Library Trends 57(2) Fall 2008 (in press). Institutional Repositories: Practice, Current Research, Data Collection, Future (final title TBD). Edited by Sarah Shreeves and Melissa Cragin.

Shedding Light on the Dark Data in the Long Tail of Science

P. Bryan Heidorn

Abstract

One of the primary outputs of the scientific enterprise is data, but many institutions such as libraries that are charged with preserving and disseminating scholarly output have largely ignored this form of documentation of scholarly activity. This paper focuses on a particularly troublesome class of data, termed *dark data*. "Dark data" is not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost. The article discusses how the concepts from long tail economics can be used to understand potential solutions for better curation of this data. The paper describes why this data is critical to scientific progress, some of the properties of this data, as well as some social and technical barriers to proper management of this class of data. Many potentially useful institutional, social and technical solutions are under development and are introduced in the last sections of the paper, but these solutions are largely unproven and require additional research and development.

Background

The majority of the work being done by scientists is conducted in relatively small projects with one lead researcher with part-time commitment to the project and perhaps two or three graduate students or part-time staff scientists. The raw product of these efforts is scientific data, the data that forms the foundation of all of scientific theory. While great care is frequently devoted to the collection, preservation and reuse of data on very large projects, relatively little attention is given to the data that is being generated by the majority of scientists. New social structures and technical developments could greatly increase the availability and value of individual scientists' data and related research. We can organize science projects along an axis from large to small. The very large projects supporting dozens or more scientists would be on the left side of the axis and generate large amounts of data, with smaller projects sorted by decreasing size trailing off to the right. The major area under the right side of the curve is the long tail of science data. This data is more difficult to find and less frequently reused or preserved. In this paper we will use the term *dark data* to refer to any data that is not easily found by potential users. Dark data may be positive or negative research findings or from either "large" or "small" science. Like dark matter, this dark data on the basis of volume may be more important than that which can be easily seen. The challenge for science policy is to develop institutions and practices such as institutional repositories, which make this data

useful for society.

When asked, almost all scientists will quickly acknowledge that they are holding dark data, data that has never been published or otherwise made available to the rest of the scientific community. An example of dark data is the type of data that exists only in the bottom left-hand desk drawer of scientists on some media that is quickly aging and soon will be unreadable by commonly available devices. The data remains in this dark desk drawer, inaccessible to the scientific community until the scientist retires. At the point of retirement some scientists rush to find a more suitable home for their data be they in the form of slides, photographs, specimens or electronic media files. More often than not, even in a well planned retirement the desk drawer is eventually emptied into a dumpster because no one including the scientist knows exactly what the data is since it lacks adequate documentation.

Many factors drive the need to pay closer attention to the long tail of science, including the growing number of scientists globally and the increase in the amount of data each scientist can generate with modern instrumentation. This vast growth in the collection of data does not in any way insure that the data is accessible now or that it will be accessible in the future. It has always been the case that scientists have generated more data than they eventually publish but as discussed below new social structures and rapidly expanding information management tools are making new modes of science data management possible.

Chris Anderson (2004) popularized the economics of the long tail in Internet commerce. Some of the same information properties and tools discussed in Internet economics apply to scientific data. Before the Internet, stores like Blockbuster™ rented movies from physical store fronts. The physical inventory was limited by the cost of space. The stores would stock only titles that would rent frequently enough to justify the storage space. Blockbuster concentrated on the head of the long tail graph (see figure 1). Less frequently viewed movies were not easily available to people in the area of the Blockbuster store. Customers may not even know of the existence of films that they would like to see because they did not see them in the local Blockbuster. These films were essentially dark data. Netflix and the Internet changed these economics by separating inventory from the point of sale. To the surprise of many, it turned out that there was a great deal of value in the rarely viewed movies. While there may only be a few dozen or hundred people interested in seeing a boutique title in a particular year, there are many thousands of such rarely viewed titles. Search tools and the Internet allowed people to find and rent boutique films and bring them to light and their television screens.

The long-tail phenomenon can repeat itself in science data. There may only be a few scientists worldwide that would want to see a particular boutique data set but there are many thousands of these data sets. Access to these data sets can have a very substantial impact on

science. It seems likely that transformative science is more likely to come from the tail than the head. For the most part, by the time large-scale projects that generate high volume data are developed, the questions to be answered are relatively well understood. The long tail is a breeding ground for new ideas and never before attempted science. Improbable and risky projects are less likely to attract large grants if they can get any grant at all. (Peck, 2008). A parallel can be seen in bibliometrics where high impact articles are not necessarily found in high impact journals (Seglen, 1997; Sun & Giles, 2007).

Of course science data is different from DVD rentals and offers unique challenges and opportunities. Searching for DVDs was relatively easy: People could search for titles, favored actors or actresses, directors, genre and a few other descriptors. The format of the returned items is also relatively limited with VHS, DVD or Blu-ray. Determining the fields that are needed to effectively describe and index scientific data is much more challenging. Once an interesting data set is found, scientists must deal with apparently infinite variability in the format of different data sets. These challenges require new practices and new technologies for data handling.

