

The Analytic Potential of Scientific Data: Understanding Re-use Value

Carole L. Palmer, Nicholas M. Weber, and Melissa H. Cragin

Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
cpalmer; cragin; nmweber@illinois.edu

ABSTRACT

While problems related to the curation and preservation of scientific data are receiving considerable attention from the information science and digital repository communities, relatively little progress has been made on approaches for evaluating the value of data to inform investment in acquisition, curation, and preservation. Adapting Hjørland's concept of the "epistemological potential" of documents, we assert that analytic potential, or the value of data for analysis beyond its original use, should guide development of data collections for repositories aimed at supporting research. Three key aspects of the analytic potential of data are identified and discussed: preservation readiness, potential user communities, and fit for purpose. Based on evidence from research from the Data Conservancy initiative, we demonstrate how the analytic potential of data can be determined and applied to build large-scale data collections suited for grand challenge science.

INTRODUCTION

Research libraries and repositories are facing the monumental task of collecting tremendous amounts of digital data. The volume of data, as well as the complexity and high cost of curating and preserving data, will require these organizations to establish priorities for what they collect. In some sense, they will do what collection developers in research libraries and archives have always done—make judgments about which information sources are of enough value to their service communities to justify the expense of collection, curation, and preservation.

The value of information has always been difficult to operationalize. It depends on who is using the information for what, when. But one important principle, identified by Uhlir in relation to the preservation of scientific data, is that "the value of data increases with their use" (2010, p.

1). Uhlir is concerned with the "information gulags" or data cemeteries—the inaccessible warehouses of data in research centers and the decaying files kept privately by scientists, treated as valueless byproducts rather than assets of research. While the principle is sound, it is difficult to follow in the practice of building a data repository, since it suggests that something that is uncertain—the potential for re-use—should be a primary factor in determining the value of a data set.

This paper addresses the question of how to assess the re-use value of data, based on research and development from the Data Conservancy initiative where the potential for re-use is considered one of several important criteria for acquisition and allocation of resources for curation. The Data Conservancy (<http://dataconservancy.org/>), based at Sheridan Libraries at Johns Hopkins University, is part of the NSF DataNet program, established to create a network of cyberinfrastructure organizations to "catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable" (NSF-OCI, 2007, p. 2). The aim of the Data Conservancy (hereafter referred to as DC) is to develop a "blueprint for research libraries" and other institutions for the provision of data curation and repository services. As such, it is developing infrastructure and professional processes to meet its mission to collect, organize, validate, and preserve observational data that will allow scientists to address grand research challenges that face society (Choudhury & Hanisch, 2009).

Preservation of observational data is a high national priority, since unlike experimental data they cannot be reproduced (NSB, 2005). "Grand challenge" research has been a concern of U.S. based science policy for several decades, with the concept first introduced in the wake of 5th generation computer development during the 1980's (Stevens, 1994). The term has since been widely adopted to refer to areas of research that have the potential to profoundly impact political and economic dimensions of society (NSF-TFGC, 2011). Scientists tackling grand research challenges, like predicting the impacts of climate change or protecting ecological biodiversity, are increasingly dependent on robust computing capabilities and ready access to high quality data of various types and

scales of analysis. They are expected to produce and consume vast quantities of observational data in the conduct of daily research activities (NSF- CIC, 2007).

For DC to support grand challenge science it needs to efficiently and reliably collect and provide services for an extensive range of highly heterogeneous data products. Therefore, an essential part of the DC initiative was development of a policy to guide the data to be acquired to meet the DC mission. As a research library based initiative, the disciplinary purview is broad, including astronomy and the earth, life and social sciences. As initial test cases representing this scope, the data selected for the first two ingests were extremes on the continuum of big and small science—a highly sophisticated terabyte-scale resource developed by the astronomy community and a complex set of heterogeneous geology data gathered by a single research laboratory over the course of one scientist’s career. Though the types of data and the phenomena represented in these two cases are quite disparate, they both exemplify observational scientific data that records historical, non-recurring events or specific states and conditions at a particular place and time.

The DC collection policy establishes general and specific domain targets. It also articulates a range of criteria by which data are considered eligible for inclusion in DC; considerations include value, uniqueness, risk of loss, funding requirements, and right of deposit. These criteria will not be applied strictly, but rather weighed against the costs of acquisition and subsequent services, prioritizing data considered valuable for re-use for targeted research areas. Not surprisingly, the greatest challenge in developing the collection policy has been articulating criteria for evaluating, or in some sense predicting, the potential for a dataset to be re-used.

Our approach to understanding re-use value draws on core library and information science work related to information seeking and subject representation, specifically Birger Hjørland’s (1997) concept of “epistemological potential.” Below, we elaborate our conceptual framework for analytic potential, but first we provide background on trends in science and related work in information science that inform our perspective and approach. The remainder of the paper presents our conceptual approach to “analytic potential” and evidence from our research on scientific data practices to illustrate key aspects of the conceptual framework, which is guiding collection and curation priorities for DC and has direct implications for data repositories more generally.

