"Data Citation is Coming" Introduction to the Special Issue on Data Citation

Gianmaria Silvello
Department of Information Engineering
University of Padua
Padua, Italy
gianmaria.silvello@unipd.it

Nicola Ferro
Department of Information Engineering
University of Padua
Padua, Italy
nicola.ferro@unipd.it

ABSTRACT

This is the introduction to the special issue on data citation of the Bulletin of IEEE Technical Committee on Digital Libraries. In this introduction we state the "lay of the land" of research on data citation, we discuss some open issues and possible research directions and present the main contributions provided by the papers of the special issue.

1. MOTIVATIONS

The practice of citation is foundational for scientific advancement and the propagation of knowledge and it is one of the basic means on which scholarship and scientific publishing rely [26]. In recent years, the nature of research and scientific publishing has been rapidly evolving and progressively relying on data to sustain claims and provide experimental evidence for scientific breakthroughs [35]. 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 [14, 15, 25].

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 [22,46], give credit to data creators, curators and publishers [13], and enabling others to better build on previous results and to ask new questions about data [12]. Furthermore, data citation plays a central role for providing better transparency and reproducibility in science [9], a challenge taken up by several fields such as Biomedical Research [1], Public Health Research [24] and Biology [11]. Computer Science is also particularly active in reproducibility. For instance, the Database community started an effort called "SIGMOD reproducibility" [32] "to assist in building a culture where sharing results, code, and scripts of database research"[1]. The Information Retrieval community in 2011 organized the DESIRE workshop [4] on data infrastructures for supporting IR evaluation with a specific focus on reproducibility; since 2015, the *European Conference in IR (ECIR)* [29, 34], allocated a whole paper track on reproducibility and in 2015 the RIGOR workshop at SIGIR was dedicated to this topic [8]. Moreover, in 2016 the "Reproducibility of Data-Oriented Experiments in e-Science" seminar was held in Dagstuhl (Germany) [2] 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" [30].

In the traditional context of printed material, the practice of citation has been evolving and adapting across the centuries [14] reaching a stable and reliable state; nevertheless, traditional citation methods and practices cannot be easily applied for citing data [17]. 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 Values (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 we cannot assume that the people citing the data understand the complexity of a dataset, know how data should be cited in a specific context, and select relevant information to form a complete and correct citation.

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

The rest of the paper is organized as follows: Section 2 briefly presents the state of the art of research in data citation. Section 3 describes some open issues and research lines. Lastly, Section 5 acknowledges the people who have worked and contributed to this special issue.

[1] http://db-reproducibility.seas.harvard.edu/

2. STATE OF THE ART

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 [15, 28, 38]; the need to connect scientific publications and the underlying data [10]; the role of data journals [23]; the definition of metrics based on data citations [25,36]; and the measurement of datasets impact [7,43].

The latter has been focusing on the infrastructures and systems required to handle the evolution of data such as archiving systems for XML [19], RDF [39] and databases [40]; the use of persistent identifiers [37, 45]; the definition frameworks and ontologies to publish data [33]; and, the creation of repositories to store and provide access to data [3, 21].

As described in [18], 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 [40] 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 [40] 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 [41]. On the other hand, this system does not provide a suitable means to automatically generate human- and machine-readable citations. In [16,20] 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 [18]. In [44] 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, [33] 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 [6, 37, 45] 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.

3. OPEN ISSUES AND RESEARCH DIRECTIONS

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 [5]. 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.

4. IN THIS SPECIAL ISSUE

This special issue presents three papers covering a wide spectrum of aspects concerning data citation and coming from both academic and industrial realities.

The paper entitled "Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use" [42] presents the 14 recommendations elaborated by the RDA Working Group on Dynamic Data Citation (WGDC) on how to adapt a data source for providing identifiable subsets of data for the long-term. This work is the outcome of a collective effort aimed at providing the community with a list of guidelines for making the data citable; it describes how to prepare data repositories and storage systems to data citation. Furthermore, it digs on the role of persistent identifiers by extending their use from pointers to static data to means for identifying queries and dynamic data with variable granularity.

The paper entitled "Well-Stratified Linked Data for Well-Behaved Data Citation" [27] focuses on the citation of RDF datasets on a theoretical level. It proposes a formalization of data citation for linked data and shows how to solve the problem of authorship attribution, subset and version identification. Moreover, this work highlights the relationships between data citation and data provenance when it comes to identify data subsets. This work represents an insight about how the problem of data citation is seen within the Semantic Web community.

