Exploring Openness in Data and Science: What is "Open," to Whom, When, and Why?

Irene V. Pasquetto
Ashley E. Sands
Christine L. Borgman
Department of Information Studies
University of California, Los Angeles
irenepasquetto@ucla.edu
ashleysa@ucla.edu
christine.borgman@ucla.edu

ABSTRACT
"Open data" is a popular phrase in research practice and science policy. While stakeholders agree on some aspects of this concept, many others remain hotly debated. As a means to identify the main themes and arguments surrounding open data, we analyzed highly cited publications from the last 10 years that address data sharing or open access to research data. We identify and synthesize eight components of open data that vary among policies, studies, and initiatives, and present problematic arguments worthy of further investigation.

Keywords
Open data, data sharing, data access, research data.

INTRODUCTION
Despite common use, the term "open data," and similar terms such as "open science" and "open research," are poorly understood. "Open" has similar vagaries to the use of "free," as in "free software." Richard Stallman (2002) famously quipped, "free speech" is not the same as "free beer." Similarly, data, publications, science, and scholarship can be "open" in multiple respects. Some uses are more aligned with free speech, some with free beer, and others are simply new business models. Establishing clear distinctions among these uses will inform research practice and public policy.

In scholarly communities, open data is most often conflated with "data sharing," itself a vague term. Data might be made available to others on demand, by request, posted, deposited, or otherwise released – all over varying time frames, license and use agreements, and degrees of usability (Borgman, 2015). The ability to make data available depends on technical, legal, economic, ethical, policy, disciplinary, and other factors. For example, the Organisation for Economic Co-operation and Development (OECD) (2007) places 13 conditions on "openness" for access to research data created under public funding. Data practices and policies vary widely by field and by stakeholders, including scholars, publishers, librarians, students, funders, policy makers, and the public at large.

While a comprehensive overview of the state of research on data sharing and open data practices is infeasible for such a fast-moving topic, framing the main themes in the current literature and policy reports is a reasonable goal. Toward this end, we selected highly cited publications from the last 10 years that address open data or data sharing. We sought an exploratory sample from multiple domains, including science and technology studies, eResearch, information studies, science policy, and humanities. We also drew on selected findings of the UCLA Center for Knowledge Infrastructures. Based on these sources, we synthesize the dimensions of "open data," drawing comparisons and contrasts. The full list of sources analyzed is provided in an open Zotero library (Pasquetto, Sands, & Borgman, 2015; UCLA KI Team, 2015).

DIMENSIONS OF OPEN DATA
We identified eight dimensions that encompass debates associated with open data and data sharing (Figure 1): definitions of open data; sources of research data; benefits of open research data; scale of data sharing; ownership, licensing, and legal status; means of dissemination; technical access; and preserving data for future access.

Definitions of open data
We found many definitions of open data in research contexts, which tend to converge on two factors: technical and legal availability. Most cited are the OECD Principles and Guidelines for Access to Research Data from Public Funding. Here, openness is defined as "access on equal terms for the international research community at the lowest

ASIST 2015, November 6-10, 2015, St. Louis, MO, USA.

possible cost, preferably at no more than the marginal cost of dissemination" (Organisation for Economic Co-operation and Development, 2007). Among stakeholders, consumers of research data are most often the international research community and occasionally the general public. These policy definitions rarely specify to what extent open data need to be technically and legally open. Rather, they offer generic expressions such as "fewest restrictions" and "lowest possible cost."

Sources of research data

"Openness" debates about research data usually refer to data collected using public funds. "Research data" are distinguished from other forms of data such as government statistics and industry records. Examples of open research data initiatives include repositories and archives (e.g. GenBank, Protein Data Bank, Sloan Digital Sky Survey), federated data networks (e.g. World Data Centers, Global Biodiversity Information Facility; NASA Distributed Active Archive Centers), virtual observatories (e.g. International Virtual Observatory Alliance, Digital Earth) and institutional repositories (e.g., PubMedCentral, arXiv). While the focus of our analysis is scientific data, Open Knowledge Foundation (2014) includes scientific data as only one category among seven possible types of open data: cultural, scientific, financial, statistical, weather, environmental, and transport.