BACKGROUND

The Science Data Imperative

While curation and preservation of scientific data are receiving considerable attention from the information science and digital repository communities, relatively

little progress has been made on approaches for evaluating the value of data to inform investment in acquisition, curation, and preservation. The social sciences have historically been at the vanguard of data preservation in response to the need to retain census data and longitudinal national survey data. Now, the demand for trusted data repositories is increasing in the sciences, due to the rise in computationally driven inquiry that depends on networked technologies and persistent access to large amounts of data. Moreover, non-computational scientists across fields will be increasingly seeking repositories for their data, in response to funding agency requirements for formal data management plans (e.g., NSF-DMP, 2010).

Repositories like the DC that are open to many different kinds of data from many disciplines face the challenge of accommodating “big sciences” like astronomy, where a given resource may consist of many terabytes of data and computational tools are often necessary for interpretation. As with other big sciences where sharing of uniform data lends itself to aggregation, visualization, and pattern analysis, the astronomy community is already largely committed to building sharable data resources.

However, by some measures, up to 80% of all science is in the long tail of scientific research made up of smaller, less costly projects (Heidorn, 2008), and this small science activity is expected to produce more data overall than big science (Carlson, 2006). In sciences like geobiology and soil science, for instance, researchers tend to work independently or in small groups, on hypothesis driven research questions, gathering data into privately held collections for local analysis. Currently, these data are rarely shared and re-used, in part because there are no suitable repositories, a problem that stems from the complexity and variation within the practice and culture of small science (Cragin, Palmer, Carlson & Witt, 2010). Small science data have potential for analysis across aggregates, similar to big science. But, their value for re-use may also be complementary, as a unique piece of a complex puzzle or an important addition to a series of measures over time.

A second principle asserted by Uhler (2010) is that “digital resources will not survive or remain accessible by accident” (p. 5). He calls attention to the well-known “memory hole” caused by the inadvertent loss of data from NASA Explorer I, but appropriately emphasizes the continual, distributed, invisible, and irreversible loss of data across science over time. In fact, the mission of most large research libraries, and, by extension, their data repositories, includes the preservation of the scientific and scholarly record. However, strategies for acquiring and retaining data as historical or cultural evidence of the conduct of science will require very different collection strategies and curatorial investment than the problem addressed in this paper—data collected to support current or future analysis to advance scientific research.

Archiving the scientific record calls for broad coverage of the entire enterprise of science, coordinated among institutions to capture the output at large without unnecessary duplication. Data to support active research calls for more selective and strategic coverage, preservation and access services, as well as assurance of quality and usefulness.

Building Data Collections

Uhlir (2010) makes a strong case for the value of data in the public sphere (see also, Arzberger, Schroeder, Beaulieu, Bowker, Casey, et al., 2004), but there has been limited work on policies and processes for collecting scientific data for public access. Archival science and records management have developed rich and useful practices for systematically collecting and appraising digital content (Eastwood, 2004; Whyte & Wilson, 2010), but archival approaches tend to be more applicable to the retention of data for the scientific record than to our focus on supporting re-use (Palmer, Renear & Cragin, 2008). Moreover, there remains a gulf between institutional and scientific notions of a record, a collection, and their representation in a digital form (Esanu, Davidson, Ross & Anderson, 2004). Library and information science collection development principles and criteria, which are driven by assessment of service community needs, may be effectively adapted to data repository practice, but existing collection evaluation techniques will need to be revised for data content.

There remains a general need to extend and develop new models for evaluating data and building data collections, especially to meet the global aims of a robust, functional, and interoperable network of cross-disciplinary data to support grand challenge science (Hey, Tansley & Tolle, 2009; NSB, 2005). At the local level, policies and processes will need to look beyond data that “sit behind” published papers, to acquire data of value for high-priority research questions by the scientific communities served. At the same time, each repository is contributing to the extensive and distributed enterprise as a whole, where data are a “fundamental infrastructural component of the modern research system” (Uhlir, 2010, p. 1; see also, Edwards, Jackson, Bowker, & Knobel, 2007).

The library and information science meta-science perspective articulated by Bates (1999) has always been fundamental to the role of providing broad, useable information collections and services, especially to support interdisciplinary research. To build an adequate base of resources to support grand challenge science, which tends to highly interdisciplinary, collection developers need to understand the landscape of information produced and its roles, interrelationships, and dependencies across fields and generations of production and use. This cross-disciplinary, longitudinal approach to collections will be more pronounced and complex for building data collections than it ever was for literature-based

collections, in part because we can no longer depend on publishers for vetting and acquisition services. Moreover, with digital data, assessments of value will not only require understanding of the content but also its structural and semantic make-up in relation to how analysis will be performed by various service communities.

