Note: This is a preprint version of the final revised paper available in the *Data Science Journal* at http://dx.doi.org/10.2481/dsj.WDS-042.

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. Opportunities emerge to re-visit long-standing approaches to managing data and to critically examine new capabilities. We describe the cognitive importance of metaphor. We describe several metaphors for managing, sharing, and stewarding data and examine their strengths and weaknesses. We particularly question the applicability of a "publication" approach to making data broadly available. 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. We close with proposed research questions and a call for continued discussion.

Keywords: data publication, data system design, data citation, semantic Web, data quality, data preservation, cyberinfrastructure.

1 INTRODUCTION

Data authors and stewards rightfully seek recognition for the intellectual effort they invest in creating a good data set. At the same time, we assert that good data sets should be respected and handled like first class scientific objects, i.e. the unambiguously identified subject of formal discourse. As a result, people look to scholarly publication—a well-established, scientific process—as a possible analog for sharing and preserving data. Data "publication" is becoming a metaphor of choice to describe the desired, rigorous, data stewardship approach that creates and curates data as first class objects (Costello, 2009; Klump et al., 2006; Lawrence et al., 2011). The emerging International Council for Science World Data System (WDS)[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 science needs to adopt the robust principles of "publication" rather than informal "sharing" as a more effective way to ensure data openness and availability. While we strongly support these efforts to recognize data providers and to improve and professionalize data science, 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.

This essay grew out of several conversations between the authors. It began with a "tweet" by Fox at the 2010 CODATA meeting that first questioned the term "publication".[3] Fox was being deliberately provocative; Parsons is easily provoked; and so the conversations began. About a year later, Parsons was invited to co-convene and speak at a session entitled simply "Data Publication" at the inaugural conference of the WDS. It was a bold move by the WDS to openly question their stated data publication paradigm, and it forced us, the authors, to begin to refine our thoughts beyond casual conversation. The presentation was politely received and generated some interest, enough for us to decide to go ahead and write an essay. We "published" the first draft of our essay on an open blog[4] in December 2011 and asked for community comment. We were overwhelmed by the response. Through comments on the blog, posts on other blogs, and direct e-mail, we received some 70 pages of review comments from more than two-dozen individuals over about six weeks. The reviews ranged from a few casual comments to very thorough and detailed critiques. The conversation was very stimulating, convincing us that it needs to continue more formally.

---

[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
[3] "okay, I'll say it. The *term* data 'publication' bothers me more and more. Am leaning toward data release and *maybe* review, #CODATA2010" (@taswegian; posted 25 Oct. 2010).
[4] http://mp-datamatters.blogspot.com/

It is now almost a year later. The world of data and informatics continues to evolve rapidly. Just in the time since we released the first draft of this essay, Thomson Reuters announced a new data citation index, several new data journals launched, the Research Councils of the UK announced a new policy on open access to research outputs,[5] and US President Obama highlighted data management as a critical new job skill for the 21st century in his State of the Union address. In this rapidly changing environment with growing expectations and challenges facing data science, we believe it is critical to be as adaptive as possible. We must do what we can to avoid the negative "path dependence" that can inhibit adaptive evolution of a robust information infrastructure (Edwards et al., 2007). In that light, we present this revised essay. It is much improved by the many cogent comments received, but we are sure we will continue to provoke some disagreement. We remain convinced of our core message that no one metaphor or worldview is sufficient to adequately conceive the entire data stewardship and informatics enterprise. All metaphors have their strengths and weaknesses, their advantages and risks, their clarification and obfuscation. Our position is that this is especially true of Data Publication (Note we deliberately capitalize Data Publication here forward to reflect its status as a recognized metaphor and data management paradigm). As the most established metaphor and narrative, Data Publication may have both the greatest strengths and the greatest weaknesses. If we do not think critically of all our metaphors, we may see only the opportunities and not the risks. Correspondingly, if we do not seek new metaphors, we may miss new opportunities.

With this revised essay, we seek to further stimulate and advance the dialog among data scientists in a way that considers multiple worldviews and helps us conceptualize diverse approaches to science data stewardship and informatics. In Section 2 we discuss briefly the critical importance of metaphor in human communication and cognition. We then explore some existing worldviews and metaphors in Section 3 and examine their strengths and weaknesses in Section 4. We examine some alternative worldviews in Section 5 and conclude in Section 6 with a call to action based on a proposed research agenda.

