The International Journal of Digital Curation
Issue 1, Volume 6 | 2011

Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects

Laura Wynholds,
University of California, Los Angeles

Abstract

Within information systems, a significant aspect of search and retrieval across information objects, such as datasets, journal articles, or images, relies on the identity construction of the objects. This paper uses identity to refer to the qualities or characteristics of an information object that make it definable and recognizable, and can be used to distinguish it from other objects. Identity, in this context, can be seen as the foundation from which citations, metadata and identifiers are constructed.

In recent years the idea of including datasets within the scientific record has been gaining significant momentum, with publishers, granting agencies and libraries engaging with the challenge. However, the task has been fraught with questions of best practice for establishing this infrastructure, especially in regards to how citations, metadata and identifiers should be constructed. These questions suggests a problem with how dataset identities are formed, such that an engagement with the definition of datasets as conceptual objects is warranted.

This paper explores some of the ways in which scientific data is an unruly and poorly bounded object, and goes on to propose that in order for datasets to fulfill the roles expected for them, the following identity functions are essential for scholarly publications: (i) the dataset is constructed as a semantically and logically concrete object, (ii) the identity of the dataset is embedded, inherent and/or inseparable, (iii) the identity embodies a framework of authorship, rights and limitations, and (iv) the identity translates into an actionable mechanism for retrieval or reference.[1]

[1] This paper is based on the paper given by the authors at the 6th International Digital Curation Conference, December 2010; received December 2010, published March 2011.
The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

As scholarship moves towards a new paradigm of data-driven science with increasingly large and complex datasets (Hey et al., 2009; Gray et al., 2005), we have seen a wide-sweeping push to establish the infrastructure to make research data available in conjunction with journal publications. This push includes major journals, such as Science and Nature; major granting agencies, such as the National Science Foundation and National Institutes of Health; major universities, such as the University of California; and major research institutions, such as the British National Library (Chavan & Ingwersen, 2009; Brase, 2009). Many academic libraries and institutional repositories are poised on the cusp of housing datasets, placing them in a position where they will need to engage with the challenge of defining and managing access to datasets (Baker & Yarmey, 2009).

This pressure to archive and then link the scholarly publications to the data has been growing, based on arguments that the data is essential for establishing validity, reproducibility and replicability, in addition to fostering value-added activities, such as democratizing access to publicly funded research and facilitating new kinds of analysis (Bell, Hey, & Szalay, 2009; National Science Board, 2005). There is little question that forming a paradigm for the inclusion of data is critical for scientific records in an age of digital records.

Data, as a digital object to be discovered, accessed and preserved, is a relatively new paradigm for scholarship. As Borgman (2007) notes, "libraries and publishers assure access to publications, but no comparable infrastructure exists for access to data and unpublished resources". While many researchers have been contributing towards the establishment of infrastructure for the curation of datasets, such as data policy, stewardship, provenance tracking, permanent identifiers, metadata, and citations, (Buneman et al., 2006; Hilse & Kothe, 2006; Wallis et al., 2010; Paskin, 2005; Zimmerman, 2007) the process has yet to be formally systematized or institutionally adopted, yielding many nodes struggling to assemble themselves into a functional system. Complex challenges, both social and technical, have continued to impede the process of knitting together these elements (Murray-Rust & Rzepa, 2004; Chavan & Ingwersen, 2009; Pollard & Wilkinson, 2010; Carlson & Anderson, 2007).

One group of data scientists went as far as to observe: "In some respects, the researcher was better off in the days of paper publication and record keeping, where there are well-defined standards for citation and some confidence that the cited data will not change." (Buneman et al., 2006). Finding this comment in one of the major scholarly forums for data engineers and computer scientists is indicative of the deep challenge of incorporating datasets into scholarly publications.

This paper argues that one central aspect of bringing these systems together is dataset identity. Despite the presence of a large amount of work on the data deluge, there has been very little work exploring how the construction of dataset identity should function within the conceptual framework of scholarly publication specifically and the scientific record more generally. As such, this paper focuses on the identity aspects of these information objects in terms of the work identity does within scholarly information systems. In doing so, the discussion intentionally steps back from specific structures and technologies in favor of looking at function. Through this novel way of

looking at dataset identity, this paper aims to establish a basis for which data structures and standards can be evaluated for their potential utility within existing and evolving scholarly practices.