Data Sharing and Re-use

The need for data to be made more openly accessible has been widely recognized, and there is a growing body of research on data sharing practices in various domains (e.g. Blumenthal, Campbell, Gokhale, Yucel, Clarridge, et al., 2006; Borgman, 2010; Whitlock, McPeck, Rausher & Moore, 2010). Studies of small science, in particular, indicate that successful digital repositories and information services will need to be responsive to the specific data practices of their user communities (Borgman, Wallis, & Enyedy, 2007).

Only a few studies have examined how scientists re-use or re-analyze data and the problems they encounter. In fields such as ecology, for example, it has been shown that standards adoption and metadata development are instrumental for assessing data, but ultimately re-use is contingent on individual field knowledge and established trust between researchers (Zimmerman, 2007). In other words, data re-use is a “context dependent” process. Moreover, cross-disciplinary studies of data sharing suggest that metadata from the perspective of the data producer is not likely to be adequate for representing the value of data for re-use in other fields (Cragin, Palmer, Carlson & Witt, 2010). Not surprisingly, access to data is a key factor in facilitating re-use, according to a study of earthquake engineers, as are integrity of a data set, reputation of the producer, confidence that the data can be easily understood, and the general relevance of the phenomenon being recorded (Faniel & Jacobsen, 2010).

Taken together these studies suggest that indicators of quality and usefulness of the data, as well as the context of data production, are necessary to support re-use. At the same time, meaningful indicators may vary for users with different levels of expertise or from different disciplines, as will be illustrated further in examples presented below.

CONCEPTUAL FRAMEWORK

Epistemological Potential of Documents

Hjørland (1997) discusses the notion of epistemological potential (hereafter referred to as EP) in the context of the subject analysis work involved in cataloging documents for access in an information system. EP refers the range of possible intellectual applications for a given book or article beyond those intended by the author. This is a particularly apt concept in regard to the development of interdisciplinary research collections in an open digital environment where many scholarly and scientific communities have direct access to a rich base of information resources.

According to Hjørland, the metadata representing the subjects of a document 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” (p. 93). A document, however, may have an infinite number of properties capable of informing users, therefore the representation provided should be based on analysis of the document’s possible contributions to various user groups (Hjørland, 1997). Those contributions should then be prioritized based on their “long-term utility” for contributing to the production of knowledge. In producing metadata records for access, these strong contributions are then encoded with categorizations appropriate for the information system, using enduring terms presumably from a controlled vocabulary.

Hjørland’s concrete examples of EP are for the subject analysis of books in History and Psychology. The first demonstrates the utility of a local history of a specific geographic location in Copenhagen between 1880-1920. The book was intended to be a social historical analysis of the dynamics between the social classes of bourgeois mistresses and working class girls, but Hjørland demonstrates the contributions it can make to research on families, sex and prostitution, police-population relationships, and unions for housemaids. His second example illustrates how a book intended to be specifically on the psychology of explanation that documents the decline of a psychological theory may, over the long term, contribute as a much broader treatment of the philosophy and methodology of psychology.

In short, given a particular document, EP is an assessment of possible intellectual contributions for the range of user communities served. Hjørland then emphasizes the need to categorize, or represent, contributions with long-term utility in a form that supports access by those communities (see Figure 1).

Analytic Potential of Data

Data sets, however, are the raw materials of research. As such, they generally are not in a form that “transfers knowledge.” Data must be analyzed before they become

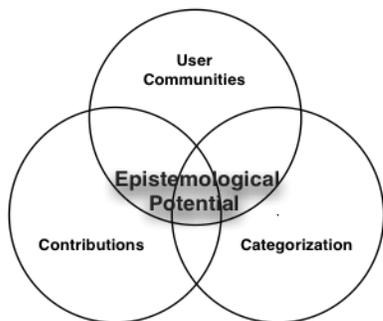


Figure 1. Epistemological potential of a document.

intellectual contributions with EP, like ideas in a book or results presented in a journal article. In other words, with data EP is dependent on analytic potential; the results of the re-use of data with analytic potential have EP. And, like epistemological potential, analytic potential reaches beyond the original intent and domain of the creator.

Analytic potential (AP) is the likelihood that a data set will be of value for future analysis by others, not just for replication, but also for new applications. Like EP, understanding user communities and contributions with “long-term utility” are central aspects of AP. In our framework, long-term utility is tightly linked to two factors that have to do with the condition of the data: preservation readiness (long-term) and fit for purpose (utility). Preservation readiness—the preparation of the data for long-term preservation and archiving—is a fundamental requirement for future use and thus a prerequisite to any further evaluation of re-use value. Fit for purpose is an established and vital area of responsibility for the curation of research data (Lord, MacDonald, Lyon & Giaretta, 2004). It refers to the alignment of a data set with the methods and tools for a given application, and, of course, the potential applications are directly associated with the user communities working on research problems that can benefit from a given set of data.