2 THE IMPORTANCE OF METAPHOR AND FRAMING

At a simple level, a metaphor is a figure of speech where a word or phrase is applied to something for which it is not literally applicable. It is something symbolic or representative of something else. But it is much more than that. Metaphor is central to how people communicate and even to how we think and react to the world around us. 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 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).*

Understanding this "conceptual system" is central to cognitive science (Lakoff and Johnson, 1980a) and the system is increasingly seen the to be fundamentally metaphorical in character (Lakoff, 1993). Lakoff and Johnson (1980) explore some of our most basic metaphors (argument is war, happy is up, sad is down, time is money, love is many things) and show how metaphors help define our modes of thought or worldviews. They show how metaphors help us create the complex narratives we use to understand our physical and conceptual experience. These complex narratives are made up of smaller, very simple narratives called "frames" or "scripts". Framing and frame analysis are often used in knowledge representation, social theory, media studies, and psychology with much of the work stemming from Erving Goffman (1974).

These frames present a set of roles and relationships between them like characters in a play. They also help us define our terms and make sense of language, because words are defined relative to a conceptual frame. The word "sell" does not make sense without some understanding of a commercial transaction and some of the other roles and terms involved like "buyer," "money," and "cost". Furthermore, by mentioning only one of these concepts like "buy" or "sell", the whole commercial transaction scenario is evoked or "activated" in the mind (Fillmore, 1976). Similarly, we can see how particular roles and our subtle understanding of them emerge from the publication metaphor with terms like "author," "editor", "publisher," "reviewer," and "librarian". We do not define these terms and let readers see what definitions emerge from their own conceptual frame.

Lakoff (2008) further argues that framing is critical to human cognition. The neural circuitry to create a frame is relatively simple and our brain essentially uses framing as a sort of cognitive processing shortcut. If things are

5 http://www.rcuk.ac.uk/research/Pages/outputs.aspx

understood in the context of a frame, much is already unconsciously understood and need not be consciously processed. We know what to expect. Indeed, the vast majority of human thought is not conscious reflective thought but unconscious reflexive thought. Lakoff (2008) explores the role of this unconscious reflexive thought in politics and morality. While he arguably carries a political bias or agenda into his work, he clearly shows how language, metaphor, and framing play critical roles in any social enterprise. He summarizes the power of language well on page 14:

> *Language is at once a surface phenomenon and a source of power. It is a means of expressing, communicating, accessing, and even shaping thought. Words are defined relative to frames and conceptual metaphors. Language 'fits reality' to the extent that it fits our body-and-brain based understanding of that reality. […] Language gets its power because it is defined relative to frames, prototypes, metaphor, narratives, images and emotions. Part of its power comes from unconscious aspects: we are not consciously aware of all that it evokes in us, but it is there, hidden, always at work. If we hear the same language over and over, we will think more and more in terms of the frames and metaphors activated by that language.*

This last point is critical. Thinking in frames is natural and unavoidable. Frames provide a structure for cognition and understanding, but they also, by their nature, present 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.

We admire and are amused that Lakoff and Johnson turn their own logic back on their own discipline. The concluding sentence of Lakoff and Johnson (1980a) states: "The moral: Cognitive Science needs to be aware of its metaphors, to be concerned with what they hide, and to be open to alternative metaphors-even if they are inconsistent with the current favorites." We seek to apply that same moral to our discipline of data science. In subsequent sections we examine Data Publication and other metaphors and worldviews around data science and stewardship. 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 WORLDVIEWS AND ASSOCIATED METAPHORS

Currently, we see (at least) five active worldviews on how to most effectively steward and share data in Earth system science. These worldviews vary in their maturity. They, and their corresponding data management approaches, are not mutually exclusive. It is common for data scientists to see themselves as actors in several narratives. Nonetheless, there is usually a dominating perspective that defines particular data management approaches. As Baker and Bowker (2007, p. 129) state "No institution is ever total, nor is any system totally closed. However, it remains true that there are modes of remembering that have very little to do with consciousness on the one hand or formal recording keeping on the other." This is understandable. As Bruce Barkstrom (2012, personal communication) points out, the data management approaches and their worldviews come from different communities and cultures and are geared toward different users and different data types. There is nothing inherently good or bad about any one approach or worldview unless it is not aligned with community views. Our intent here is not to simply criticize particular systems or methods but rather to unpack our assumptions and understand our frames of thinking and underlying values. Furthermore, we present an admittedly cursory and even stereotypical assessment of the different worldviews. It was clear that our initial draft of this essay offended data scientists from all perspectives with its blithe analysis of the worldviews. As professional data scientists, we do not trivialize the complexity of our discipline, but we do seek to understand how we frame and conceptualize our challenges and strategies. So we must examine some of the stereotypes in which we operate. Broad conceptual understanding can sometimes be at odds with technical precision, but only through understanding our underlying modes of thought, even at a crude level, can we hope to expand and adapt those modes of thought to address the dynamic, complex challenges of data science.

