Appendix to
Why data citation is a computational problem

Comm. ACM, September 2016

Peter Buneman
University of Edinburgh

Susan Davidson
University of Pennsylvania

James Frew
University of California, Santa Barbara

June 26, 2016

1 An annotated bibliography

Many of these references were suggested by participants in the workshop on "Computational Challenges in Data Citation" held at the University of Pennsylvania on 17-18 April 2014.[1]

Please note that all the references in the appendices are to papers listed in the appendix, some of which also appear in the references of the main paper.

Standards and Principles

A1 Rauber, A., Pröll, S. Scalable Dynamic Data Citation. Position Paper, Working Group on Data Citation (WG-DC), Research Data Alliance (RDA), 2015-03-23 draft version. `http://rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html`

generalizes [**??**, **??**]

A2 Uhlir, P.F. (Ed.) *For Attribution- Developing Data Attribution and Citation Practices and Standards.* National Academies Press, Washington, DC, 2012. `http://doi.org/10.17226/13564`

workshop summary

[1] `http://datacitation.eri.ucsb.edu`

A3 Bilder, G. DOIs unambiguously and persistently identify published, trustworthy, citable online scholarly literature. Right? Crossref Blog, September 20, 2013. http://blog.crossref.org/2013/09/dois-unambiguously-and-persistently-identify-published-trustworthy-citable-online-scholarly-literature-right.html

explores edge cases in DOI definition and registry implementations

A4 Chavan, V.S., Ingwersen, P. Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community *BMC Bioinformatics 10(Suppl 14),* S2 (November 10, 2009). http://doi.org/10.1186/1471-2105-10-S14-S2

proposes an architecture for a distributed data publishing environment

A5 Mooney, H., Newton, M. The Anatomy of a Data Citation. *Journal of Librarianship and Scholarly Communication 1,* 1 (2012), eP1035. http://doi.org/10.7710/2162-3309.1035

argues that information professionals must promote data citation as "an essential component of data publication, sharing, and reuse."

A6 Allen, L., Scott, J., Brand, A., Hlava, M., Altman, M. Credit where credit is due. *Nature 508* (17 April 2014), 312–313. http://doi.org/10.1038/508312a.

introduces a taxonomy of contributor roles, to facilitate fine-grained assignment of citation credit

Citation systems design and implementation

B1 Aalbersberg, I.J., Kähler, O. Supporting Science through the Interoperability of Data and Articles. *D-Lib Magazine 17,* 1/2 (January/February 2011). http://doi.org/10.1045/january2011-aalbersberg

discusses Elsevier's SciVerse ScienceDirect architecture

B2 Buneman, P. How to cite curated databases and how to make them citable. In *SSDBM 2006: 18th International Conference on Scientific and Statistical Database Management* (Vienna, 3–5 July 2006). IEEE Computer Society, Los Alamitos, CA, 2006, 195–203. http://doi.org/10.1109/SSDBM.2006.28

the original work on IUPHAR citation

B3 Altman, A., Crosas, M. The Evolution of Data Citation: From Principles to Implementation. *IASSIST Quarterly 37,* 1-4 (2013), 62–70. http://www.iassistdata.org/iq/evolution-data-citation-principles-implementation

background of the FORCE11 Data Citation Principles

B4 Pröll, R., Rauber, A., Scalable data citation in dynamic, large databases: Model and reference implementation. *Proceedings of the 2013 IEEE International Conference on Big Data*, pages 307–312, 2013.

B5 Pröll, R., Rauber, A., A scalable framework for dynamic data citation of arbitrary structured data. *DATA 2014 - Proceedings of 3rd International Conference on Data Management Technologies and Applications, Vienna, Austria, 29-31 August, 2014*, pages 223–230, 2014.

Data Citation and Linked Data

C1 Berners-Lee, T. Linked Data. (June 18, 2009); http://www.w3.org/DesignIssues/LinkedData.html.

rules for publishing data on the Web

C2 Berners-Lee, T. Cool URIs Don't Change. (1998); http://www.w3.org/Provider/Style/URI.html.

notes on designing URIs for stability and longevity

C3 Ayers, D., Völkel, M. Cool URIs for the Semantic Web. (December 3, 2008); http://www.w3.org/TR/cooluris/.

guidelines for using URIs with RDF

C4 Thompson, H.S. Naming on the Web: What scholars should want, and what they can have. In *CERN Workshop on Innovations in Scholarly Communication (OAI8)* (University of Geneva, 19–21: June 2013). http://indico.cern.ch/event/211600/session/3/contribution/2/attachments/331924/463111/scroll.pdf

overview of naming issues: binding, resolution, and management

C5 Heath, T., Bizer, C. Linked Data: Evolving the Web into a Global Data Space (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology 1,* 1 (2011), 1–136. http://doi.org/10.2200/S00334ED1V01Y201102WBE001.

