IS DATA PUBLICATION THE RIGHT METAPHOR?

M A Parsons[1]* and P A Fox[2]

**[1]National Snow and Ice Data Center, University of Colorado, UCB449, Boulder, CO 80309*
Email: parsonsm@nsidc.org
[2]Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180
Email pfox@cs.rpi.edu

ABSTRACT

International attention to scientific data continues to grow. New opportunities emerge to re-visit long-standing approaches to managing data and to critically examine proposed new capabilities. We describe the importance of metaphor and discuss several existing data management metaphors. We question the applicability of a "publication" approach to making data broadly available. We examine several metaphors and compare and contrast some of their attributes to stimulate an open dialog among international data stakeholders. Our preliminary conclusions are that no one metaphor satisfies enough key data system attributes and that multiple metaphors need to co-exist in support of a healthy data ecosystem.

Keywords: data publication, data system design, data citation, data on the Web, data preservation, cyberinfrastructure.

1 INTRODUCTION

Data "publication" is all the rage. Data authors and stewards rightfully seek recognition for the intellectual effort they invest in creating a good data set. At the same time, good data sets should be respected and handled like first class scientific objects. As a result, people look to scholarly publication—a well-established, scientific process—as a possible analog for sharing data. Data publication is becoming a metaphor of choice to describe the desired, rigorous, data management approach that creates and curates data as first class objects. The emerging International Council for Science World Data System[1] and the American Geophysical Union[2] both explicitly advocate data publication as a mechanism to facilitate data release and recognition of providers. Costello (2009) even argues that the principles of "publication" rather than "sharing" are necessary to address concerns about data openness and availability. While we strongly support these efforts to recognize data providers and to improve and professionalize data management, we argue in this essay that the data publication metaphor can be misleading and may even countermand aspects of good data stewardship. We suggest it is necessary to consider other metaphors and frames of thinking to adequately address modern issues of data science.

2 THE IMPORTANCE OF METAPHOR

Metaphor is important not only in how we communicate, but also in how we think. As Lakoff and Johnson (1980) state in their seminal book *Metaphors We Live By*:

> Metaphor is for most people a device of the poetic imagination and the rhetorical flourish — a matter of extraordinary rather than ordinary language. Moreover, metaphor is typically viewed as characteristic of language alone, a matter of words rather than thought or action. For this reason, most people think they can get along perfectly well without

1 http://wds-kyoto-2011.org/WDS_Conference_Preliminary_Report.pdf

2 AGU position statement on "The Importance of Long-term Preservation and Accessibility of Geophysical Data" at http://www.agu.org/sci_pol/positions/geodata.shtml

> metaphor. We have found, on the contrary, that metaphor is pervasive in everyday life, not just in language but in thought and action. *Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature* (p. 3, our emphasis).

Lakoff (1980; 2008) further argues that metaphors contain their own little stories that create "frames" or "scripts"—cognitive structures that we think with and use to understand and respond to events and concepts[3]. These frames present a set of roles and relationships between them. Consider, for example, "author," "editor", "publisher," "reviewer," and "librarian" as obvious roles that emerge from the contemporary publication metaphor. Most importantly, thinking in these frames provides a structure that presents a limited number of possible scenarios. Therefore, metaphors and framing can be extremely useful for describing and conceptualizing new ideas or paradigms, but they can also restrict our thinking and prevent us from seeing necessary alternatives or new possibilities.[4] In this light, we examine data publication and other metaphors as part of modern approaches to data science. We seek to stimulate a dialog among data scientists that helps us conceptualize more diverse approaches that address issues of science data stewardship *and* data-intensive science. We focus on observational and modeled (rather than experimental) sciences, especially interdisciplinary Earth system science, but we believe our ideas, our metaphors, apply broadly.