With those considerations in mind, we examine five active worldviews on science data that we name with five metaphors: Data Publication, Big Iron, Science Support, Map Making, and Linked Data. We discuss their attributes in turn below and summarize them in Table 1.

The Data Publication approach seeks to be analogous to scholarly literature publication, and generally emerges from the culture of academic research and scholarly communication. Its focus is often on "research collections" (NSB, 2005) where data are extremely diverse in form and content but tend to be relatively small in size. Data Publication seeks to define discrete, well-described data sets, ideally with a certain level of quality assurance or peer-review. The data sets often provide the basis for figures and tables 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 to organize the data. Further, the standards used across different data systems are fairly high level, e.g. exchange of Dublin Core metadata using protocols such as OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting). Data citation has been an important standards emphasis in Data Publication. 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®[6] system in Germany. A very explicit form of Data Publication is seen in newly emerging data journals such as *Earth System Science Data*. As mentioned, Data Publication is the most mature of the metaphors in play. Costello (2009) and especially Lawrence et al. (2011) provide much more rigorous descriptions of the paradigm, but it is important to recognize that they describe a desired, not fully realized situation. For example, Lawrence describes a data peer review scheme that is not yet fully or broadly adopted. Furthermore, despite these efforts, there is still incomplete agreement on the definitions and assumptions that arise from the frames of Data Publication.

The Big Iron approach is akin to industrial production and often comes from more of an engineering culture found with large-scale data producers such as NASA. Big Iron typically deals with massive volumes of data that are relatively homogenous and well defined but highly dynamic and with high throughput. The Big Iron itself is a large, sophisticated, well-controlled, technical infrastructure potentially involving supercomputing centers, dedicated networks, substantial budgets, and specialized interfaces. It may also be a simpler collection of relatively common commodity software and hardware, but the focus is still on large volumes, reducing actual data transfer, computational scaling, etc. Historically, less emphasis was placed on archiving, but it 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 and organizational schemes. Significant emphasis is placed on consistent, rich, data formats and data production concerns such as careful versioning. Examples of the Big Iron approach include the European Space Agency's Science Archives[7] or NASA's Earth Observing System Data and Information System (EOSDIS)[8]. To be fair, nobody usually refers to such data systems as Big Iron. We use the term to be illustrative of a large-scale, production-oriented mode of thinking. "Big data" may be the more common term describing this worldview. It is also worth considering cultural differences across different production paradigms. For example, there are very different concerns around latency, data quality, spatial and temporal resolution, and other issues when addressing operational weather forecasting as opposed to long-range climate analysis, even though the data streams may ostensibly be very similar.

Science Support is viewed as an embedded, operational support structure typically associated with a research base or lab. In environmental sciences, the focus is often on place-based research such as is conducted at long term research bases or sites. Data management is seen as a component or function of the broader "science support" infrastructure of the lab or the project. Science support for a lab is defined differently in different contexts and tends to be very broadly conceived. It may include many things such as facilities management, field logistics, administrative support, systems administration, equipment development, etc. Often, there is no clear line between what is the science and what is the support. For example, data collectors at a field site may be lead investigators on a given research project or lab technicians supporting many projects. In this context, data tend to be the research collections similar to those in the Data Publication metaphor but there is often a focus on creating community collections by characterizing important fundamental processes or particular representative conditions over time. The data are organized in myriad ways, usually geared towards a specific set of intended uses and local reuse in conjunction with other local data. The historical Long Term Ecological Research (LTER) network is a good example of this approach where local science support functions remain constant over time even while a broader, network-level data system is added. Baker and Millerand (2010) describe the process of how the LTER information systems developed both locally and nationally and illustrate the Science Support perspective, where data management is both integrated into the science process, yet also partially outside the process in a support role. (Lynn Yarmey, 2012 personal communication)

