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: parsons.mark@gmail.com
[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.

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

[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/

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 & Johnson, 1980a), and the system is increasingly seen 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, such as "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 such as "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 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

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

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 record 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 station or lab. In environmental sciences, the focus is often on place-based research such as is conducted at long term research stations 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 & 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 the 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 through a geographical perspective, notably the map[9]. Data in this approach tend to be more

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

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 & 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.

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

Table 1. Summary of attributes (rows) of some major data management related metaphors (columns)

	Data Publication	*Big Iron*	*Science Support*	*Map Making*	*Linked Data*
Analog	scholarly publication	industrial production	artisanal, task-specific production	cartography	World Wide Web creation
Data characteristics	small volume and diverse form, scale, and topics	high volume and more homogenous in form	small and diverse	geospatial features and attributes	many disparate and named entities
Data organizational models	hierarchical or relational	hierarchical	geospatial, hierarchical, and relational	geospatial and relational	linked graph
Primary focus	data quality, certification, and preservation	throughput and manageable access	data synthesis and reproducibility,	map-based visualization and intercomparison	interoperability and interconnection
Standards emphasis	data citation	data formats, versioning	local processes	coordinate reference systems, spatial transforms	ontologies
Examples in science	PANGEA, university repositories	EOSDIS,	LTER	INSPIRE, Geodata.gov	IODP, MyGrid, Linked Open Government Data
Metaphorical terminology	data author, publisher, data citation	data producer, processing level, version release	data collector, support staff	data source, feature, layer	data provider, name, link, resource
Cultural context	libraries and university research groups (e.g. NSF science directorates)	system engineering and project management (e.g. NASA, DoD)	place-based research (e.g. focused institutes, NSF)	land use and management (e.g. USGS, local agencies)	computer science and commercial applications (e.g. NSF CISE and W3C)

Data Publication builds from the familiar and conceptually simple model of scholarly literature publication. "Publishers" are distributed and can act autonomously or in concert. Published data are usually well cared for and often carry assertions that data are of high, or at least well-described, quality. The approach builds from the norms of scientific research and can be well trusted, but there is a corresponding lack of strong governance across systems. There is also little emphasis on data discovery and interoperability across systems. Data are often presented as they were created without explicit considerations of data integration or significant reuse beyond the scientific community. The approach works well for relatively stable data sets, but systems can be difficult to automate and do not always scale well. The attention is on preservation and formal recognized scholarly contribution with less attention to "big data" issues such as latency, rapid versioning and reprocessing, and computational demands.

Big Iron approaches tend to be highly automated and hence well suited to formal audits and reprocessing. By design, the systems handle large volumes and streaming data well and can provide very short latency when necessary. The systems usually have defined governance mechanisms with some sort of controlling authority or policy-level certification. On the other hand, Big Iron systems do not handle heterogeneous data well. They tend to be designed around a very consistent data model such as gridded fields. The systems are sometimes overly reliant on automation and tend to assume a certain type of use. Roles are not always well defined, and systems

are generally not very adaptive. More critically, Big Iron systems tend to underplay the need for preservation (although this is beginning to change). In general, there is more of an engineering focus than science focus, which is both a strength and a weakness in its own right.

Science Support is inherently localized in its focus. Systems may be well established and very useful for the designated community they are supporting. An important strength is the focus on data integration for that community, but data and systems are often not designed for use or access beyond the community. Governance structures across sites are only emerging and completely lacking in some disciplines. Data preservation is variable and largely dependent on the knowledge and interest of local science support staff. In contrast to Big Iron, there is a very strong science focus that makes the data very useful for its intended purpose, but systems are more ad hoc and may lack design and preservation rigor.

Map Making is obviously well suited to and correspondingly limited to geospatial representation of data. It can be very useful for integrating data over geographic space, but it typically does not handle temporally dynamic data well. There is an established history of geospatial governance mechanisms with variable success. Of note is the emergent success of the Open Geospatial Consortium in the last decade at establishing widely adopted standards of interoperability. A history of proprietary systems and data formats has hindered data preservation, but that is rapidly improving. Nevertheless, core aspects of data stewardship, such as preservation, access, and trust, largely depend on the institutional context where the Map Making metaphor is applied.