Figure 1: Distribution of NSF Awards by Dollar Value

The long tail of science and science's dark data share the same economic and social constraints. There is a wealth of science data that is almost impossible to see. This is science's dark data. We can find much of this dark data in the long tail of science data. Because it is difficult to find dark data it is underutilized and routinely

lost. With appropriate planning and technology this data can be brought to light and made more useful to the scientific community. For an initial analysis of the scope of this issue, we can sort and graph science projects by the volume of data that they generate. At the left side of the graph in figure 1 are projects such as super colliders and Earth observing satellites that routinely collect hundreds of gigabytes of data each day. On the right side of the graph are science projects that produce small data sets. We hypothesize that this data collection size distribution follows many of the other distributions in nature and human endeavor, such that it follows a power law distribution[i].

The characteristics of the data sets in the left, high volume head of the graph differ from the characteristics of the data on the right, low volume tail of the graph. The high volume data tends to be well planned, well-curated and highly visible to scientists worldwide. Data gathering at the head of the graph tends to be highly automated with specialized instrumentation. Data on the right, in the tail of the graph, tends to be less well planned, more poorly curated and less visible to other scientists. The graph's tail also contains a higher proportion of dark data.

A study of the size of research awards granted by the U.S. National Science Foundation (NSF) is a measure of the distribution of size of research projects. Figure 1 was created by sorting all research grants awarded by NSF in 2007 based on the dollar value of the grant. Grants range in size from multi-million dollars to just several hundred dollars. While we do not know the quantity or type of the data generated by each of these projects, we can assume that each dollar of investment in scientific research does generate some relatively constant amount of data. If this is true, then a plot of data generation by science projects would have roughly the same shape as the funding curve in Figure 1 and the individual projects would be in approximately the same location relative to one another. Table 1 provides a breakdown of NSF funding in 2007 in terms of the 80/20 rule. This can be viewed both the dollar value of the largest 20 percent of grants and as the number of grants that are needed from the high end of the curve to account for 20 percent of the revenue. There are a total of 12,025 grants in 2007 awarding a total of just under $2.9 billion.[ii] The top 20 percent of grants, 2405 grants received a little over 50 percent of the total funds. If viewed not by the number of grants but by the percent of the dollar value, the top 254 grants received 20 percent of the revenue. That is 2 percent of the largest grants received 20 percent of the total amount awarded. We argue that the top 20 percent measured either way has better curated data.

Total Grants over $500	12,025 $2,865,388,605	
	20% by number of grants	80% by number of grants
Number Grants	2404	9621
Total Dollars	$1,747,95,7451	$1,117,431,154
Range	$38,131,952- $300,000	$300,000- $579
	20% by total value = $573,077,721	80% by total value = $2,292,310,884
Number of grants	254	11,771
Range	$38,131,952- 1,034,150	1,029,9984- $579

Table 1: Grant Size Distribution

A special issue of Nature (2008), titled "Big Data", provides a good overview of current successes and challenges in managing large data sets. However, an important observation is that most scientists work on the right side of the graph on smaller research projects usually generating small data sets as can be seen in table 1. Because of the length of the tail, while the data volumes are small when viewed individually, in total they represent a very significant portion of the country's scientific output. In fact, the frequently used term big science is somewhat misleading. Many smaller science projects in the tail are actually intellectually interlinked efforts running under a distributed funding model. While GenBank and Long Term Ecological Research (LTER) site data repositories are centrally financed, the research projects that provide their data are independently financed and certain aspects, but not all, of the resulting data are pooled, in essence creating big science from organized small science. The percentage of data from the tail that is collected into well-managed repositories is unknown. While it is critical to continue to curate the data in the head of the graph, it is important to improve curation of the data in the tail. Curation is an old concept that is being applied to digital information. Digital curation is the management and appraisal of data over the life-cycle of scientific interest. "[C]uration embraces and goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage: placing emphasis on publishing data in ways that ease re-use and promoting accountability and integration" (Rusbridge et al., 2005). To accomplish this level of accessibility, the data in the long tail of

science requires different curation practice than monolithic large volume data sets. These curatorial differences are discussed in the sections below on barriers and solutions to data access.

Other researchers have pointed out properties of scientific data related to the long tail. For example, John Porter makes the distinction between deep and wide databases (Porter, 2000, p 63). "Deep" databases specialize in a few data types making up relatively homogeneous collections of data and allowing the development of sophisticated search tools. "Wide" databases collect many types of data, making tool development to deal with the data much more difficult. We speculate that the head of the data curve might tend to hold more "deep" data sets, since much of this information is collected with dedicated instrumentation. The tail tends to be much more heterogeneous as a whole, but the tail may also contain many projects working with similar data types. This scattered but similar data in the tail represents an opportunity - which some fields of science have already capitalized on - to collect data into "deep" data collections; this is discussed below in the section on technical solutions.

