. Arguello, J., Crane, M., Diaz, F., Lin, J., Trotman, A.: Report on the SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). SIGIR Forum **49**(2), 107–116 (2015)
13. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: EvaluatIR: an online tool for evaluating and comparing IR systems. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proceedings of 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), USA, p. 833. ACM, New York (2009)
14. Badan, A., Benvegnù, L., Biasetton, M., Bonato, G., Brighente, A., Cenzato, A., Ceron, P., Cogato, G., Marchesin, S., Minetto, A., Pellegrina, L., Purpura, A., Simionato, R., Soleti, N., Tessarotto, M., Tonon, A., Vendramin, F., Ferro, N.: Towards open-source shared implementations of keyword-based access systems to relational data. In: Ferro, N., Guerra, F., Ives, Z., Silvello, G., Theobald, M. (eds.) Proceedings of 1st International Workshop on Keyword-Based Access and Ranking at Scale (KARS 2017) - Proceedings of the Workshops of the EDBT/ICDT 2017 Joint Conference (EDBT/ICDT 2017). CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613–0073 (2017). http://ceur-ws.org/Vol-1810/
15. Badan, A., Benvegnù, L., Biasetton, M., Bonato, G., Brighente, A., Marchesin, S., Minetto, A., Pellegrina, L., Purpura, A., Simionato, R., Soleti, N., Tessarotto, M., Tonon, A., Ferro, N.: Keyword-based access to relational data: to reproduce, or to not reproduce? In: Greco et al. [39]
16. Baggerly, K.: Disclose all data in publications. Nature **467**, 401 (2010)
17. Bardi, A., Manghi, P.: A framework supporting the shift from traditional digital publications to enhanced publications. D-Lib Magaz. **21**(1/2) (2015). http://dx.doi.org/10.1045/january2015-bardi
18. Bloom, T., Ganly, E., Winker, M.: Data access for the open access literature: PLOS's data policy. PLoS Biol. **12**(2), e1001797 (2014)
19. Borgman, C.L.: The conundrum of sharing research data. JASIST **63**(6), 1059–1078 (2012). http://dx.doi.org/10.1002/asi.22634
20. Borgman, C.L.: Why are the attribution and citation of scientific data important? In: Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences (eds.) Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop, pp. 1–8. National Academies Press, Washington DC (2012)
21. Borgman, C.L.: Big Data, Little Data, No Data. MIT Press, Cambridge (2015)
22. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Commun. ACM (CACM) **59**(9), 50–57 (2016)
23. Buneman, P., Khanna, S., Tajima, K., Tan, W.C.: Archiving scientific data. ACM Trans. Database Syst. (TODS) **29**(1), 2–42 (2004)
24. Buneman, P., Silvello, G.: A rule-based citation system for structured and evolving datasets. IEEE Data Eng. Bull. **33**(3), 33–41 (2010). http://sites.computer.org/debull/A10sept/buneman.pdf

25. Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., Schindler, U.: On bridging data centers and publishers: the data-literature interlinking service. In: Garoufallou, E., Hartley, R.J., Gaitanou, P. (eds.) MTSR 2015. CCIS, vol. 544, pp. 324–335. Springer, Cham (2015). doi:10.1007/978-3-319-24129-6_28
26. Candela, L., Castelli, D., Manghi, P., Tani, A.: Data journals: a survey. J. Assoc. Inf. Sci. Technol. **66**(9), 1747–1762 (2015). http://dx.doi.org/10.1002/asi.23358
27. Carr, D., Littler, K.: Sharing research data to improve public health: a funder perspective. J. Empir. Res. Hum. Res. Ethics **10**(3), 314–316 (2015)
28. Davidson, S.B., Deutsch, D., Milo, T., Silvello, G.: A model for fine-grained data citation. In: Greco et al. [39]
29. Davidson, S.B., Deutsch, D., Tova, M., Silvello, G.: A model for fine-grained data citation. In: 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017) (2017)
30. Davidson, S.B., Buneman, P., Deutch, D., Milo, T., Silvello, G.: Data citation: a computational challenge. In: Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2017), USA, pp. 1–4 (2017). http://doi.acm.org/10.1145/3034786.3056123
31. De Roure, D.: The future of scholarly communications. Insights **27**(3), 233–238 (2014)
32. Dussin, M., Ferro, N.: Managing the knowledge creation process of large-scale evaluation campaigns. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 63–74. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04346-8_8
33. Ferro, N.: Reproducibility challenges in information retrieval evaluation. ACM J. Data Inf. Qual. (JDIQ) **8**(2), 8:1–8:4 (2017)
34. Ferro, N., et al. (eds.): ECIR 2016. LNCS, vol. 9626. Springer, Cham (2016)
35. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing reproducibility in IR: findings from the dagstuhl seminar on "reproducibility of data-oriented experiments in e-science". SIGIR Forum **50**(1), 68–82 (2016)
36. Ferro, N., Silvello, G.: Rank-biased precision reloaded: reproducibility and generalization. In: Hanbury et al. [41], pp. 768–780
37. FORCE-11: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. FORCE11, San Diego, CA, USA (2014)
38. Freire, J., Bonnet, P., Shasha, D.: Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, pp. 593–596 (2012). http://doi.acm.org/10.1145/2213836.2213908
39. Greco, S., Saccà, D., Flesca, S., Masciari, E. (eds.): Proceedings of 25th Italian Symposium on Advanced Database Systems (SEBD 2017) (2017)
40. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use **30**(1–2), 51–56 (2010)
41. Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.): ECIR 2015. LNCS, vol. 9022. Springer, Cham (2015). doi:10.1007/978-3-319-16354-3
42. Hanbury, A., Müller, H., Balog, K., Brodt, T., Cormack, G.V., Eggel, I., Gollub, T., Hopfgartner, F., Kalpathy-Cramer, J., Kando, N., Krithara, A., Lin, J., Mercer, S., Potthast, M.: Evaluation-as-a-service: overview and outlook. CoRR abs/1512.07454, December 2015
43. Harman, D.K.: Information Retrieval Evaluation. Morgan & Claypool Publishers, San Rafael (2011)

44. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2009)
45. Huang, Y.H., Rose, P.W., Hsu, C.N.: Citing a data repository: a case study of the protein data bank. PLoS ONE **10**(8), e0136631 (2015)
46. Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.): CLEF 2014. LNCS, vol. 8685. Springer, Cham (2014). doi:10.1007/978-3-319-11382-1
47. Klump, J., Huber, R., Diepenbroek, M.: DOI for geoscience data - how early practices shape present perceptions. Earth Sci. Inform. 1–14 (2015). http://dx.doi.org/10.1007/s12145-015-0231-5
48. Lipani, A., Piroi, F., Andersson, L., Hanbury, A.: An Information Retrieval Ontology for Information Retrieval Nanopublications. In: Kanoulas et al. [46], pp. 44–49
49. Papavasileiou, V., Flouris, G., Fundulaki, I., Kotzinos, D., Christophides, V.: High-level change detection in RDF(S) KBs. ACM Trans. Database Syst. **38**(1), 1 (2013)
50. Potthast, M., Gollub, T., Rangel Pardo, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the reproducibility of PAN's shared tasks: plagiarism detection, author identification, and author profiling. In: Kanoulas et al. [46], pp. 268–299
51. Pröll, S., Rauber, A.: Scalable data citation in dynamic, large databases: model and reference implementation. In: Hu, X., Young, T.L., Raghavan, V., Wah, B.W., Baeza-Yates, R., Fox, G., Shahabi, C., Smith, M., Yang, Q., Ghani, R., Fan, W., Lempel, R., Nambiar, R. (eds.) Proceedings of the 2013 IEEE International Conference on Big Data, pp. 307–312. IEEE (2013)
52. Pröll, S., Rauber, A.: Asking the right questions - query-based data citation to precisely identify subsets of data. ERCIM News **100** (2015)
53. Robinson-Garcia, N., Jiménez-Contreras, E., Torres-Salinas, D.: Analyzing data citation practices according to the data citation index. J. Am. Soc. Inf. Sci. Technol. (JASIST) **67**, 2964–2975 (2015)
54. Silvello, G.: A methodology for citing linked open data subsets. D-Lib Magaz. **21**(1/2) (2015). http://dx.doi.org/10.1045/january2015-silvello
55. Silvello, G.: Learning to cite framework: how to automatically construct citations for hierarchical data. J. Am. Soc. Inf. Sci. Technol. (JASIST), 1–28 (2017)
56. Silvello, G., Bordea, G., Ferro, N., Buitelaar, P., Bogers, T.: Semantic representation and enrichment of information retrieval experimental data. Int. J. Digit. Libr. (IJDL) **18**(2), 145–172 (2017)
57. Silvello, G., Ferro, N.: Data citation is coming. Introduction to the special issue on data citation. Bullet. IEEE Tech. Committee Digit. Libr. (IEEE-TCDL) **12**(1), 1–5 (2016)
58. Simons, N.: Implementing DOIs for research data. D-Lib Magaz. **18**(5/6) (2012). http://dx.doi.org/10.1045/may2012-simons
59. Vernooy-Gerritsen, M.: Enhanced Publications: Linking Publications and Research Data in Digital Repositories. Amsterdam University Press, Amsterdam (2009)
60. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Inf. Process. Manage. **36**(5), 697–716 (2000)
61. Voorhees, E.M., Rajput, S., Soboroff, I.: Promoting repeatability through open runs. In: Yilmaz, E., Clarke, C.L.A. (eds.) Proceedings of 7th International Workshop on Evaluating Information Access (EVIA 2016), pp. 17–20. National Institute of Informatics, Tokyo, Japan (2016)