The Linked Data approach is still fairly new and has not really considered the full data life cycle. Its primary strength is that it is built on a simple, highly scalable model that allows for broad data dissemination and very flexible machine processing. There is no *a priori* assumption of how data are to be used, and the model handles extremely diverse data well. In a sense, the approach is data model independent (unlike Map Making for example), but it typically achieves this through a change from the original data model (to RDF). This creates issues for preservation. Indeed preservation is largely ignored in the Linked Data worldview. The approach suffers from poor versioning, auditability, and accountability, and it is generally not very human friendly. It also lacks a controlling central authority; this allows great flexibility but limits preservation and accountability.

We summarize our simplistic analysis in Table 2. While we recognize that most all of our assertions can be countered, we trust the reader can recognize some of the strengths and weaknesses we describe in the systems they are familiar with. More importantly, we have illustrated that by focusing on limited aspects or perspectives of a problem, one can often miss other important issues. None of the metaphors are complete, and most data scientists operate in spaces that could be characterized by several of the worldviews. Nonetheless, many might argue that Data Publication is the most mature, well-understood worldview; therefore by better defining and refining Data Publication practice we best serve data science and stewardship. We do not believe that is a complete or wise approach.

Despite well-considered descriptions of formal Data Publication (Lawrence et al. 2011, Costello, 2009), it was clear in the review of this essay that there is no widely understood and accepted definition of what exactly Data Publication means. It was equally clear that "publication" carries many, differing, implicit assumptions that may not be true. A central argument for Data Publication is that the metaphor resonates with researchers. They understand their role in the process, it is said. Yet researchers are not knowledgeable of the refined definition of Data Publication. We argue that this creates false understanding; that the frames and roles of Data Publication create false assumptions that what is true for scholarly literature publication applies to data publication. Furthermore, the metaphor may be too restrictive and not allow researchers *or* data scientists to fully understand and adapt to the modern challenges of data driven science.

To illustrate our concerns we examine three frames that emerge from Data Publication that can create false assumptions and misguided approaches. First, peer review. Data Publication implies some level of imprimatur (Callaghan et al., 2009), and 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). Indeed, de Waard et al. (2006, 2008) demonstrate a rhetorical model of scientific publication that indicates that peer-review of data cannot truly parallel peer review of literature. The model makes the important distinction between the article, which is designed to persuade (Kuhn, 1996; Latour, 1987), and the data, which are intended to be simple fact.

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). The *Earth System Science Data* journal and other emerging data journals and overlay journals combine the review of the data set with review of a more conventional article that is closely linked to the data (e.g., Callaghan et al., 2009;

Pfeiffenberger and Carlson, 2011). Lawrence et al. (2011) examine peer-review in depth and provide a useful data review checklist. These are valuable contributions, but we still find the peer-review frame to be limiting. The review of data is fundamentally different than the review of an argument in a paper, and the different approaches have different meaning and levels of (implied) certification. Traditional human refereeing is appropriate for certain major data sets, but it is too slow, and it will not scale to handle the growing deluge of data. We need to consider other models of what is essentially a quality assurance/quality control process and automate where possible. Thinking outside the peer-review frame can help us conceive of these models. For example, tracking how a data set is used over time may be more revealing of its quality and fitness for use than the formal opinion of two or three disciplinary experts. The quality of data depends on the application. Unlike with literature, there may still be value in releasing "poor quality" data because they may be useful for certain applications or because broad exposure of the data may lead to creative solutions to their prior limitations. Too often we have seen purported insufficient data quality used as an excuse to restrict data access. Data quality is a critical and difficult issue fundamentally different from the intellectual merit of a scholarly article. We should not let the Data Publication metaphor limit our thinking of how data quality can be addressed.

Table 2. Summary of strengths and weaknesses of the data management worldviews.