see chapter 6 "Consuming Linked Data", section 6.4 "Effort Distribution between Publishers, Consumers and Third Parties"

C6 Memento Guide - Resource Versioning and Memento. (January 19, 2015); http://mementoweb.org/guide/howto/

supporting the functionality of time-versioned URIs with the Memento protocol

C7 Van de Sompel, H., Nelson, M. Thoughts on Referencing, Linking, Reference Rot. (December 28, 2013); http://mementoweb.org/missing-link/.

requirements for robust citations to web resources

Data Citation and Reproducibility

D1 Peng, R. Reproducible Research in Computational Science. *Science 334,* 6060 (2 December 2011), 1226–1227. http://doi.org/10.1126/science.1213847.

introduction to the reproducibility problem

D2 Schopf, J. Treating data like software: a case for production quality data. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (Washington, DC, June 10–14, 2012). ACM, New York, 2012, 153–156. http://doi.org/10.1145/2232817.2232846.

argues that software release engineering principles should be applied to data

D3 Mesirov, J.P. Accessible Reproducible Research. *Science 327,* 5964 (22 January 2010), 415–416. http://doi.org/10.1126/science.1179653.

proposes a generic framework for supporting reproducible computational science

D4 Alper, P., Belhajjame, K., Goble, C., Karagoz, P. Enhancing and abstracting scientific workflow provenance for data publishing. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops* (Genoa, Italy, March 18–22, 2013). ACM, New York, 2013, 313–318. http://doi.org/10.1145/2457317.2457370.

discusses the relationship between provenance and data citation

Use Cases

In this section we have included additional references to the examples in the main paper. In addition we have added references to databases that may present further challenges. Many RDF databases have been extracted from existing data sets, and in this process the data and metadata needed for citation have been lost; however the Experimental Factors Ontology [**??**] is directly represented in RDF and presents an interesting challenge. Also it has been suggested to us that the Encoded Archival Description [**??**] presents some interesting aspects of citation for semistructured data.

E1 Southan, C., Sharman, J.L., Benson, H.E., Faccenda, E., Pawson, A.J., Alexander, S.P.H., Buneman, P., Davenport, A.P., McGrath, J.C., Peters, J.A., Spedding, M., Catterall, W.A., Fabbro, D., Davies, J.A. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Research* Advance Access (October 12, 2015). http://doi.org/10.1093/nar/gkv1037.

the most recent version of the GtoPdb (formerly IUPHAR) database

E2 NASA. MODIS Web. (2015); http://modis.gsfc.nasa.gov/.

main website for the MODIS instruments and data products

E3 NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center. Citing Our Data. (April 14, 2014); http://lpdaac.usgs.gov/citing_our_data.

how to cite USGS MODIS data

E4 Oak Ridge National Laboratory (ORNL) DAAC. Data Product Citation Policy. (2015); http://daac.ornl.gov/citation_policy.html.

how to cite ORNL MODIS data

E5 Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H. Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics 26,* 8 (2010), 1112–1118. http://doi.org/10.1093/bioinformatics/btq099.

an ontology for gene expression data

E6 Pitti, D.V., Encoded archival description: The development of an encoding standard for archival finding aids. *The American Archivist*, pp.268-283. 1997.

E7 Chavan, V. *Recommended practices for citation of data published through the GBIF network. Version 1.0.* Global Biodiversity Information Facility, Copenhagen, 2012. http://links.gbif.org/gbif_best_practice_data_citation_en_v1

how to cite global biodiversity data

E8 Duerr, R.E., Downs, R.R., Tilmes, C., Barkstrom, B., Lenhardt, W.C., Glassy, J., Bermudez, L.E., Slaughter, P. On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics 4,* 3 (September 2011), 139–160. http://doi.org/10.1007/s12145-011-0083-6.

applicability of multiple identifier schemes for citing Earth science data

Background reading on Databases and XML

Most good textbooks on databases (e.g. [**??**, **??**]) cover both relational databases and the basics of XML. There is also plenty of on-line material related to these. Of relevance to the ideas in this paper are the systems that convert or represent relational databases in some hierarchical form such as XML. [**??**] describes a sophisticated approach to this and also reviews what is practically available. Going in the other direction, a number of database systems provide for ingesting XML into tables, see [**??**,**??**].

F1 Ramakrishnan, R., Gehrke, J. *Database Management Systems (3rd Edition)*. McGraw-Hill, 2002.

F2 Garcia-Molina, H., Ullman, J.D., Widom, J. *Database Systems: The Complete Book (2nd Edition)* Pearson, 2008.

F3 Benedikt, M., Chan, C.Y., Fan, W., Rastogi, R., Zheng, S., Zhou. A. DTD-directed publishing with attribute translation grammars. *VLDB*, 2002.