In short, given a particular data set, AP is an assessment of the possible analytic contributions for the range of possible user communities. Analytic contributions are dependent on the data being preserved and fit for the analytic purpose of a user community. Data with high analytic potential would be applicable to multiple research communities. As with EP, these contributions need to be represented in a form that allows access to the data by multiple communities, but this aspect is outside the scope of our treatment of AP in this paper. The fundamental point is that in assessing re-use value, EP is dependent on a sequence of AP factors.

Preservation Readiness

The criteria for preservation readiness adopted by DC are based on the “preservation description information,” defined by NASA’s reference model for an Open

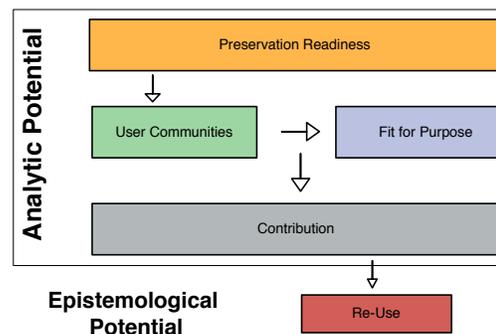


Figure 2. Re-use value of data.

Archival Information System (OAIS, 2002, 2-6). Preservation ready data include representation, provenance, context, reference, and fixity information. Representation combines structural and semantic information to provide meaning for a data object. Provenance describes the source of the deposit and its processing history. Context provides information on why the data were produced and relationships to other objects. Fixity provides for stability of the content within the repository, managed, for instance, through check sums over the digital information packages.

These relatively routine expectations for data are well documented in the digital repository literature (OAIS, 2002; RLG-OCLC TRAC, 2002). However, they are not widely understood by scientists and are best generated through interactions between data curators and data producers. Therefore, preservation readiness processing is an important part of curatorial services and provides the first step in determining the analytic potential of a data submission by assessing the initial condition of a data set for preservation and estimating the resources required for assuring long-term preservation.

Fit For Purpose

Preservation readiness is closely related to fit for purpose. In fact, the common notion of fit for purpose for data curation is concerned with data quality for the intended use, in line with the primary goal of preservation. In the context of AP, however, fit for purpose needs to be extended to new applications. For example, raw data may be preferred for re-use purposes over data that have been transformed or derived as part of a previous analysis process. Or, in some cases, part of a data set may need to be re-used, requiring decomposition of a compound object or aggregation. Assuring that data are fit for purpose may also consist of tracing statistical calculations back to the tools and methods used in generating the data. This type of authentication is related to the idea of “appropriate use,” as described by Parsons and Duerr (2005).

Appropriate use involves verifying the credibility of a data set to be accurately and meaningfully re-used given what it purports to measure or represent. The case described by Parsons and Duerr is consistent with our interpretation of fit for purpose for AP. The data consist of 38 sets (mostly images) on sea ice concentration, derived from passive microwave remote sensing (p.34). They are held by the National Snow and Ice Data Center (NSIDC <http://nsidc.org>), which has created a context for the data sets through linkages to relevant scientific literature, special reports devoted to a “deeper analysis of passive microwave derived sea ice products,” and a website that provides access to these products (p.34). Given the granularity and the time series these images represent, the data are “fit” for answering a certain range of scientific questions. They are not fit for questions that require detailed observations, such as those needed by a

biologist tracking polar bear migration where definite locales and exact daily time scales would be necessary (p. 34).

Clearly, it is not possible to anticipate all possible re-use options for a data set, and ruling out applications may be easier than identifying promising new ones. However, an AP evaluation is concerned with determining, as far as possible, the range of scientific questions, and the associated communities, to which a data set can contribute and how much investment is required in making it fit for that range of purposes.

Potential User Communities

Fit for purpose is directly associated with a user community, their research questions, and analysis techniques. Identifying the range of potential communities that might find a data set useful over time requires the meta-science expertise of information professionals discussed above, since data producers are not well positioned to assess the extent to which their data may contribute to the work of other researchers (Cragin, Palmer, Carlson and Witt, 2010; Baker & Bowker, 2007). Observational data, as an immutable source for recording an event (NSB, 2005), are recognized as having strong potential for re-use beyond their discipline of origin. For example, some fields use observational data from numerous sources and disciplines to create complex data sets or integrate a range of data for simulations and modeling (Sundberg, 2011).

The potential reach of data beyond its intended use (Chao 2011) is not articulated by the notion of the designated community promoted by OAIS (2002), which has been widely adopted as the de-facto standard for developing service based digital libraries and preservation repositories. A designated community is defined by OAIS as “an identified group of potential consumers who should be able to understand a particular set of information” (OAIS, 2002, 1-10). What a given designated community is able to understand without, “the assistance of the experts who produced the information” is considered their Knowledge Base (OAIS, 2002, 3-1), which needs to be understood (or monitored) to provide services that satisfy the needs of a designated community. By these standards, a data set that is collected by a repository must either contain, or be curated to include, all of the necessary tools, software packages, and contextual information necessary for a designated community to meaningfully use that data.

