Analytic Potential of Data: Assessing Reuse Value

Carole L. Palmer Nicholas M. Weber Melissa H. Cragin
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
clpalmer; nmweber; cragin@illinois.edu

ABSTRACT
Realizing the vision of networked data collections and services requires large bodies of scientific data that can be used in new ways. Adapting the concept of epistemological potential, we illustrate an approach for assessing the value of data for reuse in new domains. Two criteria for this analytic potential—integrity and fit-for-purpose—are recognized aspects of data curation, however identifying potential domains of interest for reuse requires knowledge of practices and needs across disciplines. Evaluating analytic potential will become increasingly important for libraries and repositories to make informed decisions about recruitment and curation of data for interdisciplinary science.

Categories and Subject Descriptors
H.3.7: Digital Libraries. Collection.
General Terms
Theory

1. INTRODUCTION
The vision of a robust network of functional data to support grand challenge science has captured the imagination and the effort of many in the scientific community [1,2]. But, realizing this vision requires the ability to collect large bodies of data that are of value for interrogating specific research questions of interest to scientists. As part of the Data Conservancy (DC), we are conducting comparative analysis of data practices across disciplines to inform curation processes. An essential part of scoping this research, and building DC collections more generally, has been the development of a data collection policy to guide data acquisition to meet the DC mission to collect, organize, validate and preserve data that will allow scientists to find new ways to address the grand research challenges that face society.

1.1 Data Conservancy Collection Context
The DC, led by the Sheridan Libraries at Johns Hopkins University, is conceptualized as a "blueprint for research libraries" for data curation and repository services [3]. As such, it aims to provide access to data from a broad range of disciplines, including the earth, life, and social sciences, and astronomy, and to collect at-risk data, as well as highly unique or valuable data, for target research areas—consistent with the traditional role of special collections. Articulating the eligibility of data for DC in the collection policy requires establishing domain targets, as well as criteria for data and metadata quality. However, a more challenging task has been determining criteria for assessing the potential of data to be reused, particularly in new domains.

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

1.2 Assessing Data for Reuse
Research libraries provide information resources to support the production of new knowledge. To fulfill this mission with data resources, they need ways to assess data for applicability to research problems and methods in the communities they serve. Current appraisal techniques for scientific data tend to rely on institutional affiliation, historical value, or some combination of user generated metadata, such as type, format, subject, size, title, and geo-spatial coordinates [4]. In addition, research on the use of existing data for studying new problems emphasizes the importance of capturing the context of data production in metadata and other data documentation [5,6,7]. Most existing studies have also only considered reuse within a single domain and metadata from the perspective of data producers.

2. CONCEPTUALIZING REUSE
2.1 Hjørland's Epistemological Potential
Our conceptual framework for examining the reuse potential of data is derived from Hjørland's [8] idea of "epistemological potential" (hereafter referred to as EP), which refers to how the subject analysis of documents should represent current and future possible uses. According to Hjørland, metadata representing a document in an information system should go beyond providing a description of its aboutness; it should expose its ability to "transfer knowledge", which requires "insight or understanding of which future problems can give rise to the use of the document in question" [9; p. 93]. A document, however, can have an infinite number of properties capable of informing a user, therefore actual description must be informed by an evaluation that considers three criteria: the range of possible user groups—beyond the originally intended audience, contributions to those groups with the most "long-term utility", and categorizations of the prioritized "potentials" as access points in the information system (Fig 1.1).

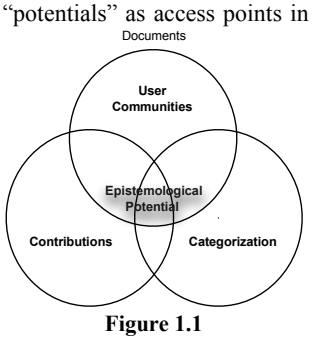

Figure 1.1

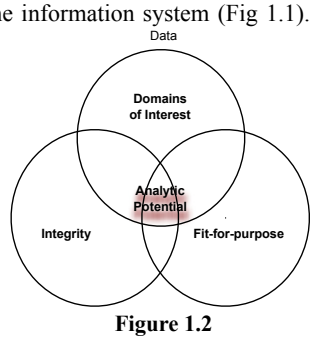

Figure 1.2

2.2 Documents vs. Data
The EP concept and the associated criteria are relevant to data sets as well as documents, but there are important differences in how data transfers knowledge to the user that need to be accounted for in determining reuse potential. With document use (outside of natural language processing or text mining applications) the intelligibility of the information object can be readily assessed and

the content can be interpreted directly by the user. Data sets, on the other hand, usually do not transfer knowledge directly, requiring processing and tools for intelligibility and interpretation. The effort and resources required for this are key considerations in assessing potential for contributing to new knowledge.