	Data Publication	*Big Iron*	*Science Support*	*Map Making*	*Linked Data*
Trust	good	moderate	good	moderate	Poor
Discovery	poor	moderate	poor	moderate	Good
Preservation	good	poor	variable	poor	Poor
Access	moderate	moderate	poor/moderate	good	Good
Usability	moderate	moderate	good	moderate/good	Moderate
Governance	poor	good	poor	moderate	Poor
Credit and accountability	good	poor/ moderate	variable	poor/moderate	Variable

The second frame of concern is the closely related concept of data citation. We strongly support the data citation concept, but we feel that the publication metaphor has created some false expectations around it. Data citation might be better termed data reference. The primary purpose is to aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study (Ball & Duke, 2012; ESIP, 2012). This means that data need to be identified and located, ideally with a persistent identifier, as soon as they are available for use by anyone other than the original creator. Data release often occurs in stages. Data may be initially shared with a small team, later released to a broader group within the discipline sometimes with caveats, and then finally the data are released to the public, i.e., "published". Technically, data need to be precisely referenced if they are used in a study at any time during those stages, but typically a DOI is not assigned until the final publishing stage. The DOI is meant to assert a sort of imprimatur. This does not seem an appropriate use of the DOI. We understand and appreciate the desire for an imprimatur, but find it odd that it be conveyed with a simple registration of an identifier. Identifiers and locators are necessary before formal, reviewed publication and there is nothing inherent in DOIs that ensures persistence of the data. It still relies on human due diligence (Duerr et al., 2011). We feel that the emphasis on "publication" underplays the often-broad use and evolution of a data set long before it may be formally published and can be making misleading assertions about the meaning and purpose of identifiers.

Another important aspect of data citation is the desire to provide fair credit for the intellectual and technical effort that goes into creating a good data set. Indeed, data citation is often seen as an incentive for researchers to release their data. Unfortunately, in our experience, scientists do not especially welcome data citation. Some like the idea; some see it as diluting citations to their paper. Also, some funding agencies question the idea of recognizing individuals as data authors. We do not necessarily agree with these detractors, but we see a problem in that the Data Publication metaphor has led us to conflate many issues into data citation, including reference,

quality assertion, credit, and data discovery. This has only made the precise identification and reference issue more difficult. We need to separate the concerns and come at them from different directions.

Our third concern is with the close association of Data Publication with copyright and restricted-access literature. 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. If citation in publications is the primary means of identifying data, an unintended side effect may be to actually limit data access and discovery because of the restrictions on the article. We end up reinforcing the hidden "deep web of data" (Wright, 2009). And while data publishers are often strong advocates of open access, some argue that Data Publication necessarily includes licensing of the data set and mandating conditions of use (Klump et al., 2006). We much prefer the norms-based, copyright-free approach of an information commons as adopted by *Earth System Science Data.* Too often, we find that the Data Publication perspective is, as John Wilbanks (2009, personal communication[14]) 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. It also implies a social contract that is not applicable for data. In traditional scholarly publishing authors relinquish certain rights and go through certain processes in exchange for receiving professional credit. The social contract for data could and probably should be much different. For example, semi-blind review may not be appropriate and rights around data are fundamentally different than copyrights on creative works. The focus on the container misses how 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. Openness and flexibility is essential to maximizing value of data, and restricted data discovery and access still remain major inhibitors of data science endeavors (ICSU, 2011). Ironically, while those who advocate data publishing tend to be some of the strongest advocates for open data, we find that the Data Publication container can restrict access, interoperability, and creative use. All the metaphors need to be explored for approaches to *unlocking* the data in the "deep web".

We have been severely critical of the Data Publication worldview. We do not suggest a wholesale rejection of the metaphor but rather recognition of how it can sometimes restrict and even misguide our thinking. Scholarly publication is in the process of re-examining its own model (see for example the European Framework LiquidPub project[15]), and we should be open to learning from that process, but we should not assume it will provide a good analog for data. None of the current worldviews described completely address the full needs of robust data stewardship. Data Publication, in particular, has major strengths and is the most evolved, but it may also be the most misleading. Data Publication efforts should certainly continue, but we must remain open to other alternatives. It is critical to avoid the stifling "path dependence" than can inhibit the development of a useful and adaptive sociotechnical infrastructure (Edwards et al., 2007). We must actively challenge our thinking and seek out other worldviews and metaphors.

5 ALTERNATIVE WORLDVIEWS AND METAPHORS

While metaphors can limit our thinking, they can also help us conceive alternatives. To say that the Data Publication or any other metaphor is limiting is insufficient. 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 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 NSF "Blue Ribbon Advisory Panel on Cyberinfrastructure" has codified the concept, in the US at least, as "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 socio-cultural elements (Bowker et al., 2010). We, therefore find metaphors typically drawn from physical infrastructure concepts like railways and electrical utilities useful but also too simplistic. Indeed, infrastructure can be very difficult to study because it typically exists in the background—invisible and taken for granted (Star & Ruhleder, 1996). We support the developing field of infrastructure studies (Bowker et al., 2010), but despite the rich, sophisticated, and holistic examinations of this literature, data practitioners still tend to view infrastructure as a physical construct rather than the body of relationships defined by Star and Ruhleder (1996).