Why Focus on Identity?

Identity, as a concept, has a long tradition within mathematics of referring to the qualities or characteristics of an information object that make it definable and recognizable, such that it can be distinguished from other objects. Identity, in this context, can be seen as the foundation from which citations, metadata and identifiers are formed. Dataset identity is used in this paper to refer to the abstract conceptualization of a dataset, under which a person can locate concrete physical and digital objects.

The conceptual understanding of what a dataset is governs how it is represented in an information system, and thus has significant implications for how datasets are managed. In their paper, "What is a Text, Really", DeRose et al (1990) argued that a flawed representation of documents can prohibit real progress in organizing and retrieving documents. They wrote: "The way in which text is represented on a computer affects the kinds of uses to which it can be put by its creator and by subsequent users." The way datasets are constructed as information objects has long term ramifications for discovery and access. Once the identity is defined, both machine and human users must be able to parse the identity correctly, e.g. that two objects are the same, are not the same, are different versions, are derivations of each other, etc.

Renear & Dubin (2003) argued for the importance of identity, engaging with document identity conditions, which are "a method for determining whether an object x and an object y are the same object." They concluded: "Identity conditions are arguably an essential feature of any rigorously developed conceptual framework for information modeling." These identity conditions form the logic structures for automating discovery and access.

This paper presents aspects of dataset identity that are foundational to their function in a scholarly information system, focusing on the example of constructing dataset identity such that it can be linked to. I argue that the construction of identity is not simply a technical challenge of metadata, permanent URLS, DOIs, or description, but a larger challenge of constructing a model for datasets that folds these irregular, evolving and obsolescing technocentric objects into an established model of scholarship and the scientific record.

Why Look at Linking?

Despite the lack of consensus for how to incorporate datasets within scholarly information systems, it seems reasonable to expect linking to be a major aspect of how datasets are managed, as evidenced by projects like OAI-ORE and Linked Data (Pepe et al., 2009; Bizer et al., 2009). In order to link to the data, significant decisions need to be made as to what parts of the research processes are represented in the linked object, which is, in turn, heavily reliant on the identity construction of the data. The following comments help illustrate the challenges of linking to data:

> "All the publications should link data. They should be interlinking between the archives and the literature. One of the goals of the Virtual Astronomical Observatory is actually to do that, and do it properly. It hasn't happened yet for many reasons. It's actually harder than you would think." (Personal Interview, NASA Data Archive Manager, Data Conservancy NSF award OCI0830976, February 11, 2010).

Example of a Dataset

For illustrative purposes the following short, highly simplified scenario is presented (see Figure 1):

> An astronomer downloads data from three specific data releases of three different telescope projects, each time querying only a specific narrow section of the sky. She cleans up the data, performing some computational transformations, and ultimately compiles the data into a dataset which she uses for analysis. She iteratively works with the data, producing findings, performing computational analyses, producing findings, until she has four articles. Upon acceptance of one of the articles, she is told she needs to submit/release her data in conjunction with the final manuscript. Which of the multiple digital objects described above does she need to link to? Should she reference the same dataset for all four articles?
>
> Eighteen months after her findings are published, a colleague has questions about how the data was handled in her computations. In order to answer her colleague's questions, the astronomer finds herself having to retrace her own steps, many of which were not fully documented and relied upon scripts that she had not looked at since the article was submitted. Which of these digital objects does the colleague need to verify her work and methods, and are they the same as those for publication (above)?
>
> Two years after her findings are published, in response to the assertion that there may be an error in the mathematical model for the computational scripts used, one of her students is compiling a survey of every article and dataset that relied on those scripts. Supposing that he was able to find the associated data, would he be able to identify which of the above digital objects were implicated, such that he could respond to the assertion?

As this example demonstrates, when attempting to capture research data the question of "what, exactly, are you representing and/or capturing for the scholarly record?" has significant and wide reaching implications for what kinds of activities the derivative digital object will support. In this case, the dataset generated by the research can be considered a compound object, consisting of multiple overlapping datasets, that together with research tools and processes, form the researcher's conceptualization of "the data".