Map Making is most readily seen in so-called spatial data infrastructures (de Sherbinin and Chen, 2005; FGDC, 1997; NRC, 1993) and their associated geographic information systems (GIS). The perspective emerges naturally from land use and survey agencies that have been creating and working with maps for centuries. Map Making shares attributes of the other paradigms. Maps are certainly used in Science Support and Map Making could be seen as a subset of Data Publication, but here the analogous publication is a map or an atlas rather than a journal article. On the other hand, national and international spatial data infrastructures often seek to operate the more centrally governed, standardized model of Big Iron. Here, however, the important metaphor is it is not the final product or the production process but rather the representation of the data and their associated science questions

6 http://pangea.de
7 http://www.sciops.esa.int/index.php?project=SAT&page=index
8 http://eospso.gsfc.nasa.gov/

through a geographical perspective, notably the map[9]. Data in this approach tend to be more fixed in time, i.e. they are more geared toward describing geospatial features rather than dynamic 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 grid-) based model with an associated (geo)database. Map Making has been especially successful in defining standards around things like coordinate reference systems, map projections, and map transfer protocols. Major examples of map-based systems include the INfrastructure for SPatial InfoRmation in Europe (INSPIRE),[10] OneGeology.org, and Geodata.gov in the US.

Linked Data is based on computer science concepts of the "Web of data" and relies on the underlying design principles behind the Semantic Web,[11] especially as described by Tim Berners-Lee.[12] The paradigm emerges from the culture of World Wide Web development, including non-science and commercial enterprises. The "data" in Linked Data are defined extremely broadly and are envisioned as small, independent bits with specific names (URIs) interconnected through defined semantic relationships (predicates) using model and language standards (e.g. the Resource Description Framework, RDF). 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 this approach is that it uses a graph model not a hierarchical or relational model of data organization. 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. Correspondingly, the standards focus, especially in the sciences, has been on the development of formal ontologies. 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]. In many ways, Linked Data is not as comprehensive a worldview as some of the others. Arguably, it may be seen as a set of techniques or tools used within a broader context such as Data Publication (Bechhofer et al., 2011) that can potentially be accessible by a broad range of data producers, e.g. an individual researcher with programming skills. Again, however, we note the focus of the metaphor. As with Map Making, the metaphorical emphasis is not on the product or the process but the data representation; this time not as a geospatial map but as a network or graph.

4 PROS AND CONS OF THE CURRENT WORLDVIEWS

Each of the worldviews described above have their strengths and weaknesses for understanding and addressing the challenges of data science. Nominally, the data management approaches that emerge from the different worldviews are fully capable of stewarding data according to defined best practice, but the varying perspectives and metaphors focus on different stages of the data life cycle, different audiences, and different challenges. We do not believe that any of the current data management paradigms fully meet all the basic criteria outlined by the ISO standard *Open Archival Information System Reference Model* (ISO, 2003), the broader guidance of the *Association of Research Libraries' Agenda for Developing E-Science in Research Libraries* (ARL, 2007) or other general community guidance (Arzberger et al., 2004; Doorn and Tjalsma, 2007, Parsons et al. 2011).

We identified seven critical attributes of an effective, comprehensive data stewardship approach, based on the aforementioned guidance and our own worldview and values:

- Established trust (of data, systems, and people).
- Data are discoverable.
- Data are preserved.
- Data are ethically open and readily accessible to humans and machines.
- Data are usable, including some level of understandability.
- Effective, distributed governance of the data system.
- Reasonable credit and accountability for data collection, creation, and curation.

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. In this section, we provide a cursory, subjective assessment of how the different worldviews address these criteria. We examine each worldview briefly and then discuss Data Publication in more detail.

9 A broader conception of this metaphor might be "Sense making". Areas like biological taxonomy and structural chemistry have different constructs for making sense of their information. Maps, however, are especially powerful metaphors and representational tools. Critical geographers have long shown how maps can be tools to assert power and authority and may be viewed as a product of authorial intent rather than objective data presentation (Harley, 1989; Koch, 2004). This is somewhat tangential, but it is another illustration of the power of metaphor in how we conceive of and represent data and their relation to broader conceptions of reality.

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

11 http://linkeddata.org

12 http://www.w3.org/DesignIssues/LinkedData