F4 Bohannon, P., Freire, J., Haritsa, J. R., Roy, P., Siméon, J. LegoDB: Customizing relational storage for XML documents. In *VLDB* 2002

F5 Teradata Database, Tools and Utilities Release 15.00 `http://www.info.teradata.com/htmlpubs/DB_TTU_15_00/index.html#page/Teradata_XML/B035_1140_015K/XML_Shredding_Publishing.09.01.html`

Views

G1 Fan, W. Chan, C.Y., Garofalakis, M.N. Secure XML Querying with Security Views. *SIGMOD* 2004.

G2 Halevy, A.Y. Answering queries using views: A survey. *VLDB J.*, 10(4):270–294, 2001.

G3 Abiteboul,S., Duschka, O.M. Complexity of Answering Queries Using Materialized Views. *PODS 1998* 1998.

G4 Deutsch, A., Popa, L., Tannen, V. Query reformulation with constraints. *SIGMOD Record*, 35(1), 2006.

G5 Lenzerini, M. Data Integration: A Theoretical Perspective. *PODS 2002*, 233–246, 2002

Archiving

Some relational database management systems provide for *time travel*, originally proposed in [**??**] – the ability to see the database at any point in the past. Moreover this can be queried through a temporal extension of SQL [**??**]. Such temporal extensions are now part of the SQL2011 standard and have been implemented by systems such as DB2. Unfortunately database developers often fail to make use of them. Moreover there is a concern that the long-term preservation of a database should be dependent on the maintenance of relatively complex database software. Work by Rauber on provenance and citation, cited in the main paper, provides an archiving system for tabular data.

For hierarchical data, [**??**, **??**] provide an approach that works by pushing temporal variation down into the hierarchy. Full web archiving [**??**] will be of enormous benefit, but for our purposes this will need to be extended to the "deep" Web.

H1 Stonebraker, M. and Kemnitz, G. The POSTGRES next generation database management system. *Communications of the ACM*, 34(10), pp.78-92. 1991

H2 Snodgrass, R.T. *Developing Time-Oriented Database Applications in SQL* Morgan Kaufmann. 1999.

H3 Buneman, P., Khanna, S., Tajima, K. and Tan, W.C. Archiving scientific data. ACM Transactions on Database Systems *TODS*, 29(1), pp.2-42. 2004

H4 Müller, H., Buneman, P. and Koltsidas, I., XArch: archiving scientific and reference data. *ACM SIGMOD 2008* (pp. 1295-1298). 2008.

H5 Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S. and Shankar, H. Memento: Time travel for the web. *arXiv preprint arXiv:0911.1112.* 2009

2 Citable units

Although the authors have heard the term "citable unit" in widespread use at meetings, there does not appear to be any authoritative description of the term. The term is used in some on-line blogs (themselves difficult to cite), but these appear to take the term as given rather than defining it. `http://serialmentor.com/blog/2015/1/2/what-constitutes-a-citable-scientific-work/` provides some criteria for a citation and appears to use "citable unit" to describe what we term the referent of the citation. `https://www.aje.com/en/author-resources/articles/nanopublications-and-mini-monographs` takes the term as given in order to describe a *nanopublication*.

We do not think that citable units coincide with the referents of DOIs. In our examples, the number of possible citatable units far exceeds the number of DOIs one would want to assign to a database. Also there is a subtlety, which we did not address, about a possible distinction between a citable unit and a citation. A citation such as "John Doe, The Impossibility of Reason, Elsinger 1954, chapter 3, paragraph 17" claims that the moon is made of green cheese." contains what one might regard as a citable unit – "John Doe, The Impossibility of Reason, Elsinger 1954" and a location – "chapter 3, paragraph 17" at which one finds the claim. Although the location is not really part of the citable unit, it is an invaluable part of the citation itself. Providing such location information is especially important for the substantial number of people who are employed as data curators to verify that citations are correct, in the sense that cited material has been correctly interpreted in the database. Finding where a particular claim is made in a long paper can be very time-consuming.

The provision of location information in a citation fits well with the machinery we have presented, but we did not discuss this in the paper.

3 Views

The original motivation for answering queries using views was for efficiency: suppose one has already computed the answers (called *materialized views*) to one or more queries against a database. Given a new query, it may be more efficient to compute the answer to that query using those views rather than applying it directly to the database. This presupposes that the query can be factored through those views. A second stimulus comes from data integration and is closely linked to partial or incomplete information in databases. These are explained in a survey paper [**??**]. The former stimulus, answering queries using materialized views, is closer to the problem presented in this paper. Another use of views, which does connect with our proposals, is for security: The "publishers" of the database associate a security level with each view. The security level needed to answer a query is determined by how the query can be answered using views. See [**??**] for details

There is a huge literature on this topic. Complexity issues are treated in [**??**]. Also relevant to data citation are how to rewrite in the presence of constraints [**??**] and how it might work in other (non relational) representations of data [**??**].