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

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). We appreciate the extension of the common data life cycle metaphor and the focus on interactions and relationships. As our late friend, Rob Raskin (2012, personal communication), stated, "A characteristic of ecosystems is their interactions with the environment. Often, this role is more than a passive one, in that the ecosystem changes the environment—similar to how data affects the underlying science." The ecosystem concept emphasizes adaptation, evolution, and diversity rather than a centralized command and control structure. It is similar to what Davenport (1997) and Nardi and O'Day (2000) call an "information ecology," and it provides a useful perspective. 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, culture …? None of these are completely clear or are likely to truly resonate with researchers and help them understand their role.

Baker and colleagues (Baker & Millerand, 2010; Baker & Bowker, 2007) bring "infrastructuring" and information ecology together. They use the science and technology studies approach of infrastructure studies to examine the infrastructure of an ecology. They may be on to something. Sometimes mixing metaphors is necessary. Perhaps we should not try and find an overarching metaphor for the whole data management process. Perhaps that misses the point. Historically, in literature publication, each publisher filled the multiple roles of archiving, registration, dissemination, and certification of the paper. Van de Sompel et al. (2004) and Priem and Hemminger (2012) 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. Priem and Hemminger 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). Van de Sompel et al. make a similar argument for a "scholarly ecology". We welcome these ideas and suggest that similarly we need to start decoupling or disaggregating the functions of data stewardship to consider each function fully. By disaggregating we can also re-aggregate in new and different ways. For example, people are beginning to consider alternative aggregations of data in ways that connect Data Publication and Linked Data concepts. Alternative forms of information aggregation have been described as "publication packages" (Hunter, 2006) or "research objects" (Bechhofer et al., 2011; Belhajjame et al., 2012). In a modern information ecosystem it is unreasonable to assume one entity would do everything. It is necessary to 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.*

Schopf (2012) argues that we should treat data the way we build production software and that this will make data more readily accessible and available for broad re-use. She states: "We should be treating data as an ongoing process," which presents a very different perspective than one that views data as a publication or an object. She further argues that this metaphor is readily understood and adopted by scientists. This is an interesting worldview. We like the emphasis on cyclical development and controlled, staged releases (e.g., development, staging, production). This perspective may not fully consider preservation, and it creates interesting, perhaps inappropriate, licensing analogs, but it is worthy of further exploration. While recognizing the limitation of Big Iron, other production models could also be worth examining. For example, Morton and Pentico (1993) describe multiple levels of heuristic scheduling systems, and Chase et al. (1998) make careful distinction between manufacturing and service firms. These different classifications of "production" could be examined, much like different classes of "publication". We also find Van de Sompel's (2004) description of a value chain useful.

Another metaphor is one of the marketplace or bazaar. We revisit Raymond's 1999 classic *The Cathedral and the Bazaar.* Metaphorically, considering a bazaar 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. We note also that a marketplace is a spatial metaphor. People use other spatial metaphors as well. We often hear discussion of an information or knowledge space. Baker and Yarmey (2009) use the concept of a "sphere of influence" to differentiate types of repositories. They introduce the intriguing concept of "sociotechnical distance" created by issues of communication, representation, filtering, and transformation rather than physical distance.

Let us not be afraid to explore, mix, and match these and other metaphors. Let us preserve data in formal, curated *archives*. Let us make data available in rapid, cyclical, carefully versioned and described *releases*. Let us *track* data as it moves through the *marketplace* or *ecosystem*. Let us use many narratives to describe and understand complex processes. 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. As such, in the next section we put forth our view of a representative (but not comprehensive) research agenda intended to stimulate further discussion, application, and critical appraisal of current and future worldviews to making data preserved and widely available.