Benefits of open research data

Most policy rationales for open access to research data focus on the potential economic benefits of sharing. Such discourse also addresses how open research data might lead to more affordable, efficient, trustworthy, and reproducible science. Other authors argue that open access to research data can have broader social benefits such as educational tools to train new scientists, shared common resources to promote capacity building in developing countries, and the ability for crowdsourced and citizen science projects to promote scientific public outreach and engagement. Borgman (2015) identified four rationales for sharing research data: to reproduce research, to make public assets available to the public, to leverage investments in research, and to advance research and innovation. Very few studies on the actual benefits of opening research data exist (Beagrie & Houghton, 2014). Sabina Leonelli (2013) suggests that the allure of big and open data "lies precisely in the impossibility to predict and quantify their potential as evidence in advance."

Scale of data sharing

Despite increasing policy efforts, a system of coordinated and efficient international data exchange is still in its infancy. We identified tensions between policy maker requests for international open research data and scholars who describe the challenges of sharing scientific data at a local level. Sharing data with unknown audiences is yet harder. Research data sharing practices vary greatly depending on the format of data, the ways that data are handled and interpreted, and the disciplines or fields involved (Wallis, Rolando, & Borgman, 2013). A primary challenge to data sharing and integration is decisions made locally about curation and documentation. According to Goodman, et al. (2014), the ability to produce reusable research data is increased if individuals conduct their research with a "data reuse" perspective in mind. Studies of local data sharing practices tend not to use terms such as "open data" when referring to data made legally and technically available.

Ownership, licensing, and legal status

Data creators can preserve their intellectual property rights while providing open access. Creative Commons licenses (CC) are widely used for data, but are an incomplete solution. While copyright protection applies to acts of creativity, many types of data are considered "facts." The Open Data Commons licensing system addresses this intellectual property and licensing conundrum (Miller, Styles, & Heath, 2008). Other legal aspects of open data include privacy and ethical issues, which vary greatly by type of data. Medical and other personal data are the obvious examples, but data about cultural heritage or environmental sites also can be sensitive. The more general problem is the lack of agreement on who owns, or should own, research datasets (Borgman, 2015).

Means of dissemination

Data can be disseminated by many means. Methods include depositing datasets in archives, repositories, domain-specific collections, or library collections; making them supplemental materials with journal articles; posting them on personal or laboratory websites, or privately exchanging them between individuals (Wallis et al., 2013). Scholarly publishers are more eager to link journal articles to datasets than they are to host data repositories themselves. The notion of "publishing" datasets is itself controversial, given the broad connotations of scholarly publishing (Parsons & Fox, 2013). Depositing data in archives is generally viewed as the gold standard, as data become more discoverable, sustainable, linkable, and citable. However, the availability of data archives varies widely by domain, data type, and country, and many of these archives have only short term grant funding. Commercial services such as Figshare, Slideshare, and SSRN provide immediate access, but not necessarily long-term sustainability. Other kinds of commercial services are appearing, as are data journals in which datasets can be contributed as citable publications. The means to disseminate open data are myriad, complex, and evolving.

Technical access to data

Technical access to research data, whether open or not, varies by means of dissemination, such as whether datasets are contributed to repositories, linked to journal articles, or posted on personal or lab web sites. Linked Open Data, based on World Wide Web Consortium standards, provides basic methods for linking data to publications, protocols, and related objects (Bizer, Heath, & Berners-Lee, 2009).