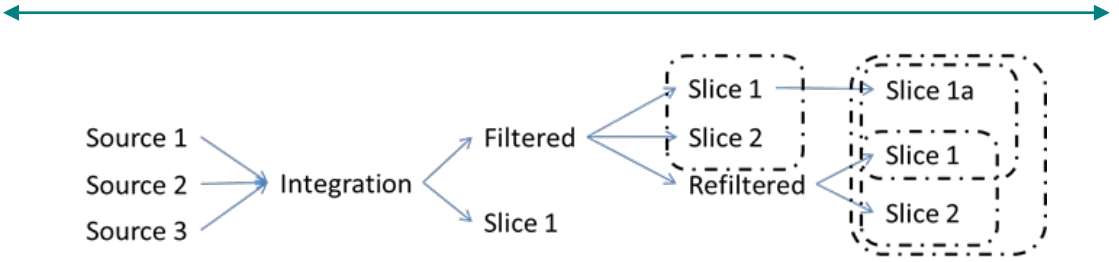

Figure 1. Data Manipulation Yielding Multiple Branching and Overlapping Versions. Diagram courtesy of Jillian Wallis.

Figure 1 shows the series of versions of datasets created as the researcher moved through different processes, leading to a complex branching of versions further complicated by the fact that these datasets all reside within one digital object (a database). The dashed boxes indicate dataset versions used for various publications. As a result, it is difficult to use a URI for the slices as the database is neither online, nor does it offer the specificity to retrieve those individual slices.

Constructing Dataset Identity for Scholarly Publication

The underlying framework of this paper lies in a synthesis of functions of dataset identity from the complex challenges described in the scholarly literature and encountered in our research. However, this paper takes a novel approach of engaging with the question of dataset identity functions across a distributed scholarly publication system. Based on the understandings of the challenges facing dataset identity construction, this paper argues for four functional requirements: (i) the dataset is constructed as a semantically and logically concrete object; (ii) the identity of the dataset is embedded, inherent and/or inseparable; (iii) the identity embodies a framework of authorship, rights, and limitations; and (iv) the identity translates into an actionable mechanism for retrieval or reference. These four functions are explained in more depth below.

The definition of what a dataset is and how it should function within the context of a structured information schema is not a priori given, even within well-established disciplines and methods (Carlson & Anderson, 2007; Collins, 1998; Cole, 2008; Renear et al., 2010). The unifying aspect of data has been in terms of its evidentiary role within the research process. This suggests that any scholarly information system intending to manage datasets will have to accommodate a great deal of variety, even within a single discipline. These challenges may be reasonably expected as a natural process of establishing information infrastructure. Star and Bowker (1999) argued that this is, by nature, an act of reification of exemplars and cognitive conventions, the process of which frequently yields different conceptualizations across different social groups. Rather than standing immutably, these information objects are "embedded within webs of socially organized, situated practices".

The Dataset is Constructed as a Semantically Concrete Object

In practice, as demonstrated by the example in Figure 1, the term "dataset" often refers to an evolving constellation of databases, files, and associated information as a single information object. Unfortunately, this fuzzy conception of datasets presents a problem, as it can yield multiple overlapping versions of what may be described as the same "thing". Pollard and Wilkinson (2010) reported on this problem, referring to the

challenge of "dynamic datasets, different renditions of datasets, and what [the presenter] referred to as the 'Russian Doll Problem', where datasets are progressively merged".

This paper proposes to address this slippage by treating datasets in terms of a mathematical set. By treating datasets as sets, then the dataset is defined by its members. Should one of the members change, then the new set must, by definition, be assigned a new identity so that the two sets can be successfully distinguished from each other. Establishing and modeling the relationships of these objects is contingent on establishing the identity of the objects.

If scientists are to be able to expect to use these objects as evidence (Collins, 1998; Scheiner, 2004), concreteness is an essential function of the citation. The identity should be formed around these objects with enough specificity (e.g., version or date of the records) such that the citation invokes one and only one, unambiguous, clearly defined dataset. This is not simply a problem of assigning identifiers or metadata, but rather for the purposes of aggregation, computation, verification, reproducibility and replicability the dataset must be defined such that it can yield a concrete search result.

The Identity of the Dataset is Embedded, Inherent and/or Inseparable

Within current scholarly practices surrounding journal articles, conventions within publishing practices embeds the identity of the article within the information object, typically within the first page.