3 CURRENT METAPHORS AND PARADIGMS

Currently, we see (at least) four broad approaches to data management in Earth system science. These approaches are not mutually exclusive, but they come from different communities, use different metaphors, and are geared toward different data types. We will call these four paradigms Big Iron, Data Publication, Map Making, and Linked Data (Table 1).

The Big Iron approach is akin to industrial production and typically deals with massive volumes of data. The data are relatively homogenous and well defined but are highly dynamic and have a high throughput. The Big Iron itself is a large, sophisticated, well-controlled cyberinfrastructure potentially involving supercomputing centers, dedicated networks, substantial budgets, specialized interfaces often focused on reducing actual data transfer, etc. The Big Iron focus tends to be on managing throughput and enabling access to large volumes, although archiving is an increasing concern. Big Iron systems rely heavily on data and metadata standards and typically use relational (e.g., MySQL) and hierarchical (e.g. HDF) data structures. Examples of the Big Iron approach include the European Space Agency's Science Archives5, NASA's Earth Observing System (EOS)[6] and the National Science Foundation's TeraGrid (now called the Extreme Science and Engineering Discovery Environment (XSEDE))[7].

The Data Publication approach seeks to be analogous to scholarly journal publication. Its focus is often on research collections where data are extremely diverse in form and content, but tend to be small in size. Data Publication seeks to define discrete, well-described data sets, ideally with a certain level of quality assurance or peer-review. These data sets often provide the basis for figures, tables, etc. in research articles and other publications. Published data are then considered to be first class, reference-able, scientific artifacts, and they are often closely associated with peer-reviewed journal articles. The Data Publication focus tends to be on curation, archiving, and data quality. Data management systems, like the data, are not well standardized but tend to use relational or hierarchical data structures like Big Iron. Further, the standards used across different data systems are very high level, e.g. Dublin Core. Examples of Data Publication can be found in a variety of libraries and university repositories. An especially recognized advocate of data publication is the PANGAEA®[8] system in Germany.

Map Making is most readily seen in so-called spatial data infrastructures and their associated geographic information systems (GIS). Map Making could be seen as a subset of the data publication paradigm, but here the

---

3 Framing and frame analysis is used in social theory, media studies, and psychology. Much of the work stems from Erving Goffman's Frame Analysis: An Essay on the Organization of Experience (1974).

4 Computer science and applications is an area where metaphor is particularly prevalent and obvious. Consider for example, the classic "desktop" metaphor used in many PC operating systems. This was initially very useful in helping new computer users understand and conduct basic operations. More recently, however, the desktop and file metaphors present a limiting frame that need to be reconsidered when developing more modern, less hierarchical, cloud-based, mobile operating systems.

5 http://www.sciops.esa.int/index.php?project=SAT&page=index

6 http://eospso.gsfc.nasa.gov/

7 https://www.xsede.org/

8 http://pangea.de

analogous publication is a map or an atlas rather than a journal article. Data are necessarily geospatially focused and tend to be more fixed in time, i.e. they are more geared toward describing geospatial features rather than processes. The Map Making focus tends to be on cartographic visualization and intercomparison with uneven attention to preservation. Data are well standardized around a map- (or occasionally grid-) based model with an associated (geo)database. Major examples of map-based systems include the INfrastructure for SPatial InfoRmation in Europe (INSPIRE)[9] and Geodata.gov in the United States. These major systems also share some traits with the Big Iron systems, such as top-down governance and well-defined data models and protocols.

Linked Data[10] is based on computer science concepts of webs of data and relies on many underlying principles behind the Semantic Web[11]. The "data" in Linked Data are defined extremely broadly and are envisioned as small disparate bits with specific names (URIs) interconnected through defined semantic relationships (predicates) using model and language standards (e.g. the Resource Description Framework; RDF[12]). The focus to date has been almost entirely on enabling interoperability and capitalizing on the interconnected nature of the web. There is also a major emphasis on "open" data. Scant attention is paid to preservation, curation, or quality. An underlying principle of the approach is that it uses a "graph" model not a hierarchical or relational model. This lends itself well to very distributed and interdisciplinary connections but also requires substantial agreement on the formal semantics, i.e. ontologies, to be useful for diverse audiences. This approach has been applied in a variety of contexts outside science and increasingly in life and medical sciences. There is growing discussion and use in the Earth sciences, such as in the Integrated Ocean Drilling Program (IODP)[13].