6 RESEARCH AGENDA

As mentioned, we seek to foster an ongoing conversation in the data science community. We, the authors, are but dilettantes in cognitive and social science. We are not theorists; we are practicing and teaching data scientists. But we believe data science can learn by examining how other disciplines and theory can inform practical data management approaches. We, therefore, end this essay with a proposed agenda for research and development. We suggest that there are important lessons to be learned from a closer examination of data science by practicing data scientists themselves. Science and Technology Studies (STS) based approaches, rooted in the principles of "Science in Action" (Latour, 1987), have shown to be very useful in understanding how science and informatics are actually conducted and how data are handled and perceived (e.g., Baker & Millerand, 2010; Bowker & Star, 2000; Harvey & Chrisman, 1998; Parsons et al., 2011; Star & Ruhleder, 1996). We need more STS-based examinations of data science practice that considers sociotechnical *and cognitive* processes and examines the particular attitudes and perceptions of data stewardship and informatics that emerge from different domain and data science worldviews and ways of knowing. More importantly, we need to use that critical examination to develop creative *solutions* to the challenges of data science and stewardship. Broad, critical, multi-facetted analyses of "data science in action" can reveal potential new sociotechnical solutions to data science challenges. And we can use this analytical framework to examine or test how different solutions are understood, adopted, and adapted by different communities.

As we examine different worldviews, we need a fuller development and understanding of all the roles in the entire data stewardship enterprise. Lawrence et al. (2011) lay out a series of defined roles from a Data Publication perspective. Baker and Bowker (2007) do the same from an ecological infrastructuring perspective. Baker and Yarmey (2009) further examine the specific roles of data curation. Schopf (2012) has yet another examination from a production software perspective. They all emphasize different roles with different terms and even seem to define the term "role" differently. A deeper comparison of these roles and how data managers and all the players in the enterprise perceive them is warranted. Are the different actors using the same frames and metaphors and in the same way? Is there a difference across disciplinary cultures? How do the worldviews and metaphors of data creators and data users align? Do the metaphors and frames of data scientists help or hinder that alignment? A particularly critical set of roles falls in the category of what Baker and Bowker (2007) call "in between" work. These roles of the intermediary and "middleware" connecting computer science and domain science are central to informatics and data science (Fox, 2011), yet they are also often hidden from view. Similarly, the role of a curator is critical, but as Fleischer and Jannaschk (2011) illustrate, curation can also introduce a bottleneck in data archiving and release processes. They suggest a closer examination of the role of the data manager or curator and automated curation services. In such an examination, we must consider not just the science domain but also the culture from which curators emerge. The culture of an academic library or archive is vastly different from that found in operational weather center, for example. Finally, in the examination of roles, we should use different worldviews to tease out what important roles we have missed. For example, the roles of the financial sponsor or the unintended non-specialist in the overall data ecosystem have not been examined in depth.

In addition to these broader explorations, more specific research ideas emerged from our critique and earlier feedback. Crosscutting issues emerged around data quality, data referencing, and the norms of data sharing. Data quality is an incredibly complex and subjective issue. Given its subjective nature, it seems appropriate to explore flexible means of community annotation and usage tracking as a means to better understand who are using the data for what purpose and how. This and many other aspects of good data stewardship require careful tracking through precise referencing of the data. This need for precise, continuous, dynamic referencing is closely related to but needs to be considered independently from issues of credit, discovery, and quality. For example, Bruce Caron (2011, personal communication) suggested that we consider how we might track "badges" of recognition for the many roles in the data value chain. This might help address tensions around individual vs. institutional credit and accountability; in essence re-evaluating yet another unexplored metaphor of contracts. Going forward, we see many ideas that need further exploration. We close with an initial, incomplete list of short- and longer-term research questions to be explored to identify key components of an enabling data infrastructure that would promote data availability across *many* worldviews and metaphors:

- What informatics and STS approaches can foster new and robust peer norms for science data stewardship and how may they be evaluated?
- What sociotechnical means exist for the precise, continuous, and dynamic referencing and sharing of data from creation all the way to discovery.