There is another previously used definition of dark data, which defines it as the unpublished data of "failed" experiments (Goetz, 2007). In this use of the term failed refers not to bad science but to the fact that only positive results tend to be published. Experiments that accurately demonstrate no effect of the treatment condition are valid findings but are less likely to be published. The data therefore becomes "dark data" and later meta-analyses of the literature provide a skewed view of the actual scientific findings. This definition of dark data is subsumed by the broader definition used in this paper. While such unpublished data indeed are difficult to find (and therefore "dark"), there are many types of "positive" research findings and raw data that lie behind published works which are also difficult or impossible to access as time progresses.

Long tail science is not synonymous with small scientific questions or even small science. The results of multiple projects in the tail can contribute to truly big data, grand accomplishments and accumulated knowledge if handled properly. A prototypical example of this type of research is biomolecular biology projects that contribute to GenBank and the Protein Data Bank (PDB). Principles established by funding bodies such as the National Institutes of Health and the publishing industry have fostered accumulation of large collections of genomic and protein data respectively. This data has helped fuel rapid development within these fields. This type of molecular data is relatively simple and standards commensurately easy to establish. While other data may be more difficult to organize, the collection of this data may lead to advances that we cannot easily predict ahead of the collection.

Dark data exists throughout the fabric of science. While some may indeed be incorrect data that should be discarded because of mistakes,

a substantial amount of time and money is spent to collect potentially useful data that is then underutilized. Much of this dark data resides in the long tail of science, in innumerable projects and labs. The next section will focus on the importance of this data. The following section will examine some of the properties of data in the tail to better understand the opportunities and problems at hand. The remainder of the paper will explore barriers and the institutional and technical solutions that can make this data more useful. Finally the paper addresses some of the outstanding research questions that must be answered before science will be able to make fuller use of the raw fruits of its efforts.

The Dark Data in the Tail Matters

The data in the long tail is an important resource for science. Most data generated and collected in science is important to the scientific process of theory development and evaluation. This fact is understood by the popular culture as expressed even in crime mysteries. "I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. (Sherlock Holmes, in "A Scandal in Bohemia" by Sir Arthur Conan Doyle). In addition, science is built on the principle of replicability. Independent researchers should be able to collect data and analyze it to produce similar results. If the results of prior research are unavailable, or only available in a highly abstracted form, then replication is difficult.

The history of science is a reflection of the history of communication of knowledge and is based on reproducible data. If evidence for a theory cannot take the form of replicable data, then the theory is in question. Theories that are not testable with verifiable data are not scientific theories but rather unsubstantiated belief systems. Scientists live in a world of changing and evolving theory with new or refined theories replacing previously accepted theories only on the evidence of hard[iii] data collected using well documented methods. The availability of the data behind experiments helps to insure scientific integrity by keeping the process open to external evaluation. The data itself is often too voluminous or varied for humans to understand by looking at the data in its raw unprocessed form, so scientists use graphs, charts, mathematical equations and statistics to "explain," "describe" or "summarize" the data. These representational tools help us to understand the world around us. The use of data simplification and data reduction methods in science is repeated at all scales of natural phenomena from the subatomic to the physics of our human scale world, to the function of a cell, a mating behavior of birds or the functioning of ecosystems. But these summary representations of data rely on the underlying data, and the published papers do not capture the richness of the original data and are in fact an interpretation of the data. If the dark data in the tail is not selectively encoded and preserved, then the underpinning of the majority of science research is lost.

The number of journal articles and the number of scientists is growing globally (Mabe & Amin, 2001), so we can infer that the amount of dark data is also growing. As the world economy expands along with the expansion in the number of scientists and science findings, it becomes increasingly unlikely that traditional publishing, word-of-mouth and ephemeral project Web pages will lead to sufficient data preservation and data sharing. The limited increase in publishing slots in comparison to the vast increase in data may be exasperating the situation. "Only a small proportion of the explosively expanded output of biological laboratories appears in the modestly increased number of journal slots available for its publication, even if more data can be compacted in the average paper now than in the past." (Young, Ioannidis, & Al-Ubaydli, 2008).

While some data can be discarded because it can be generated again by replicating the conditions of the original data gathering, the fact that data can be regenerated does not necessarily mean that data should be discarded. One reason to keep replicable data is economics. Here the interests of the individual scientists and the interest of the society of science may diverge. Once the originator of scientific data has exercised its utility through theory validation and publishing, he may decide that the data no longer has value and therefore discard it. However, from the perspective of the larger scientific community this data may have value and utility above the cost required to preserve it. Some data can not be duplicated because it records unique events. Finally, data may contain a hidden signal that is lost in the summary statistics. For example, some genes are expressed through protein production following circadian rhythm. Special procedures are required to discern which fraction of the expression is due to circadian rhythm and which part of the protein production is due to other processes. As a consequence the signal can be overlooked (Refinetti, Cornélissen, & Halberg, 2007).