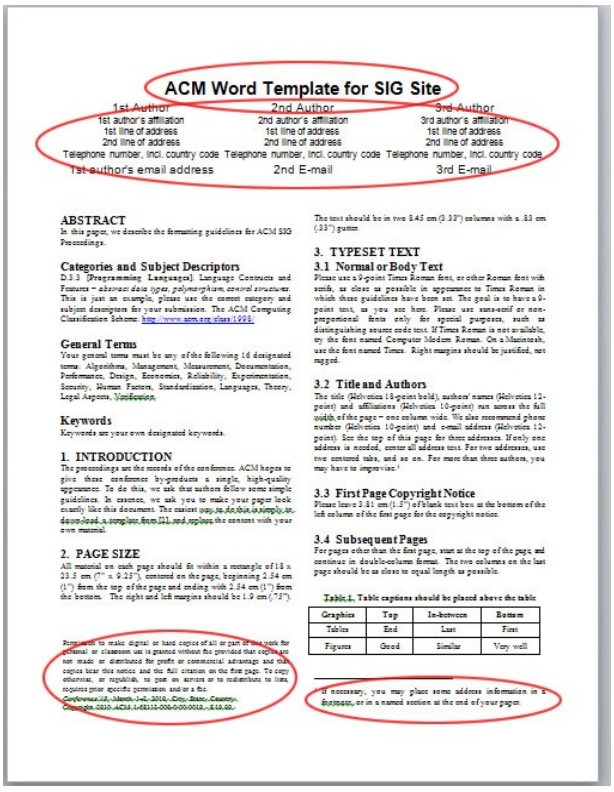

Figure 2. Example of Embedded Identity in Journal Articles

For example, the ACM conference proceedings article template shown in Figure 2 embeds identity elements in on the first page (circled) such that the following elements are included within the content: author name, author affiliation, title of paper, conference name, conference date, location, copyright/permission and contact information. Compare this with a common form of tabular data, the spreadsheet (such as a csv or excel file). The spreadsheet, emailed from one scientist to another, easily sheds unembedded aspects of identity in the act of being passed, renamed or downloaded.

Gilliland-Swetland (2000) articulated the challenge of provenance for digital objects: "The integrity of the evidential value of materials is ensured by demonstrating an unbroken chain of custody, precisely documenting the aggregation of archival materials." Within the context of data and datasets, this suggests that identity of derivative datasets needs to be captured in such a way that the identity of the data is preserved during download and use. Many types and granularities of technology support functions of this nature, from eScience provenance management systems , (Groth et al., 2008) to embedded metadata headers. Given the wide variety of technologies and formats deployed across the sciences and the relatively small market share for research applications, one can reasonably anticipate a long tail of data that is not going to fall under the domain of officially supported cyberinfrastructure (Heidorn, 2008). However, while identity construction should benefit from advances in technology, such as provenance systems, the lack of such technology should not cripple our ability to construct dataset identities. Therefore, in order to remain identifiable, the identity must be inherent enough to support changes in structures and formats, as well as the changes in context. In summary, within scholarly work there is a value to being able to identify the object across changing contexts, and, as such, there is an argument to make the object tightly coupled with its identity.

The Identity Embodies a Framework of Authorship, Rights, and Limitations

Central to researchers' discussions of data reuse and sharing is the recognition that multiple factors, such as authorship, rights, limitations, trust, and permissions, play a central role in their relationship to data. (Cambell et al., 2002; Carlson & Anderson, 2007; Chin & Lansing, 2004). These constraints can be as complex and exacting as the practices from which they are derived. The limitations may be derived from the technology, such as error and accuracy ranges for a sensor; from social mores, such as privacy protections; or from concerns that the data may be misused to support claims for which the data is misapplied. These concerns may derive from the assumptions or parameterizations of the models, or from the politicization of the science, or both, as demonstrated with climate modeling (Edwards, 1999).