- How can we (or should we) evolve peer norms of data sharing to a more "commons" (Bailey & Tierney, 2002, Beagle, 1999) based approach built around ethical rather than proprietary concerns where data are viewed as a networked public good rather than an owned object?
- Can a Contract metaphor (as applied to social networks, for example) be articulated for making data widely available?
- We know that researchers from different disciplines have different attitudes toward data sharing (Key Perspectives Ltd, 2010; Parsons et al., 2011). Do they also have different attitudes when they are presented with different data management metaphors? Do they have different expectations about reuse, credit, user-responsibility, etc. when they "release" data rather than "publish" data?
- How can we identify and track data and related contextual information immediately upon creation? For example, some have suggested the concept of a "Dropbox[16] for the field scientist", where data and related information are deposited as they are created and are immediately available to curators and potentially other researchers.
- What value-added steps in the data life cycle need to be explicitly credited? How? What approaches (e.g., capability or maturity models) are applicable to determine when computer and information science innovations are ready for data science communities?
- What approaches are needed to bridge the needed domain, data, and computer science disciplines into cohesive collaborations when needed?
- With increasing data intensity, what approaches in the data life cycle need to scale (in numbers of data sets, across disciplines, etc.)?
- What form of improved preservation for large-scale systems are available or need to be developed?
- How can research collections be discovered beyond the context of the scholarly article?
- What are the essential elements of data quality? What standards and technical means are available to capture these elements? How are these balanced or augmented by annotations, recommendations, or qualified citations?
- Should we reexamine our base metaphor of data as a first-class object? Is it sufficient that data simply be accessible, preserved, and usable? Does scientific rigor really require that we give data such formal, independent attention?

It is time for the all stakeholders in the Data Ecosystem (yes, our metaphor) to step outside their comfort zone, examine their worldview, clarify and share it with others, listen to alternate approaches and views, and integrate, assimilate, and evolve. The ideas in this essay must only be a beginning. We hope we have provoked a range of responses and a few ideas from the reader. We look forward to continued discussion, research, *and* action. More metaphors, please.

7 ACKNOWLEDGEMENTS

We sincerely thank the dozens of reviewers who have guided us formally and informally over the past year. In particular, the critiques of Bruce Barkstrom, David Carlson, Bryan Lawrence, Chris Rusbridge, and Lynn Yarmey made this a much better essay.

8 REFERENCES

ARL Joint Task Force on Library Support for E-Science (2007) *Agenda for Developing E-Science in Research Libraries.* Retrieved January 16, 2013 from the World Wide Web: http://www.arl.org/bm~doc/arl_escience_final.pdf

Arzberger P., Schroeder P., Beaulieu A., Bowker G., Casey K., Laaksonen L., Moorman D., Uhlir P., & Wouters P. (2004) Science and government: An international framework to promote access to data. *Science. 303*(5665): pp. 1777-1778.

Atkins D.E., Droegemeier K.K., Feldman S.I., Garcia-Molina H., Klein M.L., Messerschmitt D.G., Messina P., Ostriker J.P., & Wright M.H. (2003) *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* Retrieved November 26, 2011 from the World Wide Web: http://www.nsf.gov/od/oci/reports/toc.jsp

Bailey, R. and Tierney, B., (2002) Information commons redux: Concept, evolution, and transcending the tragedy of the commons. *The Journal of Academic Librarianship 28*(5): pp. 277–286.

16 http://www.dropbox.com/

Baker K.S. & Bowker G.C. (2007) Information ecology: open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems.* 29(1): pp. 127-144. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1007/s10844-006-0035-7

Baker K.S. & Millerand F. (2010) Infrastructuring ecology: challenges in achieving data sharing. In Parker J., Vermeulen N., & Penders B. (Eds.). *Collaboration in the New Life Sciences.* Surrey, England: Ashgate Publishing

Baker K.S. & Yarmey L. (2009) Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation* 4(2). Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/doi:10.2218/ijdc.v4i2.90

Ball A. & Duke M. (2012) *How to Cite Datasets and Link to Publications. DCC How-to Guides.* Digital Curation Centre. Retrieved January 16, 2013 from the World Wide Web: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Beagle, D. (1999) Conceptualizing an information commons. *The Journal of Academic Librarianship 25*(2): pp. 82–89. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1016/S0099-1333(99)80003-2

Bechhofer S., Buchan I., De Roure D., Missier P., Ainsworth J., Bhagat J., Couch P., *et al.* (2011) Why linked data is not enough for scientists. *Future Generation Computer Systems.* Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1016/j.future.2011.08.004

Belhajjame K., Goble C., & De Roure D. (2012) Research object management: opportunities and challenges. *Data Intensive Collaboration in Science and Engineering (DISCOSE) workshop, collocated with ACM CSCW 2012.*

Bowker, G.C. & Star S.L. (2000) *Sorting Things Out: Classification and its Consequences.* Boston, MA: MIT Press

Bowker G.C., Baker K., Millerand F., & Ribes D. (2010) Toward information infrastructure studies: Ways of knowing in a networked environment. *International Handbook of Internet Research.* Springer Science+Business Media.