The Properties of Dark Data in the Tail

There are a number of important features that distinguish data in the head and data in the tail of science as presented in table 2.

Head	**Tail**
Homogeneous	Heterogeneous
Mechanized	Hand Generated
Uniform Procedures	Unique Procedures
Central Curation	Individual Curation
Disciplinary and Reference Repositories	Institutional Repositories
Maintained	Not Maintained
Open Access	Obscured or Protected
Immediately Reused	Seldom Reused
Make Careers	Currently Unnoticed

Table 2: Differences between Head and Tail Data

As noted by Porter (2000), high volume data, such as those that appear in the large projects, tends to be more homogeneous. In well-coordinated projects of this type researchers agree ahead of time on what data will be generated, how it will be formatted and stored for later access. This is in part because there are immediately other users of the data, the other scientists on the project. As noted earlier, data from large projects also tends to be homogeneous and voluminous because data gathering is often mechanized with instrumentation. Of course there can be huge variation in the instrumentation being used among projects, but within a project the same instruments tend to be used. Also, where projects are large, instrument manufacturers, be they the scientists themselves or commercial entities, have a greater motivation to coordinate instrument data format. As the size of projects exists on a continuum from large to small, the quality of this organization exists on a continuum from highly structured industry-wide standards to relatively independent proprietary data formats. This uniformity in the head makes it much easier for the data to be stored in structured databases. This makes the data more accessible. In contrast, the hand-collected data of the smaller science projects in the tail are not uniform. Each project might invent their own data format, and this may or may not be committed to a structured database. More often the highest level of structure attained in this end of the graph is encoding into spreadsheets. In fact, many scientists make no distinction between databases and spreadsheets although databases are much more amenable to selection, sorting and data merging than are spreadsheets. This extra labor is a barrier to reuse.

In the head it becomes economical to centrally curate the data of the project. Since the scientists of the project are working together it is natural for the information to be gathered together. Sometimes disjoint projects in the tail are organized by funding agencies to answer scientific questions or for economies of scale. This essentially moves these projects from the tail toward the head of the graph whether you count the dollars in the joint projects or the joint data production, as is the case with disciplinary or reference repositories such as GenBank or the Protein Data Bank (PDB) and each of the model organism databases. These repositories have relatively uniform internal structure. With important exceptions discussed below such as LTER, data from projects in the tail generally do not make it into repositories and fall into disuse and darkness. In those cases where they do make it into repositories, these data are generally put into institutional repositories, not because of the similarity of the data format to the other information in the repository but because of properties extrinsic to the data such as the university where it was generated. Queries to access data in the disciplinary collections typically allow future users to pull out individual records of measurements from the previous studies. For example, a researcher can retrieve a sequence of nucleotides in a particular gene from a particular species or individual in GenBank. In contrast, many institutional repositories save many different file formats. For example, both Fedora[iv] and DSpace[v] treat data such as relational

databases as a unitary digital object with metadata describing the set of records rather than the individual records. Later users cannot retrieve individual records but must retrieve an entire collection of records as they were deposited by the original researcher. If this data is properly documented with metadata, then a person retrieving the data will be able to perform operations on their local computer to access particular records. For example, an Excel spreadsheet of the measurements of trees in a study plot could be retrieved from the institutional repository and then the user would need to manually load this into compatible spreadsheet software and manually select tree or records of interest.

When data does not make it into a repository, then it is much less likely to be maintained over time. Data distributed on individual researcher web sites are not maintained and often quickly vanish. This is much more likely to happen in the tail.

Private and government funding agencies for larger projects often pay special attention to the resulting data. When public funds are used or when private agencies are trying to insure the greatest impact of their research investment, they often require explicit data sharing plans and frequently require open access to the data for the rest of the scientific community. Smaller projects are not as likely to attract such oversight so the resulting data can end up in obscure locations (such as desk drawers) or protected from access, again leading to dark data from the perspective of the broader science community.

A final difference between data in the head and data in the tail is the impact that access to the data has on people's careers. For data in the tail, the only definition of success is the publication of a very abstract representation of the data (e.g. graph or statistics) in a journal. In the large science projects, the data management itself is frequently the object of academic and social interest so people's careers can be built on successful management of the primary data so that it can be effectively reused. However, in some cases the person receiving credit may be an informaticist and not the scientists.

Barriers to bringing dark data to light

An irony is that dark data is initially very visible, at least to one individual. Scientists spend time and effort carefully collecting, formatting and saving data about some phenomenon of interest. This may be the location of a planet, the flow of water in a river, the behavior of a mammal, a virus or a molecule. While there is already much dark data that can be exposed and reused, there are also forces that move data into obscurity that must be recognized and addressed. Table 4 in the next section lists some of the barriers.