9 http://inspire.jrc.ec.europa.eu/

10 http://linkeddata.org

11 http://linkeddata.org

12 http://w3.org/RDF

13 http://www.iodp.org/

Table 1. Summary of attributes of the major data management metaphors or paradigms

	Big Iron	*Data Publication*	*Map Making*	*Linked Data*
Analog	industrial production	scholarly publication	cartography	World Wide Web documents
Data	large and homogenous	small and diverse	geospatial features	disparate and named
Data models	hierarchical	hierarchical or relational	geospatial and relational	linked graph
Focus	throughput and manageable access	quality, certification, and preservation	data visualization and intercomparison	interoperability and interconnection
Examples in science	EOSDIS, TeraGrid	PANGEA, university repositories	INSPIRE, Geodata.gov	IODP, MyGrid, Linked Open Government Data
Metaphorical terminology	data producer, processing level, version release	data author, publisher, data citation	data source, feature, layer	data provider, name, link, uri?

4 LIMITS OF THE CURRENT PARADIGMS

These short descriptions of the different data management paradigms (further summarized in Table 1) are cursory, simplistic, and even stereotypical. That is the point. People tend to think and speak in the frames they know and understand, and that is in itself revealing. We, the authors, have often struggled and witnessed others struggle with communicating across paradigms. As an example, one of us coming from the data publication world encountered serious confusion and miscommunication with a new partner from the Big Iron world. We were talking past each other for days until we finally realized we had very different conceptions of the basic notion of "data set."

This example illustrates why it's important to consider the language and metaphors of the different paradigms and how they may limit understanding. Sometimes the limitations are obvious like trying to use a geographic map to describe a new genotype. But sometimes the limits are more subtle. For example, a data "producer" may not carry the same intellectual weight as a data "author." Perhaps most critical is the actual term data "publication." A publication is typically viewed as complete and only subject to minor updates, whereas data can be very dynamic, even long after they are made available. More seriously, publication implies some level of imprimatur (Callaghan et al., 2009). Especially when associated with literary publication, a "published" data set may be assumed to have undergone some sort of peer-review, yet there are no standards or even agreement on what peer-review of a data set might mean (Parsons et al., 2010). Some communities have made admirable efforts to peer review data, but it is not really the same as traditional peer-review of literature and the approaches vary. For example, the Planetary Data System has a long established peer-review scheme, but it is actually more like an audit that assures that a data set adheres to best practices of documentation, format, error characterization, etc. (McMahon, 1996). Other data peer review efforts tend to combine the review of the data set with review of a more conventional literary publication (e.g., Callaghan et al., 2009; Pfeiffenberger and Carlson, 2011).

Data publication is also closely associated with data citation, a concept we strongly support, but we feel that the publication metaphor has created some false expectations around data citation. A primary purpose of data citation is to aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study.[14] This means that data need to be citable, ideally with a persistent identifier like a Digital Object Identifier (DOI), as soon as they are available for use by anyone other than the original creator. Despite this, we have heard time and again from different data managers that they do not want to assign a DOI to a data set until it has reached some level of stability and quality control. They really do see publication and the assignment of DOI as a sort imprimatur. They do not seem to recognize the often-broad use and evolution of a data set long before it may be formally published. It appears that the publication metaphor limits how a data set may be conceived and represented. Indeed, even when the dynamics of the data cycle are more explicitly addressed, concepts from literary publication such as "peer review, registration, persistence, bibliographic description, etc." still dominate (e.g., Penev et al., 2009).