Callaghan S., Hewer F., Pepler S., Hardaker P., & Gadian A. (2009) Overlay journals and data publishing in the meteorological sciences. *Ariadne.* (60).

Chase, R.B., Aquilano N.J., & Jacobs F.R. (1998) *Production and Operations Management: Manufacturing and Services. 8th edition.* Boston, MA: Irwin/McGraw-Hill

Costello M.J. (2009) Motivating online publication of data. *Bioscience.* 59(5): pp. 418-427. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1525/bio.2009.59.5.9

Davenport, T.H. & Prusak L. (1997) *Information Ecology: Mastering the Information and Knowledge Environment.* Oxford, UK: Oxford University Pres

Doorn P. & Tjalsma H. (2007) Introduction: archiving research data. *Archival Science.* 7(1): pp. 1-20. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1007/s10502-007-9054-6

Duerr R., Downs R., Tilmes C., Barkstrom B., Lenhardt W., Glassy J., Bermudez L., & Slaughter P. (2011) On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics.* 4: pp. 139-160. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1007/s12145-011-0083-6

Edwards P.N., Jackson S. .J., Bowker G.C., & Knobel C.P. (2007) *Understanding Infrastructure: Dynamics, Tensions, and Design.* National Science Foundation. Retrieved May 30, 2012 from the World Wide Web: http://hdl.handle.net/2027.42/49353

ESIP (Federation of Earth Science Information Partners) (2012) *Data Citation Guidelines for Data Providers and Archives.* Parsons M.A., Barkstrom B., Downs R.R., Duerr R., Tilmes C., & ESIP Preservation and Stewardship Committee (Eds.) ESIP Commons. Retrieved September 1, 2012 from the World Wide Web: http://commons.esipfed.org/node/308

FGDC (Federal Geographic Data Committee) (1997) *A strategy for the NSDI.* Retrieved September 3, 2012 from the World Wide Web: http://www.fgdc.gov/policyandplanning/A%20Strategy%20for%20the%20NSDI%201997.doc/at_download/file

Fillmore C.J. (1976) Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences.* 280(1): pp. 20-32. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1111/j.1749-6632.1976.tb25467.x

Fleischer D. & Jannaschk K. (2011) A path to filled archives. *Nature Geoscience.* 4(9): pp. 575-76. http://dx.doi.org/10.1038/ngeo1248

Fox, P. (2011) The rise of informatics as a research domain. *Proceedings of the Water Information Research and Development Alliance.* CSIRO e-Publication, retrieved January 19, 2012 from the World Wide Web: http://www.csiro.au/WIRADA-Science-Sympsosium-Proceedings pp. 125-132.

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

Harley J.B. (1989) Deconstructing the map. *Cartographica: The International Journal for Geographic Information and Geovisualization.* 26(2): pp. 1-20. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.3138/E635-7827-1757-9T53

Harvey F. & Chrisman N. (1998) Boundary objects and the social construction of GIS technology. *Environment and Planning A.* 30(9): pp. 1683-1694.

Hunter J. (2006) Scientific publication packages: A selective approach to the communication and archival of scientific output. *The International Journal of Digital Curation.* 1(1): pp. 33-52.

ICSU. (2011*) Interim Report of the ICSU ad-hoc Strategic Coordinating Committee on Information and Data*. Retrieved January 16, 2013 from the World Wide Web: http://www.icsu.org/publications/reports-and-reviews/strategic-coordinating-committee-on-information-and-data-report/SCCID_Report_April_2011.pdf

ISO (2003) *ISO Standard 14721:2003, Space Data and Information Transfer Systems—A Reference Model for an Open Archival Information System (OAIS).* International Organization for Standardization

Key Perspectives Ltd (2010) *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability.* Digital Curation Center. Retrieved February 5, 2011 from the World Wide Web: http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf

Klump J., Bertelmann R., Brase J., Diepenbroek M., Grobe H., Höck H., Lautenschlager M., Schindler U., Sens I., & Wächter J. (2006) Data publication in the open access initiative. *Data Science Journal.* 5: pp. 79-83. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.2481/dsj.5.79

Koch T. (2004) The map as intent: Variations on the theme of John Snow. *Cartographica.* 39(4): pp. 1-13.