2.3 Analytic Potential

While the results of research (as reported in documents) have epistemological potential, we assert that data sets have *analytic potential* (AP) and have identified three criteria for evaluating AP (Fig. 1.2): the *domains of interest* and associated research problems, *integrity* or intelligibility and functionality for reuse, and *fit-for-purpose* for the new application. Determining integrity and fitness for purpose are recognized roles in the data curation process [10]. In DC, for example, integrity is assessed in part by applying OAIS criteria for preservation description information. Fit-for-purpose requires alignment of data units and levels of processing with the methods and tools to be applied. Identifying potential domains is particularly challenging, however. They cannot be inferred from the semantics of the content, as Hjørland demonstrates with the subject analysis of documents. Moreover, scientists often fail to recognize the potential of their data for reuse by others, especially by those outside their field [11].

3. IDENTIFYING DATA & USERS

Through qualitative studies of data practices in domains targeted for inclusion in DC, we are examining the spheres of context [12] around data to assist in operationalizing AP. Fig. 2 presents an outline of three subfields in the earth sciences, specifying the types of data objects to be collected and potential domains of interest. This table is informed by ongoing qualitative work where we are articulating different kinds of curation work to ensure integrity for long-term preservation and to support use for new purposes in other domains [13].

	Geobiology	Volcanology	Soil Ecology
Data objects	Site-specific time series: -"reduced spreadsheets": rock, water chem, microbial; - microscopy images - annotated digital "field photos"	Rock profile -physical rock -thin section -chemical analysis -photographs -field notes	Database - multiple abiotic soil measurements -associated metadata
Domains of interest	Geo/Microbiology Geology Chemistry U.S. Park Service	Igneous geology Petrology Geophysics Geochemistry	Biochemistry Earthworm ecology Sensor Sciences

Figure 2. Domains of Interest for Earth Science Data

4. Conclusion

Evaluating the analytic potential of data will become increasingly important as data repositories strive to make informed decisions about curatorial investment and recruitment of data with high value for interdisciplinary science. As DC continues to develop empirically derived accounts of AP, we will build a base of knowledge to allow more routine assessments, and patterns of use will provide additional evidence of potential domains of interest. Next steps in our research will focus on identifying additional indicators beyond the three criteria derived from Hjørland's model and streamlining protocols for analyzing AP for different kinds of data products and uses. Building high quality, large-scale data collections with long-term value will require much more than acquisition and description of data. It will demand a professional meta-science perspective [14], which includes comprehension of broad cross-disciplinary epistemological trends and historical-cultural dynamics of research areas [15] to anticipate what data will be needed and how they will be analyzed by researchers in the future.

5. ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (OCI- 0830976).

6. REFERENCES

[1] Hey, T., Tansley, S., Tolle, K. (2009). The 4th paradigm: Data-intensive scientific discovery. Microsoft Research, Redmond, WA. http://research.microsoft.com/enus/collaboration/fourthparadigm/

[2] National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. http://www.nsf.gov/pubs/2005/nsb0540/

[4] Choudhury, G.S. & Hanisch, R. (2009, December). Data Conservancy: Building a sustainable system for interdisciplinary scientific data curation and preservation. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data. Presentation given at PV 2009 conference, Madrid, Spain..

[5] NASA Socioeconomic Data and Applications Center (SEDAC) Long-Term Archive. (2005). Appraisal for accession to the SEDAC LTA. http://sedac.ciesin columbia.edu/lta/Appraisal.html.

[6] Borgman, C., Wallis J., & Enyedy N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 7(1/2), 17-30.

[7] Faniel, I.M., & Jacobsen, T.E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work, 19(3-4), 355-375.

[8] 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(1-2), 5-16.

[9] Hjørland, B. (1997). Information seeking and subject representation: An activity-theoretical approach to information science. Westport, CT : Greenwood.

[10] Lord, P., MacDonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. Proceedings of the UK e-Science All Hands Meeting, Nottingham, September 2004.

[11] Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science, and institutional repositories. Philosophical Transactions of the Royal Society A, 368(1926), 4023-4038.

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

[13] Cragin, M.H., Palmer, C.L., & Chao, T.C. (2010). Relating data practices, types, and curation functions: An empirically derived framework. Proceedings of the ASIS&T annual meeting, Pittsburgh, PA, Oct. 22-27, 2010.

[14] Bates, M.J. (1999). The invisible substrate of information science. Journal of the American Society for Information Science, 50(12), 1043-1050.

[15] Hjørland, Birger. (1998). Theory and metatheory of information science: A new interpretation. Journal of Documentation 54: 606-62