In fact, it is the close association with copyright, restricted-access, literary publications, etc. that troubles us most about the data publication metaphor. While most scholarly publishers agree that data should be openly available regardless of the restrictions on the article, they still assume most data discovery comes through the article and that most data sets have at least one peer-reviewed article associated with them—an arguable assumption at best. Some even argue that data publication necessarily includes licensing of the data set and mandating conditions of use (Klump et al., 2006). Further, if publication citations become a primary means of identifying data, an unintended side effect may be to actually limit data access and discovery and reinforce the hidden "deep web of data" (Wright, 2009). We believe that the Data Publication perspective is, as John Wilbanks (2009, personal communication[15]) says, focused on "the container and not the customer". It requires publishers to spend undue time managing the definition of and access to the container, be it an article or a data set. Yet in a networked world, the proliferation of copies and the customer's ability to annotate, federate, transform, and integrate the content makes the content more valuable. In short, the data publication container can restrict access, interoperability, and creative use, while we believe *unlocking* the data in the "deep web" should be a major priority.

We have focused on the limitations of the data publication metaphor because it is often advocated as a desired goal of data science, but all the described paradigms are lacking in some ways. Big Iron approaches do not handle heterogeneous data well. They tend to be designed around a very consistent data model such as gridded

---

14 See the Data Citation Guidelines from the Federation of Earth Science Information Partners at http://bit.ly/data_citation.

15 See brief discussion and slide set at http://scholarlykitchen.sspnet.org/2009/09/24/john-wilbanks-its-the-customer-not-the-container/.

fields. The systems are overly reliant on automation, tend to assume a certain type of use, and are generally not very adaptive. More critically, Big Iron systems tend to underplay the need for preservation (although this is beginning to change). Map Making is perhaps the most limiting as it shares some of the limitations of both Data Publication and Big Iron approaches. It can be very useful for integrating data over geographic space, but it typically does not adequately handle temporally dynamic data well. The Linked Data approach is still fairly new and has not really considered the full data life cycle. The approach suffers from poor versioning and auditability and is generally not very human friendly. More troubling, it typically results in a change from the original data model.

Overall, while each of the metaphors we present have their strengths, none of the current data management paradigms fully meet all the basic criteria outlined by the *Association of Research Libraries' Agenda for Developing E-Science in Research Libraries* (ARL, 2007) or other community guidance (Arzberger et al., 2004; Doorn and Tjalsma, 2007, Parsons et al. 2011). To illustrate this shortcoming, we identified six critical attributes of an effective approach, based on the aforementioned existing guidance and our own experience:

- Established trust (of data, systems, and people)
- Data are discoverable.
- Data are preserved.
- Data are accessible to humans and machines.
- Data are usable, including some level of understandability.
- Effective, distributed governance of the data system.

These are by no means all the desirable attributes, but we do not think that any of the current models fully address even these basics. For example, Data Publication does not handle discovery well beyond the publication and lacks a coordinated governance mechanism. Big Iron, Map Making, and Linked Data on the other hand all lack robust preservation approaches, and the usability of data can be highly variable. Of the four models, Data Publication may actually perform the best against these criteria, but we believe we need to consider alternatives.

5 ALTERNATIVE PARADIGMS AND METAPHORS

While metaphors can limit our thinking, they can also help us conceive alternatives. It is not sufficient to say that the Data Publication (or any other) metaphor is limiting. We need to recognize other existing metaphors and actively seek new metaphors that complement each other and help us conceive of all aspects of the e-science data challenge. We believe this needs to be an ongoing conversation in the community, but we offer some initial ideas here.