This definition of designated community is not well aligned with the DC context of “discipline crossing,” grand challenge research. It is most applicable to data repositories developed to archive particular kinds of data for a primary user community with shared data practices. In contrast for example, within the DC we expect small science observational data collected in the field to be of interest to climate change modelers. Since these data are

complex, heterogeneous, and often recorded in non-standard formats, they are likely to be inconsistent with the knowledge base and tools of an entire modeling community. Nonetheless, they may hold enough re-use value to justify acquisition, preservation, and curation. The challenge is determining if and how much potential there is for re-use and how much investment should be put into curation to fill the knowledge base gap and improve the fit for new purposes. Ultimately, the success of DC in supporting grand challenge science will depend on efficient curation of high quality data for numerous, diverse user communities, which may require varying levels of curation for the same data for different communities and purposes.

At present, our research on data practices, presented in the next section, serves as the primary means for understanding the knowledge base and data needs of the broad range of DC user communities. Our studies examine practices and the spheres of context around data production and use (Baker & Yarmey, 2010) for the range of targeted scientific domains associated with DC. This landscape view of data and practice is providing evidence of potential user communities for certain types of data, which is informing the work of the infrastructure team in developing preservation processes and anticipating fit for purpose demands for re-use.

Methods

The domain profiles discussed in this section are based on our ongoing empirical research on disciplinary differences in data practices across the domains expected to be early data contributors to DC. Our unit of analysis is a research sub-discipline, based on our previous work indicating that this level best captures small science data practices and evidence of re-use value (Cragin, Palmer, Carlson & Witt, 2010). To date we have engaged with twenty participants who are active researchers in geology, oceanography, and environmental science, focused on areas such as climate, water, soil, and magma. We use a sequenced, multi-method approach for data collection, employing multiple semi-structured interviews, data inventorying, and artifact analysis, which work together to produce dense, high-quality case study units of evidence (Cragin, Chao & Palmer, 2011).

A Pre-Interview Worksheet is used to orient participants to our specific interests in their data, setting the scene for the questions that follow in the Research Interview. The worksheet responses often provide important domain-specific information necessary for the curation process. They also facilitate deep discussion on the participant's research practices. A subsequent Follow-up Interview is used to clarify or address gaps. This interview has been especially important for probing on specific data types identified by participants as having value for other users

and to document deposition requirements and curation needed to support re-use.

Like most case study research, the volume or density of each case varies, and the rigor emerges from the systematic, iterative approach across cases. The sequencing and multiple modes of interaction allow us to identify overlaps in research problems, methodologies, and types of data associated with user communities, and to conduct deeper analysis to address the important "how" and "why" questions about data needs and uses (Cragin, 2009). Comparative cross-case analysis is ongoing, but here we draw on initial within-case analyses to demonstrate aspects of AP across areas of earth science, and in relation to particular geophysics data sets, specifically in the area of volcanology. Interview excerpts have been slightly refined to improve readability, while references to scientists are obscured to maintain participant confidentiality.

ANALYSIS

Earth Science Précis

The case studies being developed are quite deep, covering a range of information aimed at informing curation and infrastructure development, but they are also useful for basic AP profiling. Table 1 (below) outlines a small sample of three sub-disciplines in the earth sciences, covering primary data types, the designated community, potential user communities, and re-use scenarios.

The first three rows in the table illustrate basic aspects of AP drawn from our data, beginning with a list of data types and materials (data objects) that given their preservation readiness qualify for ingest and need consideration for fit for purpose. For Geobiology, many types of data are collected to investigate the ways that microorganisms influence the geology of the earth and how earth environments influence the behavior of the microbes. For Volcanology, the primary sources are physical rock samples, from which a range of analog and digital data are generated. In Soil Ecology, data are collected on target species, as well as environmental variables, much of which is gathered to support interpretation of the primary biological data. The Designated Community presented in the second row represents the domains of researchers who generated the data, and the Potential User Communities in the third row are additional researchers with interests in the kinds of observations represented in the data. The fourth row provides examples of the value of the data for re-use based on our studies and further engagement with scientists in Geobiology. In the following section we discuss one particular case, the volcanology group, to explain how our formulation of the concept of analytic

	Geobiology	Volcanology	Soil Ecology
Data objects (original analytic purpose, for preservation)	Site-specific time series: -“reduced spreadsheets” of averaged rock, water chemistry measures; - microscopy images; - annotated field photos; - microbial genomics	Rock profile - physical rock - thin section - chemical analysis - photographs (35mm slides; digital) - field notes and maps	Database, work and soil samples - multiple abiotic soil measurements - sensor and network function - associated metadata - soil composition - worm population counts
Designated Community	Geobiology Geology (general) Microbiology	Igneous petrology Geophysics Geochemistry	Biogeochemistry Earthworm ecology Soil Science
Potential User Communities	Evolutionary biology Bioprospecting U.S. Park Service Public Health	Glaciology	Biodiversity Environmental sciences
Examples of Re-sue Potential	Microbial data can be analyzed to assess the presence and extent of disease.	Field photos with spatio-temporal stamp can be used to assess changes to glaciers over time.	While spatially sparse, the taxonomic, soil carbon, and ground cover data are sampled over time and can be useful indicators of local environmental change.

