Measuring Dataset Impact: Data Citation as an Economic Process

GIANMARIA SILVELLO, University of Padua

This paper reports the extended abstract of the talk I'll give at the Information Retrieval and Interaction Fest in Honour of Peter Ingwersen. The main topic is data citation. I'll talk about some connections between data citation for structured database, economic theory and measures for dataset impact. Peter Ingwersen's contributions provided both inspiration for starting this work and perspectives for future work.

• **Information systems→Database management system engines.**

Additional Key Words and Phrases: Data citation

ACM Reference Format:

Gianmaria Silvello, 2016. Measuring Dataset Impact: Data Citation as an Economic Process. *Information Retrieval and Interaction Fest in Honour of Peter Ingwersen*. (October 2016), 3 pages.

1. EXTENDED ABSTRACT OF THE TALK

Citations are the cornerstone of scientific progress and knowledge propagation, and they can be considered as the currency of the of the system of sciences. Indeed, from a *social-constructivist* viewpoint citations are seen as one rhetorical device employed by scientists to support their results and convince people of their claim; whereas, from a *normative* viewpoint they serve as a symbolic payment of intellectual debts [Baldi, 1998].

Citation rules and practices have stratified for centuries for text, but they cannot be applied as they are for data [Borgman, 2015]; yet data are as vital to scientific progress and propagation as traditional publications are. The shift towards data-intensive scientific discovery [Hey, 2009] is challenging scholarly publications, the way in which scientific credit is attributed and how the impact of the scientific work is estimated. Recently, [Abadi et al., 2016] highlighted that the increasing emphasis on citations to assess science quality is discouraging large system projects, end-to-end tool building and large curated dataset sharing. Before this, [Ingwersen and Chavan, 2011] pointed out that, due to the lack of a "*deep and persistent mechanism for citing data*", there is no consistent and established way to know how a given dataset has been used in time, what analyses and results have been conducted on it and who contributed to its creation and curation. This affects also the fields of "webometrics" [Björneborn and Ingwersen, 2001; 2004] that, for this reason, does not deal with scientific datasets. Furthermore, [Chavan and Ingwersen, 2009] was among the firsts to individuate in the "lack of incentive for data originator(s) and manager(s)", one of the main aspects impairing "scientists and research institutions to make concerted efforts in the management, archiving and publishing of primary scientific data".

Relying on these important observations, we can state that data citation hugely impacts on knowledge building, scientific communication and research investment decisions. Nevertheless, despite of its relevance and the attention dedicated to the topic, the scientific community still does not exactly know what it means to cite data and how to cite data. Indeed, data citation poses a number of new questions:

- (*Identification problem*) What data can be cited? How do we define a data citable unit?
- (*Completeness problem*) When data is extracted from a large, complex, evolving database, how do we create an appropriate and informative citation for it?

- (*Fixity problem*) How do we guarantee that cited data will be accessible in their cited form?
- (*Correctness problem*) How do we verify if a citation is correct and if two citations refer to the same data?

[Buneman et al., 2016] explained why data citation is a computational problem and outline the main problem to tackle as: *"Given a database D and a query Q, automatically generate an appropriate citation C"*. This problem is faceted because here the term "database" is used very broadly referring to relational databases, hierarchical datasets (e.g., XML) and graph-based datasets (e.g., RDF) among others. Each of these types of databases present heterogeneous structures and functions and, as a consequence, seem to require specific solutions for addressing data citation problems [Silvello and Ferro, 2016]. In the last years, we have proposed several solutions for targeting the problem of citing data from the computational viewpoint. The first is a *rule-based citation system* that exploits the hierarchical structure of XML to provide human- and machine-readable citations to XML elements [Buneman and Silvello, 2010]. The second solution is built on top of the previous one and uses the idea of *database views* to define citable units as a key to specifying and generating citation of XML elements [Buneman et al., 2016]; this solution was then extended to handle relational databases in [Davidson et al., 2017]. The third uses a *machine learning approach* that learns a model from a training set of existing citations to generate citations for previously unseen XML elements [Silvello, 2016]. None of these solutions can be straightforwardly adopted for citing RDF for which we defined a methodology based on *named meta-graphs* to cite RDF sub-graphs, and create human-readable and machine-actionable data citations [Silvello, 2015].

The common aspect of these solution is that they are more or less explicitly based on views over databases (defined by hand or automatically learned from data). Views are the basis of state-of-the-art solutions for data citation and by means of them we define who is gaining credit for which a specific data subset. Given that citations are the currency of the science system, we must ensure that the credit attribution mechanism operated through a citation system is *fair*, *truthful* and *efficient* [Othman et al., 2016].

In this context, we see tight connections between data citation and economic principles. The question to address is if it possible to define an economic theory for credit attribution that guarantees an equilibrium making the data citation mechanism computationally efficient as well as fair and convenient for scientists and scientific institutions.
Up to now, all the discussion is oriented on data citation principles and on data citation computational solutions, but the fairness of a citation model is an important milestone for all the subsequent uses of data citations. One of the most relevant implications is that data citations will constitute the basis to define new impact measures and will directly influence future research investment and people careers.
In this talk we discuss data citation in general and its implications for measuring the impact, both economical and scientific, of scientific datasets.

The life-long contribution of Peter Ingwersen to the field of scientometrics as well as connected fields casts a light on the path we should follow for defining fair, truthful and efficient impact measures based on citations to data. Moreover, Peter's contribution will be fundamental in the future for defining a new field that may be

called *datametrics*, which will define impact indicators based on data citation patterns.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Peter Buneman for introducing him to the topic of data citation and Prof. Susan Davidson for all the discussions and join work to which this talk owns a lot.

REFERENCES

Abadi, D., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P. A., Carey, M. J., Chaudhuri, S., Dean, J., Doan, A., Franklin, M. J., Gehrke, J., Haas, L. M., Halevy, A. Y., Hellerstein, J. M., Ioannidis, Y. E., Jagadish, H. V., Kossmann, D., Madden, S., Mehrotra, S., Milo, T., Naughton, J. F., Ramakrishnan, R., Markl, V., Olston, C., Ooi, B. C., Suciu, D., Stonebraker, M., Walter, T., and Widom, J. (2016). The Beckman Report on Database Research. *Comm. of the ACM (CACM)*, 59(2):92–99.

Björneborn, L. and Ingwersen, P. (2001). Perspective of Webometrics. *Scientometrics*, 50(1):65–82.

Björneborn, L. and Ingwersen, P. (2004). Toward a Basic Framework for Webometrics. *JASIST*, 55(14):1216–1227.

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

Chavan, V. and Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, 10(S-14):S2.

S. B. Davidson, D. Deutsch, T. Milo, and G. Silvello (2017). A model for fine-grained data citation. In *Proc. of the biennial Conference on Innovative Data Systems Research (CIDR 2017)*, *accepted for publication.*

Hey, T. (2009). The Fourth Paradigm: Data-intensive Scientific Discovery, Microsoft Pr.

Ingwersen, P. and 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.

Othman, A., Papadimitriou, C. H., and Rubinstein, A. (2016). The Complexity of Fairness Through Equilibrium. *ACM Trans. Economics and Comput.*, 4(4):20.

Silvello, G. and 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.

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

Silvello, G. (2016). Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. *Journal of the American Society for Information Science and Technology (JASIST)*, in print:1–28.

Received October 2016;