We see two high-level metaphors that go beyond the data management enterprise and consider the larger whole of science communication: the large concepts of infrastructures and ecosystems. The data infrastructure metaphor is well established. The geospatial data community has referred to national and global "spatial data infrastructures" since at least the early 1990s (NRC, 1993). More recently, the concept has been codified, in the US at least, as "cyberinfrastructure" by the NSF "Blue Ribbon Advisory Panel on Cyberinfrastructure" (Atkins et al., 2003). Considering an entire infrastructure helps us recognize the scale of our endeavor—it truly needs to reach across the entire scientific enterprise. But in many ways the concept of a data or information infrastructure is still being defined. More critically, conceptions of infrastructure too often ignore or underplay its socio-cultural elements (Bowker et al., 2010). As such, we find metaphors typically drawn from physical infrastructure concepts, like railways and electrical utilities, while useful can also be too simplistic. Indeed, infrastructure can be very difficult to study because it typically exists in the background—invisible and taken for granted (Star and Ruhleder, 1996).

More recently we have become intrigued by the metaphor of a "data ecosystem – the people and technologies collecting, handling, and using the data and the *interactions* between them" (Parsons et al., 2011, p. 557 original emphasis). We appreciate the extension of the common data life-cycle metaphor and the focus on interactions and relationships. The ecosystem concept emphasizes adaptation, evolution, and diversity rather than a centralized command and control structure. Yet while the obvious metaphors like the seeding and growth of an idea or the evolution of a technology give us a holistic view, they are sometimes lacking in specifics. What is the equivalent of publishing a data set in an ecosystem? Data sprouting, growth, birth, release, graduation, …? None of these are completely clear or truly resonate as a complete or compelling solution.

Perhaps we should not try and find an overarching metaphor for the whole data management process. Perhaps that misses the point. Historically, in literary publication, each publisher filled the multiple roles of archiving, registration, dissemination, and certification of the paper. Priem and Hemminger (2011) argue that this model, with thousands of independent publishers each filling all roles, resists innovation and makes it difficult to change any one aspect of the system. They argue that we need to "decouple" the journal to create a "Web-like environment of loosely joined pieces—a marketplace of tools that, like the Web, evolves quickly in response to new technologies and users' needs" (p.1). We welcome this idea and suggest that similarly, we need to start

decoupling or disaggregating the functions of data stewardship to consider each function fully. In a modern information ecosystem it is unreasonable to assume one entity would do everything. It is necessary take multiple approaches to manage different types of data. We need to consider all the available paradigms and consider the various functions of data stewardship individually in their own right and as a whole. *We need not one metaphor but many.*

For example, within a portion of the research life cycle, a core function may be data preservation and therein the obvious and common metaphor is the formal, curated *archive*. Data must also be made available for use. Here we see the need for rapid, carefully versioned and described *releases* (akin to software) rather than fixed publications. Potential data users need some form of assurance or certification that the data are useful and of certain quality, but rather than thinking of a formalized peer-review system with appointed reviewers, we suggest *tracking* data through a *democracy* or a *marketplace* to see how and where data are used. We revisit Raymond's 1999 classic *The Cathedral and the Bazaar.* Considering a marketplace or bazaar also illustrates the need for specialist *shopkeepers*, *mediators*, or *brokers*, who help users understand and make effective use of the data. Indeed, bazaars evolve and thrive on the needs of customers as much as the actual goods and services they initially provide. These are nascent ideas, but we hope they spark some conversation.

In closing, we note that metaphors are prevalent and powerful across the research enterprise. They can help us see new aspects of a problem, but they also create frames of thinking that can limit our perspective and perceived choices. We suggest that, at the present state of evolution toward data as a first class citizen, it is important not to be hidebound by the idea of 'data publication' or any *one* metaphor. We need to disaggregate the roles of data stewardship and reassemble them in new ways. We must be open-minded and consider many metaphors, paradigms, and ways of knowing to fully address the data science challenges of the 21st Century.

6 REFERENCES

ARL Joint Task Force on Library Support for E-Science. (2007). *Agenda for Developing E-Science in Research Libraries.* http://www.arl.org/bm~doc/ARLESciencefinal.pdf. Accessed 27 Feberuary 2011.