More complex methods of modeling relationships between datasets and other research objects are being adopted. These include Object Reuse and Exchange (Van de Sompel et al., 2012), Resource Sync (Pepe, Mayernik, Borgman, & Van de Sompel, 2010), and Scholarly Research Objects (Bechhofer et al., 2010). Yet another model is "Linked Open Science" (Kauppinen & Espindola, 2011) that supports "executable papers" in which tools and data for reproducing analyses are embedded. Digital Object Identifiers are being applied to datasets, usually at the point of ingest into a data archive or repository ("What is DataCite?," 2012). Technical strategies for opening access to data are evolving rapidly.

Preserving data for future access

The Open Archival Information System (OAIS) is the most widely adopted interoperability standard across disciplines (Consultative Committee for Space Data Systems, 2012). For data to remain open, they must be managed, stored, and curated. Accompanying contextual information is necessary for access and retrieval. Little agreement exists on how long any individual dataset might be useful and thus how long it should be preserved, by whom, or for whom. Given the potential long-lived nature of research data, librarians and archivists are particularly concerned about who will take responsibility for data management, storage, curation, access, and sustainability.

DISCUSSION AND CONCLUSION

Data sharing, open access to data, and open data are topics of debate among a broad range of stakeholders. These stakeholders agree on a few points, such as a primary concern for access to research data collected using public funds. Most promote the potential economic and qualitative research benefits of open access to research data. However, stakeholders disagree on most other aspects of how, when, and why to provide open access to research data.

We found that policies, studies, and initiatives on open research data vary by eight dimensions (Figure 1 and Appendix 1): definitions of open data (from more to less inclusive); sources of research data (type of funding and domains); benefits of open research data (economic, social, political, ethical); scale of data sharing (global, local); ownership, licensing, and legal status (responsibility, control); technical access to data (models, standards, protocols); preserving data for future access (manage, store, curate); and means of disseminating data (depositing, linking, posting).

Through our analysis, we identified three disjointed perspectives on open data. First, the relationship between open access to research data and the necessity of public engagement varies between stakeholders. Second, few studies have empirically tested the advantages of open data. Third, the terms "open data" and "data sharing" are used in many competing ways; openness and sharing are sometimes conflated, other times used as synonyms, and still other times used to refer to distinct practices. This poster is our first publication examining what openness means for data in science, with fuller explications in progress.

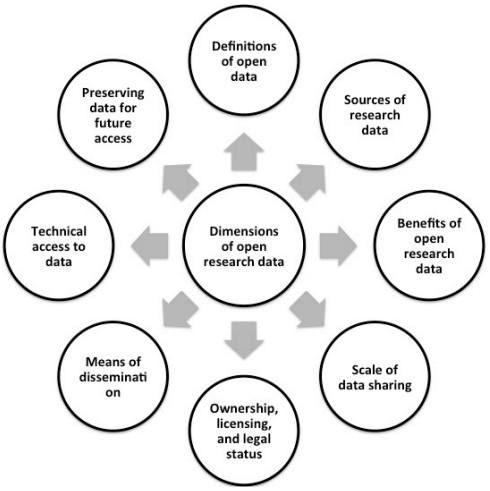

Figure 1. Eight dimensions of debate about "open data"

ACKNOWLEDGEMENTS

This research is funded by the Alfred P. Sloan Foundation, #20113194: The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective. We thank other members of the UCLA Center for Knowledge Infrastructures for comments on earlier drafts of this poster: Peter T. Darch, Milena Golshan, and Sharon Traweek.

REFERENCES

Beagrie, N., & Houghton, J. W. (2014). *The Value and Impact of Data Sharing and Curation - A synthesis of three recent studies of UK research data centres* (pp. 1–26). JISC (Joint Information Systems Committee). http://repository.jisc.ac.uk/5568/

Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I., Couch, P., Cruickshank, D., … Sufi, S. (2010). Why Linked Data is Not Enough for Scientists. In *Proceedings of the 2010 IEEE Sixth International Conference on e-Science* (pp. 300–307). Brisbane, Australia: IEEE Computer Society. http://doi.org/10.1109/eScience.2010.21

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. http://doi.org/10.4018/jswis.2009081901

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press. http://mitpress.mit.edu/big-data

Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System* (No. CCSDS 650.0-M-2 Magenta Book). Washington, D.C. http://public.ccsds.org/publications/archive/650x0m2.pdf

Goodman, A. A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A.