Failure to build structures for recognition into dataset records has the potential to hobble the development of datasets as first class objects, as it is one of the mainstays of academic incentive structures. Authorship and attribution/citation are two of the central measures of scholarly productivity, and yet, current technologies offer little hope to aggregate either of these for datasets. Birnholtz, (2008) while studying large high energy physics (HEP) collaborations at CERN, observed that without a publication or attribution system it was "quite easy to get lost or even crushed in the crowd of a large HEP collaboration. Breakdowns in informal systems of recognition, of course, are not a novel result on their own. What distinguishes the present discussion is the almost complete absence of a formal record to fall back on." In an

earlier paper describing the difficulties of authorship in large collaborations, Birnholtz (2006) noted: "authorship has multiple functions in the sciences. We can describe these as follows: 1) attributing credit for discoveries to a person or group of people; 2) assigning ownership to this person or persons; and 3) enabling the accrual of reputation." While there are differences in how authorship functions between publications and data, Birnholtz's work suggests that there are critical functions for dataset identity in regards to credit, attribution and reputation.

These aspects of reputation and authorship are foundational to scholarship and should not be ignored for dataset identity. Chavan and Ingwersen (2009) saw this gap as well, lamenting "we lack consistency in data citations, which is sure to provide much needed high visibility to data. It is difficult or impossible given the existing citation metrics system to identify who originally created or added value to a datum".

This becomes particularly difficult for data, as while it is clear that data operates under a framework of rights, authorship, and limitations, it is unclear how authorship, rights and limitations should be applied to something which falls outside of the copyright driven models of the print world. The identity of the dataset must therefore embody functions of a system of authorship, rights, limitations, and permissions in a way that is heritable (meaning that a derivative of a restricted dataset should also be restricted), transparent, and flexible in the face of a changing landscape.

The Identity Translates into an Actionable Mechanism for Retrieval and Citation

Formal, traditional citations for journal articles are an excellent example of presenting an object's identity such that it can be translated into an actionable mechanism for either retrieval or citation. Citations, with their highly structured formats, can be considered a formalized identity for scholarly products. A complete citation should act to uniquely identify the exact document in question, in cases where the document is not uniquely identified, the citation is considered to be flawed. A complete citation may also act as an access key, providing the essential information required to retrieve the document from an information system. As such, it can be said that citations perform identity functions within information systems, even though these functions may not be fully automated within the system.

DOIs, persistent URLs and Handles all offer solutions for permanent identifiers and offer many retrieval and citation functions in an elegant way (Hilse & Kothe, 2006). However, these technologies all rely on assigning a URI (Berners-Lee, 2006) to the object, and there are several reasons to believe that URIs, by themselves, will not provide answers to how dataset identities should be constructed. First, the infrastructure for assigning URIs to data is still under development, albeit with considerable investment (Brase, 2009; Paskin, 2005). Second, there are semantic granularity questions based on what level an identifier should be assigned such that they offer reasonable efficiency for the system and a concrete citation (Pollard & Wilkinson, 2010). For example, the dataset may be a subset of data within a digital object, as discussed in Figure 1. The technology that houses the data may not support a uniform manner for automatically returning the exact set (such as the results of a federated search query), or the data may reside solely in an offline location, unincorporated with any public information systems, due to necessary access restrictions, such as privacy or proprietary restrictions.

Conclusions

In recent years we have seen a push for the infrastructure to incorporate data as a first class information object within scholarly information systems. Correspondingly, we are engaged in building and vetting new additions to the scholarly publication infrastructure, which are modular and distributed in nature. Representation of identity is critical in this process, and has implications for every function that depends on identifying the information object, including access, retrieval, metrics and aggregation.

In these early stages of infrastructural adoption it is critical that we engage with broad conceptual questions of what is intended to be retained within these information objects, how data should be represented within the system, and within what kinds of structures before these nodes are hardened into rigid infrastructures, classification schemes and standards. Unfortunately, there are many overlapping and intersecting systems that contribute to the ways in which data are unruly and poorly bounded objects within scholarly work. As such, the challenges of dataset identity are not simple to address. While this paper was able to explore and argue for the functions that dataset identity might accommodate within scholarly publications, it was hampered by the lack of well developed conceptual model for data.

Even though these infrastructural technologies are in motion on a horizon between development and obsolescence, in order to evaluate their utility within a larger information infrastructure, we need to understand and evaluate the expectations we have of these objects in within the system. Those expectations are derived directly from the conceptual object of what a (published) dataset is or should be within the scholarly record, and are not simple questions of technology or infrastructure, but rather a vision of how academic work should be done and recorded. It is therefore unconscionable that these important decisions could potentially be left to software developers and publishers as a purely technological challenge.