Arzberger, P, P Schroeder, A Beaulieu, G Bowker, K Casey, L Laaksonen, D Moorman, P Uhlir, and P Wouters. (2004). Science and government: An international framework to promote access to data. *Science. 303*(5665): 1777-78.

Atkins, DE, KK Droegemeier, SI Feldman, H Garcia-Molina, ML Klein, DG Messerschmitt, P Messina, JP Ostriker, and MH. Wright. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* http://www.nsf.gov/od/oci/reports/toc.jsp. Accessed 26 November 2011.

Bowker, GC, K Baker, F Millerand, and D Ribes. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. *International Handbook of Internet Research* Springer Science+Business Media. pp. 97-117.

Callaghan, S, F Hewer, S Pepler, P Hardaker, and A Gadian. (2009). Overlay journals and data publishing in the meteorological sciences. *Ariadne.* (60). http://www.ariadne.ac.uk/issue60/callaghan-et-al/

Costello, MJ. (2009). Motivating online publication of data. *Bioscience. 59*(5):418-427. http://dx.doi.org/10.1525/bio.2009.59.5.9

Doorn, P, and H Tjalsma. (2007). Introduction: archiving research data. *Archival Science. 7*(1):1-20.

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience.* New York: Harper & Row.

Klump, J, R Bertelmann, J Brase, M Diepenbroek, H Grobe, H Höck, M Lautenschlager, U Schindler, I Sens, and J Wächter. (2006). Data publication in the open access initiative. *Data Science Journal. 5*:79-83. http://dx.doi.org/10.2481/dsj.5.79

Lakoff, G. (2008). *The Political Mind: Why You Can't Understand 21st-century Politics with an 18th-century Brain.* New York: Penguin Group.

Lakoff, G, and M Johnson. (1980). *Metaphors We Live By.* Chicago: The University of Chicago Press.

McMahon, M. (1996). Overview of the Planetary Data System. *Planetary and Space Science. 44*(1):3-12. http://dx.doi.org/10.1016/0032-0633(95)00101-8

NRC (National Research Council). (1993). *Toward a Coordinated Spatial Data Infrastructure for the Nation.* Washington, DC: National Acadamies Press. 192 pp.

Parsons, MA, R Duerr, and JB Minster. (2010). Data citation and peer-review. *Eos, Transactions of the American Geophysical Union. 91*(34):297-98. http://dx.doi.org/10.1029/2010EO340001

Parsons, MA, Ø Godøy, E LeDrew, TF de Bruin, B Danis, S Tomlinson, and D Carlson. (2011). A conceptual framework for managing very diverse data for complex interdisciplinary science. *Journal of Information Science. 37*(6):555-569. http://dx.doi.org/10.1177/0165551511412705

Penev, L, M Sharkey, T Erwin, S Van Noort, M Buffington, K Seltmann, N Johnson, M Taylor, C Thompson, and M Dallwitz. (2009). Data publication and dissemination of interactive keys under the open access model. *ZooKeys. 21*:1-17.

Pfeiffenberger, H, and D Carlson. (2011). "Earth System Science Data" (ESSD) — A peer reviewed journal for publication of data. *D-Lib Magazine. 17*. http://dx.doi.org/10.1045/january2011-pfeiffenberger.

Priem, J, and BM Hemminger. (2011). Decoupling the scholarly journal. *Frontiers in Computational Neuroscience.* . http://www.frontiersin.org/computational_neuroscience/abstract/14455 Accessed 28 November 2011.

Raymond, ES. (1999). *The Cathedral and the Bazaar: musings on Linux and open source by an accidental revolutionary.* Cambridge, MA, USA: O'Reilly.

Star, SL, and K Ruhleder. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research. 7*(1):111.

Wright, A. (2009). Exploring a 'Deep Web' that Google can’t grasp. *The New York Times*. 22 February 2009. http://www.nytimes.com/2009/02/23/technology/internet/23search.html. Accessed 12 December 2011.