The paper entitled "Research Data in Journals and Repositories in the Web of Science: Developments and Recommendations" [31] provides insights about the functioning of Web of Science^{TM} and the Data Citation Index^{SM}. In particular, it is described how the papers included in Web of Science^{TM} as well as the repositories considered in the Data Citation Index^{SM} are selected. Furthermore, this paper discusses the role of persistent identifiers as well as metadata in the construction of citations. This work gives us an industrial perspective on how data citation is treated by the commercial world.

5. ACKNOWLEDGMENTS

Special Issue Editorial Board.

- Christine Borgman, UCLA, USA
- Paul Clough, University of Sheffield, UK
- Stefano Ferilli, University of Bari, Italy
- Edward Fox, Virginia Polytechnic Institute and State University, USA
- Norbert Fuhr, University of Duisburg-Essen, Germany
- Paul Groth, Elsevier Labs, The Netherlands
- Bradley Hemminger, University of North Carolina, USA
- Jaap Kamps, University of Amsterdam, The Netherlands
- Noriko Kando, National Institute of Informatics, Japan
- Christina Lioma, University of Copenhagen, Denmark
- Paolo Manghi, Consiglio Nazionale delle Ricerche, ISTI-CNR, Italy
- Andreas Rauber, Vienna University of Technology, Austria
- Seamus Ross, University of Toronto, Canada
- Heiko Schuldt, University of Basel, Switzerland
- Costantino Thanos, Consiglio Nazionale delle Ricerche, ISTI-CNR, Italy

6. REFERENCES

[1] Reproducibility and reliability of biomedical research: improving research practice. Technical report, The Academy of Medical Science, 2015.

[2] Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. In J. Freire, N. Fuhr, and A. Rauber, editors, *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science*, Dagstuhl Reports, Volume 6, Number 1. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Germany, 2016.

[3] M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello. DIRECTions: Design and Specication of an IR Evaluation Infrastructure. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012)*. Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany, 2012.

[4] M. Agosti, N. Ferro, and C. Thanos. DESIRE 2011 Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation. *SIGIR Forum*, 46(1):51–55, June 2012.

[5] M. Altman and M. Crosas. The evolution of data citation: From principles to implementation. *IAssist Quarterly* , 37(1–4):62–70, 2013.

[6] M. Altman and G. King. A Proposed Standard for the Scholarly Citation of Quantitative Data. In *IASSIST 2006 - Data in a World of Networked Knowledge*. IASSIST, 2006.

[7] M. Angelini, N. Ferro, B. Larsen, H. Müller, G. Santucci, G. Silvello, and T. Tsikrika. Measuring and Analyzing the Scholarly Impact of Experimental Evaluation Initiatives. *Procedia Computer Science*, 38:133–137, 2014.

[8] J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum*, 49(2):107–116, December 2015.

[9] K. Baggerly. Disclose all Data in Publications. *Nature*, (467):401, 2010.

[10] A. Bardi and P. Manghi. A Framework Supporting the Shift from Traditional Digital Publications to Enhanced Publications. *D-Lib Magazine*, 21(1/2), 2015.

[11] T. Bloom, E. Ganly, and M. Winker. Data Access for the Open Access Literature: PLOS's Data Policy. *PLoS Biol*, 12(2), 2014.

[12] C. L. Borgman. The Conundrum of Sharing Research Data. *JASIST*, 63(6):1059–1078, 2012.

[13] C. L. Borgman. Why are the Attribution and Citation of Scientific Data Important? In National Academy of Sciences' Board on Research Data and Information, editors, *Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop*. National Academies Press: Washington DC, 2012.

[14] C. L. Borgman. *Big Data, Little Data, No Data*. MIT Press, 2015.

[15] J. Brase, Y. Socha, S. Callaghan, C. L. Borgman, P. F. Uhlir, and B. Carroll. *Research Data Management: Practical Strategies for Information Professionals*, chapter Data Citation: Principles and Practice, pages 167–186. Purdue University Press, 2014.

[16] P. Buneman. How to Cite Curated Databases and how to Make Them Citable. In *Proc. of the 18th International Conference on Scientific and Statistical Database Management, SSDBM 2006*, pages 195–203, 2006.

[17] P. Buneman, S. Cohen-Boulakia, S. B. Davidson, J. Frew, and V. Tannen. Computational challenges in data citation. Technical report, University of Pennsylvania, 2014.

[18] P. Buneman, S. B. Davidson, and J. Frew. Why data citation is a computational problem. *Communications of the ACM (CACM)*, forthcoming, 2016.

[19] P. Buneman, S. Khanna, K. Tajima, and W.-C. Tan. Archiving Scientific Data. *ACM Transactions on Database Systems (TODS)*, 29(1):2–42, March 2004.