Table 1. Data with analytic potential from three sub-disciplines.

potential has grown out of our study of data practices. Specifically we highlight the attributes of preservation readiness resulting from our artifact inventory, and aspects of fit for purpose and domains of interest emerging from our interviews with scientists.

Volcanology Example

Due to extremely low humidity and ice cover, geological structures in Antarctica have undergone relatively little erosion since their formation millions of years ago, providing a unique opportunity for gathering igneous rocks and studying magmatic differentiation. A seminal figure in this field, Reginald Aldworth Daly stated early in the 20th century: “A final philosophy of earth history must be largely founded upon the unshakeable facts known about igneous rocks” (1933, p. 1). Formed from the cooling of magma, the study of igneous rocks and magmatic flow are fundamental to understanding the history and formation of the earth’s crust (Marsh, 2004). For volcanologists, the Antarctic site under study provides access to rock that represents a sort of time series of historical magmatic events—the geological structures are exposed but well preserved, as if someone had sliced vertically through a volcano to show its layers and extended plumbing system.

This large trove of data from Antarctica is being processed for ingest into the DC repository. The research group that produced this collection has made over ten data gathering trips to the region since the early 1990s, resulting in an abundance of physical materials: literally tons of igneous rock samples, finely cut cubes (or billets), billet slices mounted on glass slides (thin sections), powdered rock, as well as supplementary field data. The powdered rock samples are used to produce “bulk rock analysis,” which results in tabular data of the chemical

composition of the samples; additional digital data from the field include photos and field notes, born-digital and digitized images from the thin section slides, resulting quantitative analysis of their contents, and recently, 3D maps of the region under study.

Preservation Readiness

Acquisition of this special collection has required creation of a detailed inventory identifying the several data types and their specifications, and then processing of analog and digital materials in anticipation of the ingest process. DC repository staff and data scientists will conduct integrity assessments for the collection, following OAIS ingest guidelines for verifying and transforming submission information packages into archival information packages.

Beyond requisite descriptive metadata, this collection requires documentary evidence related to provenance, reference, and context. To identify locations where rocks were gathered, samples are marked with an identifier that signifies the specific field campaign and the collector. All derived samples and subsequent digital data are labeled with this identifier, which is then used in the records created for the object. Recording accurate metadata for collected rock samples is crucial for accurate analysis in the laboratory setting and for validating findings disseminated to the geophysical community.

In addition to the sample number (identifier), descriptive information, such as date, time, weather condition, surrounding environment, sill position, GIS coordinates, and the condition of the specimen when collected, all contribute to preservation readiness for these data. Much of this information has been recorded in field notebooks produced during data gathering trips, which becomes part of the complete data package.

Fit for purpose

The details in the field notes are similar to a map's legend—they give an orientation to the collected samples and guide interpretation of content. Additionally, the narratives include numerous clues as to why a certain oddity or outlying data point may appear in later chemical analysis. For instance, one participant described the value of the record of the data gathering conditions for accurate re-use in situations where abnormalities occur:

...if you see an outlier, you have to wonder, 'ok what's going on there,' so you go check the field notes, well if the field notes don't have anything written down about 'oh this sample was really weird because of these characteristics' or something, then they really have nothing to draw on to help them understand what at all is going on.

In evaluating fit for purpose for this collection, completeness will be an important factor to assure access to the compilation of products necessary to facilitate accurate re-use. Use of the digitized thin sections, for example, generally requires a minimum of three images—the plane polarized view, and then two cross-polarized images generated at different angles to ensure that the minerals present are visible. These data will only be re-usable for volcanology purposes if the field notebooks account for specimen gathering conditions, if the thin-sections are digitized in a way that allows for crystals to be accurately judged, and if the chemical analysis includes the documentation of any normalization performed on the original data set.

However, another domain might not require such specific field gathering information. The same participant noted that while contextual information was essential for his work, it might not be necessary for other areas of research:

...somebody for example who's more knowledgeable about isotopes than I am can take the data that I produced and do a whole different series of investigations that I would never do myself because it's not really even my field. There's a lot of geochemical work that's done that relies less on field context.

Clearly, fit for purpose may vary with the application by the user community and new applications might emerge over the course of the data lifecycle.

Potential User Communities

For a broad based repository like DC, identification of potential user communities is complex, but will need to precede analysis of, and investment in, making data fit for purpose. A solid understanding of the scope and extent of the user communities will be required to generate accurate cost estimates for curatorial activities. For the volcanology data, interest in the data will be associated with the location, time period, and phenomenon of study that the collection represents.