Acknowledgements

I would like to thank my research team and colleagues for providing feedback on the many drafts of this paper: Christine L. Borgman, Jillian Wallis, Matt Mayernik, Katie Shilton, Alberto Pepe, David Fearon, and Karen Wickett.

References

Baker, K. S., & Yarmey, L. (2009). Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation*, *4, (2).*

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323, (5919)*.

Berners-Lee, T. (2006). What do HTTP URIs identify? *Design Issues. W3C.* Retrieved February 9, 2011, from http://www.w3.org/DesignIssues/HTTP-URI2.

Birnholtz, J. (2008). When authorship isn't enough: Lessons from CERN on the tmplications of formal and informal credit attribution mechanisms in collaborative research. *Journal of Electronic Publishing*, *11, (1)*.

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology*, *57, (13)*.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - The story so far. *International Journal on Semantic Web and Information Systems*, *5, (3)*.

Borgman, C.L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Brase, J. (2009). DataCite-A global registration agency for research data. *Proceedings of the Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology.* Beijing, China.

Buneman, P., Chapman, A., & Cheney, J. (2006). Provenance management in curated databases. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. Chicago, IL, USA.

Campbell, E.G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. *Journal of the American Medical Association*, *287, (4)*.

Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, *12, (2)*.

Chavan, V., & Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, *10 (Supplement 14), S2*.

Chin, G.J., & Lansing, C.S. (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. *Proceedings of the Conference on Computer Supported Cooperative Work*, *6, (3)*. Chicago, Illinois.

Cole, F.T.H. (2008). Taking "data" (as a topic): The working policies of indifference, purification and differentiation. *Proceedings of the Australian Conference on Information Systems.* Christchurch, Australia.

Collins, H.M. (1998). The meaning of data: Open and closed evidential cultures in the search for gravitational waves. *American Journal of Sociology*, *104, (2)*.

DeRose, S.J., Durand, D.G., Mylonas, E., & Renear, A.H. (1990). What is text, really? *Journal of Computing in Higher Education*, *1, (2)*.

Edwards, P.N. (1999). Global climate science, uncertainty and politics: Data-laden models, model-filtered data. *Science as Culture*, *8, (4).*

Gilliland-Swetland, A.J. (2000). *Enduring paradigm, new opportunities: The value of the archival perspective in the digital environment.* Washington, DC: Council on Library and Information Resources.

Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., & Heber, G. (2005). Scientific data management in the coming decade. *SIGMOD Rec.*, *34, (4).*

Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., et al. (2008). The provenance of electronic data. *Communications of the ACM*, *51*, *(4).*

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57, (2).*

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft.

Hilse, H., & Kothe, J. (2006). *Implementing persistent identifiers*. Report for the Consortium of European Research Libraries, European Commission on Preservation and Access.

Murray-Rust, P., & Rzepa, H.S. (2004). The next big thing: From hypermedia to datuments. *Journal of Digital Information*, *5, (1).*

National Science Board (U.S.). (2005). *Long-lived digital data collections: Enabling research and education in the 21st century*. Arlington VA: National Science Foundation. Retrieved February 9, 2011, from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

Paskin, N. (2005). Digital object identifiers for scientific data. *Data Science Journal*, *4*.

Pepe, A., Mayernik, M., Borgman, C.L., & Van de Sompel, H. (2009). From artifacts to aggregations: Modeling scientific life cycles on the semantic web. *JASIST*, *61, (3)*.

Pollard, T.J., & Wilkinson, J.M. (2010). *Making datasets visible and accessible: DataCite's first summer meeting*. Ariadne, 64. Retrieved February 9, 2011, from http://www.ariadne.ac.uk/issue64/datacite-2010-rpt/.

Renear, A., & Dubin, D. (2003). Towards identity conditions for digital documents. *Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications: Supporting communities of discourse and practice.* Seattle, Washington: Dublin Core Metadata Initiative.

Renear, A.H., Sacchi, S., & Wickett, K.M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology Annual Meeting.* Pittsburgh, PA.

Scheiner, S.M. (2004). Experiments, observations, and other kinds of evidence. In M.L. Taper & S.R. Lele (Eds.) *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago: University of Chicago Press.

Star, S.L., & Bowker, G.C. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Wallis, J.C., Mayernik, M.S., Borgman, C L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: from vision to practical reality. *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. Gold Coast, Queensland, Australia.

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7, (1)*.