[20] P. Buneman and G. Silvello. A Rule-Based Citation System for Structured and Evolving Datasets. *IEEE Data Eng. Bull.*, 33(3):33–41, 2010.

[21] A. Burton, H. Koers, P. Manghi, S. La Bruzzo, A. Aryani, M. Diepenbroek, and U. Schindler. On Bridging Data Centers and Publishers: The Data-Literature Interlinking Service. volume 544 of *Communications in Computer and Information Science*, pages 324–335. Springer, 2015.

[22] S. Callaghan, S. Donegan, S. Pepler, M. Thorley, N. Cunningham, P. Kirsch, L. Ault, P. Bell, R. Bowie, A. M. Leadbetter, R. K. Lowry, G. Moncoiffé, K. Harrison, B. Smith-Haddon, a. Weatherby, and D. Wright. 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, 2012.

[23] L. Candela, D. Castelli, P. Manghi, and A. Tani. Data Journals: A Survey. *Journal of the Association for Information Science and Technology*, 66(9):1747–1762, 2015.

[24] D. Carr and K. Littler. Sharing Research Data to Improve Public Health: A Funder Perspective. *Journal of Empirical Research on Human Research Ethics*, 10(3):314—316, 2015.

[25] R. Costas, I. Meijer, Z. Zahedi, and P. Wouters. The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. Technical report, Danish Agency for Culture, April 2013.

[26] B. Cronin. *The citation process. The role and significance of citations in scientific communication*. London: Taylor Graham, 1984.

[27] D. De Nart, D. Degl'Innocenti, and M. Peressotti. Well-Stratified Linked Data for Well-Behaved Data Citation. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, 2016.

[28] Martone M. (ed.), editor. *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles*. FORCE11, San Diego CA, 2014.

[29] N. Ferro, F. Crestani, M.-F. Moens, J. Mothe,

F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, editors. *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*. Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany, 2016.

[30] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1), June 2016.

[31] M. Force, N. Robinson, M. Matthews, D. Auld, and M. Boletta. 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*, 2016.

[32] J. Freire, P. Bonnet, and D. Shasha. 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*, pages 593–596, 2012.

[33] P. Groth, A. Gibson, and J. Velterop. The Anatomy of a Nanopublication. *Inf. Serv. Use*, 30(1-2):51–56, 2010.

[34] A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, editors. *Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015)*. Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany, 2015.

[35] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, USA, 2009.

[36] Y.-H. Huang, P. W. Rose, and C.-N. Hsu. Citing a Data Repository: A Case Study of the Protein Data Bank. *PLoS ONE*, 10(8), 2015.

[37] J. Klump, R. Huber, and M. Diepenbroek. DOI for Geoscience Data – How Early Practices Shape Present Perceptions. *Earth Science Informatics*, pages 1–14, 2015.

[38] CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of Cite, Out of Mind: The Current State of Practice, Polocy, and Technology for the Citation of Data. *Data Science Journal*, 12:1–67, September 2013.

[39] V. Papavasileiou, G. Flouris, I. Fundulaki, D. Kotzinos, and V. Christophides. High-Level Change Detection in RDF(S) KBs. *ACM Trans. Database Syst.*, 38(1):1, 2013.

[40] S. Pröll and A. Rauber. Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In X. Hu, T. Lin Young, V. Raghavan, B. W. Wah, R. Baeza-Yates, G. Fox, C. Shahabi, M. Smith, Q. Yang, R. Ghani, W Fan, R. Lempel, and R. Nambiar, editors, *Proc. of the 2013 IEEE International Conference on Big Data*, pages 307–312. IEEE, 2013.

[41] S. Pröll and A. Rauber. Asking the Right Questions - Query-Based Data Citation to Precisely Identify Subsets of Data. *ERCIM News*, (100), 2015.

[42] A. Rauber, A. Ari, D. van Uytvanck, and S. Pröll. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, 2016.

[43] N. Robinson-Garcia, E. Jiménez-Contreras, and D. Torres-Salinas. Analyzing data citation practices according to the Data Citation Index. *Journal of the American Society for Information Science and Technology (JASIST)*, 2015.

[44] G. Silvello. A Methodology for Citing Linked Open Data Subsets. *D-Lib Magazine*, 21(1/2), 2015.

[45] N. Simons. Implementing DOIs for Research Data. *D-Lib Magazine*, 18(5/6), 2012.

[46] M. Vernooy-Gerritsen. *Enhanced Publications: Linking Publications and Research Data in Digital Repositories*. Amsterdam University Press, 2009.