The chemical analysis component of the data is a useful example of a data product generated for one purpose that is likely to have broad appeal beyond the original research community. One scientist in this lab noted that the chemical analysis of an igneous rock often has numerous fields of data that their laboratory is not concerned with, but which might be of interest to someone from another discipline:

...there are people who might work on little iron and titanium oxides which I don't really care about; there might be a couple of them in my rocks, but I'm not looking at that...they might be really interested in that and want to take the data that I've provided and study that, and it might be a completely different thing than I would ever do with it.

The chemical data represent a unique phenomenon with implied value for other domains. For DC, this may justify allocation of resources for preservation and curation to extend fit for purpose only for this segment of the larger data collection.

Summary

The volcanology group discussed above conducts site-specific research that generates rare observational data. DC has prioritized these data because of the high cost of data gathering, the unique location studied, and the importance of the phenomena investigated. But, there is still a need to understand AP as an additional criterion for investment in curation. Evaluation of AP accounts for the conditions of the data for preservation and fit for purpose for the designated community, but understanding the potential user communities that can make new use of the data is the foremost challenge

At present, our pre-interview worksheet is capturing information that informs the types of quality evaluations necessary for a given data type and the designated community, but the staged interviews are instrumental in developing a fuller understanding of the potential value of data for re-use in new communities. Fit for purpose assessments will need to be iterative and evolve around evidence of contextual dependence for appropriate use by new users. As our knowledge grows about the qualities and types of data produced from DC service communities, identifying data of value for broader consumption will become more systematic and routine, and will be a vital part of the growing base of curatorial expertise.

CONCLUSION

There is no doubt that digital scientific data will not survive or become accessible without repositories dedicated to preserving these raw materials of research. And, if the value of data does, in fact, increase with their use, then repositories that commit to collecting data with analytic potential could become among the most significant resources for the production of new scientific

knowledge. However, for this to happen repository developers will need to take an active approach to building data collections, encouraging sharing but also recruiting data of value for investigation of high-priority research questions by the scientific communities they serve and perhaps by those they that were previously considered out of scope.

As we continue to elaborate and refine our framework and techniques for evaluating the re-use value of data, we will integrate the outcomes with established curatorial techniques to facilitate meaningful cross-disciplinary re-use of DC data and develop systematic processes that can be shared with information professionals involved with data collection and curation in research libraries and repositories. There are complications inherent in using analytic potential as a framework, including the large investment in understanding the data practices of both data producers and consumers. But, as was true with bibliographic materials in traditional research libraries, collections and services for data sources need to be understood as part of the broad cross-disciplinary epistemological trends and socio-cultural dynamics of the overall research enterprise (Hjørland, 1998, p.618).

Preventing “memory holes” in the scientific record will require a different approach to data collection and curation that can capture data at the point of production, since gaps arise when there are discontinuities in stewardship. Our strategy for this is to develop next a site-based model of curation principles and processes for channeling curated series of data into the repository for preservation and access. There is tremendous need to curate for preservation and fit for purpose for the designated community at the site of data production, thus freeing up repositories to concentrate on aggregating data and adding value for re-use. This division of labor would promote progress on both the retention of scientific data and facilitating research re-use, toward a future system of scientific production that makes efficient and innovative use of all the assets of research.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (grant # OCI-0830976) and draws on previous work supported by the Institute of Museum and Library Services (LG-06-07-0032-07). We wish to thank Tiffany Chao for her extensive contributions to the Data Practices team and members of the Data Conservancy Collection Policy group, especially Ruth Duerr for her insights on designated communities.

REFERENCES

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laakosonen, L., Moorman, D., Uhler, P., & Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3, 135-152.

- Baker, K. S. & Bowker, G. C. (2007). Information ecology: Open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems*, 29, 127-144.
- Baker, K. S. & Yarmey, L. (2009). Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation*, 4(2).
- Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, 50(12), 1043-1050.
- Blumenthal, D., Campbell, E., Gokhale, M., Yucel, R., Clarridge, B., Hilgartner, S., et al. (2006). Data withholding in genetics and the other life sciences: Prevalences and predictors. *Academic Medicine*, 81(2), 137-145.
- Borgman, C. (2010). Research data: Who will share what, with whom, when, and why? China-North America Library Conference, Beijing. Retrieved on May 30, 2011 from: <http://works.bepress.com/borgman/238>
- 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.
- Carlson, S. (2006). Lost in a sea of science data. *Chronicle of Higher Education*, 52 (42), A35.
- Chao, T. (2011) Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. Proceedings of the 2011 American Society of Information Science and Technology. October 9-11, New Orleans.
- Choudhury, G.S. & Hanisch, R. (2009). Data Conservancy: Building a sustainable system for interdisciplinary scientific data curation and preservation. *PV 2009 Conference*, Madrid, Spain. Retrieved on May 30, 2011 from: <http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/talks/20091203-1445-Choudhury.pdf>
- Cragin, M. H. (2009). Shared scientific data collections: Use and functions for scientific production and scholarly communication. Unpublished doctoral thesis. University of Illinois.
- Cragin, M. H., Chao, T. C., & Palmer, C. L. (2011). Units of evidence for analyzing sub-disciplinary difference in data practice studies. *Proceedings of the Joint Conference of Digital Libraries*. June 13-17, 2011, Ottawa, Canada.
- 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, 4023-4038.
- Consultative Committee for Space Data Systems (2002). Reference model for an Open Archival Information