(2014). Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology*, *10*(4), e1003542. http://doi.org/10.1371/journal.pcbi.1003542

Kauppinen, T., & Espindola, G. M. de. (2011). Linked Open Science-Communicating, Sharing and Evaluating Data, Methods and Results for Executable Papers. *Procedia Computer Science*, *4*, 726–731. http://doi.org/10.1016/j.procs.2011.04.076

Leonelli, S. (2013). Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology. *Bulletin of Science, Technology & Society*, *33*(1-2), 6–11. http://doi.org/10.1177/0270467613496768

Miller, P., Styles, R., & Heath, T. (2008). Open Data Commons, a License for Open Data. In *Proceedings of the WWW2008 Workshop on Linked Data on the Web* (Vol. Session: Populating the Web with Linked Data). Beijing, China. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/

Open Knowledge Foundation. (2014, October 10). Open Knowledge: What is Open? https://okfn.org/opendata/

Organisation for Economic Co-operation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding* (p. 24). Paris: Organisation for Economic Co-Operation and Development. http://www.oecd.org/dataoecd/9/61/38500813.pdf

Parsons, M. A., & Fox, P. A. (2013). Is Data Publication the Right Metaphor? *Data Science Journal*, *12*, WDS32–WDS46. http://doi.org/10.2481/dsj.WDS-042

Pasquetto, I. V., Sands, A. E., & Borgman, C. L. (2015). Open Data in Science Bibliography. https://www.zotero.org/groups/asist_2015_pasquetto_etal

Pepe, A., Mayernik, M., Borgman, C. L., & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology*, *61*(3), 567–582. http://doi.org/10.1002/asi.v61:3

Stallman, R. M., & Gay, J. (2002). *Free Software, Free Society: Selected Essays of Richard M. Stallman* (1st. ed). Boston, Mass: Free Software Foundation.

UCLA KI Team. (2015). Knowledge Infrastructures: UCLA. https://knowledgeinfrastructures.gseis.ucla.edu/

Van de Sompel, H., Sanderson, R., Klein, M., Nelson, M. L., Haslhofer, B., Warner, S., & Lagoze, C. (2012). A Perspective on Resource Synchronization. *D-Lib Magazine*, *18*(9/10). http://doi.org/10.1045/september2012-vandesompel

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, *8*(7), e67332. http://doi.org/10.1371/journal.pone.0067332

What is DataCite? (2012). https://www.datacite.org/about-datacite/what-do-we-do

APPENDIX 1 Agreement and Disagreement among stakeholders in regards to eight dimensions of open research data

DIMENSION	AGREEMENT AMONG STAKEHOLDERS	DISAGREEMEMENT AMONG STAKEHOLDERS
Definitions of open data	Research data should be technically and legally available.	Potential consumers of research data: international community, local scientists, the general public, etc.
Sources of research data	Research data that are publicly funded and used for research purposes.	Relationships between research, government, and industry data.
Benefits of open research data	Open access to data may make science more trustworthy and reproducible.	Potential social benefits of opening research data: education, crowdsourcing, and citizen science projects.
Scale of data sharing	Varies by type of data, domain, and country.	Scales of time, geography, domain, licensing, and assorted policy factors.
Ownership, licensing, and legal status	Creative Commons and Open Data Commons licenses are popular.	Who owns, who should own data, how to assign rights; rights stacking.
Means of dissemination	Archives, repositories, and libraries can play important roles.	Disparate ways to disseminate data in the short and long term.
Technical access to data	Digital Object Identifiers and Linked Open Data are lowest common denominator solutions.	Competing technologies and models that vary by domain, type of data, policy, and technical environments.
Preserving data for future access	Data should be documented following community practices.	Practices vary by stakeholder, community, availability of archives, economics, and many other factors.