Data becomes dark because no one is paying attention. There is little professional reward structure for scientists to preserve and disseminate raw data. Scientists are rewarded for creating high-

density versions of their data in statistics, tables and graphs in scholarly journals and at conferences. These publications in some ways are the sole end product of scientific inquiry. These products, while valuable, may not be as useful as some authors hope. In a comparison of offline and online papers Lawrence (2001) showed that the mean number of citations to offline articles is 2.74, and the mean number of citations to online articles is 7.03. If a scientist were to have a data set used at least as often, it might justly be judged to be a greater service to science than the publication. We might expect that if it is available online it would be more likely to be used than if it is kept private. Yet there is largely no reward built into the promotion and tenure process for developing valuable data. Perhaps part of this barrier of lack of reward is because mechanisms are not in place to track data use as closely as we track citation of a paper.

It should also be noted that provision of data and sufficient documentation/metadata for its reuse requires time and financial expense, yet frequently scientists are not provided with the financial resources required to properly curate data. In some cases researchers may receive funds for data maintenance but only for the duration of the grant with an average duration of three years. Where there is some uncommitted capital, priorities and barriers being as they are the capital is spent on some other cause than data. This lack of financial resources to keep data visible may be rooted in part in a broader lack of appreciation of the worth of the data. The true worth of the data is not determined by the cost for gathering it but in the savings incurred by someone else not needing to gather it again. Sometimes this value is immeasurable, as in historic data on climate for example, or the genetic makeup of a rare species, which can not be recreated without far greater expense - even if it can be gathered again. For example, a recent study of impact of climate change on bloom time and consequent impact on pollinators was based on dark data from personal nature diaries of nature lovers from the end of the nineteenth century to the beginning of the twentieth: dark data that could easily have been lost, but was able to be used to demonstrate a relationship between early bloom time and higher temperatures over a century ago (Primack, Miller-Rushing, Primack, & Mukunda, 2007).

Every scientific endeavor is different so there is a tendency to develop specialized data repositories. Even in the area of molecular biology that has GenBank and PDB there are also numerous specialized databases including one for each model species, databases for functional annotation and individual gene and protein families. Each database has special features that add value for the particular data within the database but there is also little coordination between collections and few sustainability plans and replication of management structure. Lack of integrated design has led to a cottage industry of computer programs and books designed to help scientists scrape information from the independent interfaces of each database.

Even if scientists were more inclined to make data available and if they had the finances to do so, they still could not because neither

the scientists nor those under their employ have the skill set required to efficiently and effectively make the data available. The enterprise of science as a whole is lacking in sufficient number of individuals who are familiar with the institutions and technologies available to make data visible to the broadest number of individuals. They are unaware of repositories, database technology, and representations such as XML or RDF. Perhaps scientists should not be required to have a deep knowledge of such things any more than they are required to understand the subtleties of journal layout or distribution channels for their paper publications. But there are few data curation specialists to handle the technical details.

There is also an educational deficit in the understanding and application of intellectual property rights (IPR). Some scientists believe that in releasing data they are forfeiting their IPR. While this may be true of trade secrets it is not true of all rights any more than an author of a book gives up rights to the book by publishing it. Some authors choose to give up limited rights in exchange for services from a publisher. Scientists need to make informed decisions about the impact of different intellectual property decisions on the dissemination of knowledge and attribution of intellectual credit.

In addition, there is a lack of tools for metadata generation, which is an example of the broader lack of tools in most aspects of data curation including acquisition tools, data migration tools, validation tools and others all the way to the end of data usefulness where there is a need for culling tools to support de-accessioning and destruction.

Some of these many barriers might be reduced with the application of technology but easy to use technology is not always available. Many of the processes of efficient data curation have yet to be automated, thus keeping the burden on the scientists. In addition, in many fields the expenditures have not been made to create the institutions that might organize and hold the data although some federal and private efforts have begun to address the issue as discussed below.

A final barrier that cannot be overlooked is the Digital Tower of Babel that we have created with seemingly countless proprietary as well as open data formats. This can include versions of the same software products that are incompatible. Some of these formats are very efficient for the individual applications for which they were designed including word processing, databases, spreadsheets and others, but they are ineffective to support interoperability and preservation.

The visibility of data is often a matter of perspective. Some data is sometimes very visible and “light” from one point of view and user community. The same data may however be “dark” from a different point of view or a different community of potential users. For example, a group of ecological genomicists may be generally aware of a set of

data and its underlying semantics. This mutual knowledge may evolve from shared projects or education. However, a group of system biologists who could use the data may be completely unaware of the data and if they did have the data would not be able to interpret the semantics.

Developments in telecommunications and computing make it much more feasible to bring this data to light than was economical in the past.