- System (OAIS). CCSDS 650.0-B-1 Blue Book, January 2002.
- Daly, R. A. (1933). *Igneous Rocks and the Depths of the Earth: Containing Some Revised Chapters of "Igneous Rocks and their Origin."* McGraw-Hill, New York.
- Edwards, P. N., Jackson, S. J., Bowker, G., & Knobel, C. P. (2007). Understanding infrastructure: Dynamics, tensions, and design. Report of a workshop on history and theory of infrastructure: Lessons for new scientific cyberinfrastructure. Retrieved May 30, 2011 from <http://deepblue.lib.umich.edu/handle/2027.42/49353>
- Eastwood, T. (2004), Appraising digital records for long-term preservation. *Data Science Journal*, 3(0), 214 - 220.
- Esanu, J., Davidson, J., Ross, S., & Anderson, W. (2004). Selection, appraisal, and retention of digital scientific data: Highlights of an ERPANET/CODATA Workshop. *Data Science Journal*, 3 (0), 227-232.
- Faniel, I. M. & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, 19 (3-4), 355-375.
- Heidorn, B. P. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, 57, (2).
- Hey, T., Tansley, S., & Tolle, K. (2009). The 4th paradigm: Data-intensive scientific discovery. Microsoft Research, Redmond, WA. Retrieved on May 30, 2011 from: <http://research.microsoft.com/enus/collaboration/fourthparadigm/>
- Hjørland, B. (1997). Information seeking and subject representation: An activity-theoretical approach to information science. Westport, CT : Greenwood.
- Hjørland, B. (1998). Theory and metatheory of information science: A new interpretation. *Journal of Documentation*, 54, 606-621.
- 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.
- Marsh, B.D. (2004) A Magmatic Mush Column Rosetta Stone: The McMurdo Dry Valleys of Antarctica. *EOS Transactions American Geophysical Union* 85 (47) 497-502.
- National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Draft Report of the National Science Board. Retrieved on May 30, 2011 from: www.nsf.gov/pubs/2005/nsb0540/
- National Science Foundation. (2007). Sustainable Digital Data Preservation and Access Network Partners (DataNET): Office of Cyberinfrastructure, Program Solicitation. Retrieved on May 30, 2011 from: <http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm>
- National Science Foundation. (2007). Cyberinfrastructure Vision for 21st Century Discovery: A Report to the NSF Cyberinfrastructure Council. Retrieved on May 30, 2011 from: http://www.nsf.gov/pubs/2007/nsf0728/index.jsp/nsf11001/gpg_2.jsp#dmp
- National Science Foundation. (2010). Grant Proposal Guide; Chapter II C.2.j. Retrieved on May 30, 2011 from: <http://www.nsf.gov/pubs/policydocs/pappguide>
- National Science Foundation. (2011). Advisory Committee for Cyberinfrastructure; Task Force on Grand Challenges: Final Report. Retrieved on May 30, 2011 from: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_GrandChallenges.pdf
- Parsons, M. A. & Duerr, R. (2005). Designating user communities for scientific data: Challenges and solutions. *Data Science Journal*, 4, 31-38.
- Palmer, C.L., Renear, A.H., Cragin, M.H. (2008). Purposeful curation: Research and education for a future with working data. 4th International Digital Curation Conference, December 1-3, 2008, Edinburgh, Scotland.
- RLG-OCLC Working Group on Digital Archive Attributes. (2002). Trusted digital repositories: Attributes and responsibilities. Mountain View, CA: Research Libraries Group (RLG). Retrieved on May 30, 2011 from: <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>
- Stevens, R. (1994). High performance computing and communication. *Future Generation Computer Systems*, 10 (2-3), 159-167.
- Sundberg, M. (2011). The dynamics of coordinated comparisons: How simulationists in astrophysics, oceanography and meteorology create standards for results. *Social Studies of Science*, 41 (1), 107-125.
- Uhlir, P. F. (2010). Information gulags, intellectual straightjackets, and memory holes: Three principles to guide the preservation of scientific data. *Data Science Journal*, 10, 1-5.
- Whitlock, M. C., McPeck, M. A. Rausher, M. D., Rieseberg, L., & Moore, A. J. (2010). Data archiving. *American Naturalist*, 175 (2), 145-146.
- Whyte, A. Wilson, A. (2010). Appraise and select research data for curation. DCC How-to-Guide for digital curators. Retrieved from : <http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data>
- 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.