Kuhn, T.S. (1996) *The Structure of Scientific Revolutions. 3rd edition.* Chicago, IL:: University of Chicago Press

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. & Johnson M. (1980) *Metaphors We Live By.* Chicago: The University of Chicago Press

Lakoff G. (1993) The contemporary theory of metaphor. In Ortony A. (Ed.). *Metaphor and Thought, 2nd edition,* Cambridge: Cambridge University Press

Lakoff G. & Johnson M. (1980a) The metaphorical structure of the human conceptual system. *Cognitive Science.* 4(2): pp. 195-208. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1207/s15516709cog0402_4

Latour, B. (1987) *Science in Action: How To Follow Scientists and Engineers Through Society.* Cambridge, MA: Harvard University Press

Lawrence B., Jones C., Matthews B., Pepler S., & Callaghan S. (2011) Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation.* 6(2).

McMahon M. (1996) Overview of the Planetary Data System. *Planetary and Space Science.* 44(1): pp. 3-12. Retrieved January 19, 2013 from the World Wide Web: http://dx.doi.org/10.1016/0032-0633(95)00101-8

Morton, T.E. & Pentico D.W. (1993) *Heuristic Scheduling Systems: With Applications To Production Systems And Project Management,* Wiley-Interscience.

Nardi, B.A. & O'Day V. (2000) *Information Ecologies: Using Technology With Heart.* Boston, MA: MIT Press

National Research Council (2007) *Environmental Data Management at NOAA: Archiving, Stewardship, and Access.* Washington, DC: National Academies Press

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

Parsons M.A., Duerr R., & Minster J.B. (2010) Data citation and peer-review. *Eos, Transactions of the American Geophysical Union.* 91(34): pp. 297-98. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/doi:10.1029/2010EO340001

Parsons M.A., Godøy Ø., LeDrew E., de Bruin T.F., Danis B., Tomlinson S., & Carlson D. (2011) A conceptual framework for managing very diverse data for complex interdisciplinary science. *Journal of Information Science.* 37(6): pp. 555-569. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1177/0165551511412705

Pfeiffenberger H. & Carlson D. (2011) "Earth System Science Data" (ESSD) — A peer reviewed journal for publication of data. *D-Lib Magazine.* 17. Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.1045/january2011-pfeiffenberger

Priem J. & Hemminger B.M. (2012) Decoupling the scholarly journal. *Frontiers in Computational Neuroscience.* 6(19). Retrieved January 16, 2013 from the World Wide Web: http://dx.doi.org/10.3389/fncom.2012.00019

Raymond, E.S. (1999) *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* Cambridge, MA: O'Reilly

Schopf J.M. (2012) Treating data like software: A case for production quality data. *Proceedings of the Joint Conference on Digital Libraries.* pp. 11-14 June 2012, Washington DC.

de Sherbinin A. & Chen R.S. (Eds). (2005) *Global Spatial Data and Information User Workshop: Report of a Workshop.* Socioeconomic Data and Applications Center, Center for International Earth Science Information Network, Columbia University. Retrieved February 5, 2011 from the World Wide Web: http://sedac.ciesin.columbia.edu/GSDworkshop/

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

de Waard, A. & Kircz J. (2008) Modeling scientific research articles--shifting perspectives and persistent issues. *Proc. ELPUB2008 Conference on Electronic Publishing.*

de Waard, A., Breure L., Kircz J.G., & Van Oostendorp H. (2006) Modeling rhetoric in scientific publications. *International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006.*

Van de Sompel H., Payette S., Erickson J., Lagoze C., & Warner S. (2004) Rethinking scholarly communication. *D-Lib Magazine.* 10(9): pp. 1082-9873.

Wright A (2009) Exploring a 'Deep Web' that Google can't grasp. *The New York Times.* February 22, 2009. Retrieved January 16, 2013 from the World Wide Web: http://www.nytimes.com/2009/02/23/technology/internet/23search.html

(Article history: Available online 31 January 2013)