Bringing Dark Data to Light: Potential Solutions

Institutional Solutions

Existing and developing institutions will play a critical role in improving access to dark data. Some of the same solutions for data management that are being developed for big science can also be applied to the data in the tail. In fact, in some disciplines the process has already begun. One solution is to create science data centers around individual disciplines. Many initiatives within federal agencies such as NSF, NASA, NOAA, and by individual principal investigators are already addressing these issues. For example, at the organizational level, NSF funded the creation of the National Center for Ecological Analysis and Synthesis (NCEAS)[vi] in part to address data issues. The mission of NCEAS is threefold. First, advance the state of ecological knowledge through the search for general patterns and principles in existing data.

Institutions	**Roles**
Science Centers	Disciplinary Repositories and Specialized Tool Development
Museums, Libraries and Archives	Institutional Disciplinary Repositories
Funding Bodies	Seeding Innovation, Technology Development
Publishing Industry	Referencing and Storing "Published" Data
Educational Institutions	Training Scientists and Science Information Specialists

Table 3: Institutional Solutions

Second, organize and synthesize ecological information in a manner useful to researchers, resource managers, and policy makers addressing important environmental issues. Third, influence the way ecological research is conducted and promote a culture of synthesis, collaboration, and data sharing. Other examples are the National Evolutionary Synthesis Center (NESCent)[vii] and the Plant Science Cyberinfrastructure Center (PSCIC - iPlant)[viii]. Select social science data is managed by the Inter-university Consortium for Political and Social Research (ICPSR)[ix]. While this is a useful step it is certainly not the case that all or most of ecological data is curated by NCEAS, or evolutionary data by NESCent, plant data by iPlant or social science data by ICPSR. Also, the long-term economic sustainability of some institutions of this type is questionable.

Libraries are already addressing long tail issues introduced by the internet for their text collections (Dempsey, 2006). Libraries are increasingly playing a role in science data curation as well (Treloar 2006, 2007) but face cultural and financial challenges (Carlson, 2006; Davis & Vickery, 2007). While many initiatives began with a focus on digital text, these experiences have paved the way for management of other scholarly output. The Association of Research Libraries (ARL) has been participating in a series of workshops and publications on data curation[x]. Many libraries have already established institutional data repositories, which are sometimes centered on particular disciplines, including GIS, ecological and chemical structure data to name a few. There is well-established library collaboration for chemical crystallography data (Coles et al., 2006). While academic libraries are nearly as stable as the academic institutions where they reside, and have funding models supported in part by a sustainable fraction of overhead collected from the research and education missions of the institutions, it is unlikely that the added burden of data curation can be managed within current funding levels.

Museums are in a similar state of development to libraries in terms of data management. Museums have been moving from independent databases run by individual investigators to at least the availability of central databases. This data can increasingly be shared across institutions as for example with natural history collection and observation data through the Global Biodiversity Information Facility (GBIF) portal[xi]. However, this type of data represents only a small portion of the data collected by museum scientists, staff and volunteers.

Publishers through the twentieth century came to dominate scientific journal production and some are now beginning to associate data with publications. In the biology literature this is well established with GenBank. Publishers avoid using valuable page space on DNA sequences because the sequences can be included now by reference. More publications now are allowing deposit of data behind graphs and statistics that is a valuable advance but leaves many issues unresolved. These include limited data searching capability, lack of storage for more voluminous underlying data, and mechanisms for storing unpublished data.

Promising Approaches

The services and organizations listed above and ones like them are working on solutions to the barriers to effective data use enumerated above. This section lists some of the solutions and the organizations working on them. While the solutions listed here are not exhaustive, and in some cases may not prove to be practicable, they do represent a sample of the solution space for the curation problem. Unfortunately, there is no one solution to the optimization of data preservation and use.

Barrier	Potential Solutions
Lack of Professional Reward Structure	Funding Body Requirements Data Citation Requirements, Data Citation Index Replace or Educate the Old Guard
Lack of Financial Reward Structure	Funding Initiatives: NSF DataNet, Interop; IMLS Data Curation Initiative
Undervaluation / Lack of Investment	Public and Private Foundation Initiatives Sociology of Science Research
Lack of Education in Data Curation	Formal Education Programs
Intellectual Property Rights (IPR)	Formal Education Programs Science Commons
Lack of Metadata Standards and Creation Tools	Metadata Working Groups, Metacat
Lack of Sustainable Technology	DataNet
Cost of Infrastructure Creation	Data Repositories Cyberinfrastructure Development (OCI, eScience) Metadata Tool Development Research Initiative e.g. DataNet Publishers, Data Federation Technology (TAPIR)
Cost of Infrastructure Maintenance	Long Term Collaborations and Institutionalization and Economies of Scale
(Babble) PDF, Excel, MS Word, ArcView, Floppy Disks	Open Formats, translation tools, migration tools (e.g. Fedora)

Table 4: Barriers and Solutions to Data Reuse

As with many of the barriers to optimal data use, many institutions need to be involved in establishing a professional reward structure for scientists to participate. Sharing and long-term preservation of data should lead to professional success. Scientists currently get credit for the citation of their published papers. Similar credit for data use will require a change in the sociology of science where data citation is given scholarly value. The publishing industry including for example “Nature” and “Science” is already beginning to provide a solution by allowing data to be connected with publications. However, space limits, format control and indexing of data remain a major problem. Institutional and disciplinary repositories need to provide facilities so that citations can return the same data set that was used in the citation without adding or deleting records. Standards bodies for the sciences can set up methods to cite data in databases and not just data in publications (Altman & King, 2007). Once publishers and index services include these citations in calculation of impact factors for data sets as they now do with journals there is little doubt that tenure and promotion committees will acknowledge the value of data and give professional credit to the scientists

responsible for gathering the data.

There must be sufficient funds to select, annotate, preserve and disseminate data. Libraries are one obvious solution but libraries face financial constraints (Carlson, 2006; Davis & Vickery, 2007). Most often, disciplinary and reference repositories also survive only on project funds that will terminate or require new funding every three to five years. These are currently addressed with project oriented grants and therefore time limited solutions with funds from the Institute of Library and Museum Services, the National Science Foundation and other agencies. But initiatives such as NSF's Sustainable Digital Data Preservation and Access Network Partners (DataNet)[xii] are intended to explore and establish more sustainable models. Also, NSF's Community-based Data Interoperability Networks (INTEROP)[xiii], designed to foster solutions for data interoperability, has fostered projects to improve data annotation. These IMLS and NSF programs, as well as e-Science programs in Europe, are funding the study of work practices and economics of different solutions, including the estimation of the value of data. Economies of scale may also help overcome the financial barriers. Corporations such as Google and Microsoft as well as traditional publishers are beginning to provide storage for data on the scale of hundreds of terabytes. As an example, for smaller data sets Google released the Google Data API[xiv]. Commercial entities are also involved in making molecular life science data tools, for example NextBio[xv]. The financial models for this work are not yet clear. In spite of all of this effort, there continue to be specialized data collections with little interoperability.

Universities and data centers are beginning to deliver data curation education programs. The United States federal agency the Institute of Museum and Library Services (IMLS)[xvi] funds programs for concentrations in Master of Science degrees and professional development workshops that are offered by the University of Illinois[xvii], the University of North Carolina[xviii] and the University of Arizona[xix] and others. The Digital Curation Centre[xx] in the UK provides professional training, and international collaboration in education is beginning through the International Data curation Education Action (IDEA) Working Group[xxi]. Conferences provide a venue for education and the dissemination of best practices. These include for example DigCCurr II[xxii] and the International Digital Curation Conference[xxiii].

To some degree intellectual property barriers can be addressed with education of the scientists and their support staff. As curation is professionalized and receives proper funding we can expect copyright mechanisms to follow those that are developing for text with a broad range of options available. It is critical, however, that scientists make informed decisions about control of intellectual property rights so that the positive impact on science is maximized.

Technologies have been slow to develop to make data curation easy. Governmental and non-governmental organizations are funding tool development and evaluation. Institutional and disciplinary

repositories are beginning to run on common platforms so that the development costs can be shared over many users. The barriers are still high enough however that the majority of scientists do not properly managed their data for the long term.

Several technologies are particularly well suited to exposing dark data. These include for example, thesauri and controlled vocabularies to describe the data to make it easier to find. An example is the Biocomplexity Thesaurus maintained by the National Biological Information Infrastructure[xxiv]. Metadata formats allow data descriptions to be integrated. An example is the Ecological Metadata Language (EML)[xxv]. Ontologies help to define the relationships among individual elements of data sets to make them interoperable. Examples include the Gene Ontology (GO)[xxvi], the Plant Ontology[xxvii] and the collection of the Open Biomedical Ontologies (OBO)[xxviii]. Confederated data sharing frameworks have been developed with free or low cost software tools including for example Dublin Core being shared over Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[xxix] for mostly text and Darwin Core[xxx] or ABCD[xxxi] being shared through DigIR or TAPIR[xxxii] for biodiversity occurrence data. These initiatives expose data that would otherwise be sequestered in individual institutional databases. The semantic Web promises to make substantial contributions to solving the problems of data access but the long tail data is not frequently the focus of these efforts. There are other existing technologies and as yet to be invented technologies that could assist with improving access and use of dark data.

Understanding Dark Data

In order to better manage the dark data of the tail of science, it is necessary to understand that data. The DataNet solicitation from NSF addresses some of the issues for making data more accessible and more useful. As is reasonable most proposals focus on the head of the graph first, the most frequent data, and then some move on to the tail. These efforts do not however directly address some of the questions about the dark data in the tail. Some important questions are:

How long is the tail?

What is the area under the tail?

What data in the tail and the head are “dark”?

How do we determine the value of dark data?

What is different between tail-science and head-science?

What is the differential distribution of sciences based on data size and funding?

Which data is more likely to contribute to transformative science?

A number of methods can be used to understand dark data but all in the end are a study of the behaviors of individual scientists. Economists and sociologists of science are currently conducting surveys, interviews and observation studies to understand how data is handled and stored in different scientific disciplines. This information will help us to design better mechanisms to support broader reuse of scientific data.

Conclusions

Data is the underpinning of the scientific method. Without data to back up theory science becomes ungrounded conjecture. While the majority of data from large scientific enterprises is well curated, there is little scientific infrastructure in place to support the storage and reuse of data created by smaller projects. In order to maximize our return on investment in scientific research we need to develop this science infrastructure through existing institutions such as libraries and museums that have traditionally been the guardians of scholarly productivity. We need to develop technologies that make it cost effective for scientists to document and deposit their data in these repositories. We also need tools that make it easy to search and retrieve data from these repositories. We need to educate a new generation of curators of our scholarly output, who are trained in appropriate computer technology, and who have an appreciation of science and the sociology of science. Most of all we need new educational initiatives and incentives that will give the next generation of scientists the knowledge they need to make informed decisions about the broader use of their data and broader impact of their research.

Acknowledgements

I would like to thank Pamela Brooks-Pope at the Division of Biological Infrastructure at NSF for helping to gather the raw data on NSF grants. Linda Smith and Melissa Cragin made valiant efforts to correct and improve earlier versions of this text.

Bibliography

Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, *13*(3/4), March/April. Retrieved from http://www.dlib.org/dlib/march07/altman/03altman.html

Anderson, C. (2004). The Long Tail. *Wired Magazine 12.10*. Retrieved from http://www.wired.com/wired/archive/12.10/tail_pr.html

Carlson, S. (2006). Lost in a sea of science data. *Chronicle of Higher Education, 52*(42), A35. Retrieved from http://chronicle.com/free/v52/i42/42a03501.htm

Coles, S. J., Frey, J. G., Hursthouse, M. B., Light, M. E., Milsted, A. J., Carr, L., De Roure, D., Gutteridge, C., Mills, H. R.,

Meacham, K., Surridge, M., Lyon, E., Heery, R., Duke, M., & Day, M. (2006). An E-Science environment for service crystallography-from submission to dissemination. *Journal of Chemical Information and Modeling 46*(3): 1006-1016.

Davis, H. & Vickery, J. (2007). Datasets, a shift in the currency of scholarly communication: Implications for library collections and acquisitions. *Serials Review, 33*, 26-32.

Dempsey, L. (2006). Libraries and the long tail. *D-Lib Magazine 12*(4). Retrieved from http://www.dlib.org/dlib/april06/dempsey/04dempsey.html

Goetz, T. (2007). Freeing the dark data of failed scientific experiment. *Wired Magazine,* 15(10). Retrieved from http://www.wired.com/science/discoveries/magazine/15-10/st_essay.

Lawrence, S. (2001). Online or invisible. *Nature, 411*, 521.

Mabe, M. & Amin, M. (2001). Growth dynamics of scholarly and scientific journals. *Scientometrics, 51*(1), 147-162.

Nature. (2008). "Big Data." *455*(7209), 1-136.

Peck, S. (2008). Science suffers when getting a grant becomes the goal. *The Chronicle of Higher Education, 55*(7), A42.

Porter, J. (2000). Chapter 3: Scientific databases. *In* W. K. Michener, & J. W. Brunt (Eds.), *Ecological data: design, management and processing* (48-69). Oxford, UK: Blackwell Science, Inc.

Primack, R., Miller-Rushing, A., Primack, D., & Mukunda, S. (2007). Using photographs to show the effects of climate change on flowing time. *Arnoldia, 65*(1), 2-9.

Refinetti, R., Cornélissen, G. & Halberg, F. (2007). Procedures for numerical analysis of circadian rhythms. *Biological Rhythm Research, 38*(4), 275-325. Retrieved from http://www.informaworld.com/10.1080/09291010600903692

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L., & Atkinson, M. (2005). The Digital Curation Centre: A vision for digital curation. In *Proceedings of Local to Global Data Interoperability - Challenges and Technologies, 2005.* Mass Storage and Systems Technology Committee of the IEEE Computer Society, June 20-24, 2005, Sardinia, Italy. Retrieved December 4, 2007 from http://eprints.erpanet.org/82/

Seglen, P. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, Education, 314*(7079), 497.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika, 42*(3/4), 425-440. Retrieved from http://linkage.rockefeller.edu/wli/zipf/simon55.pdf

Sun Y. & Giles, L. (2007). Popularity weighted ranking for academic digital libraries. In *Proceedings of the 29th Annual European Conference on Information Retrieval Research*, ECIR 2007, Rome, Italy in April 2007, pp 605-612.

Treloar, A. (2006). The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: building the new collaborative e-research infrastructure. *Proceedings of AusWeb06*, the Twelfth Australian World Wide Web Conference, Southern Cross University Press, Southern Cross University, July. Retrieved from http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/

Treloar, A. (2007). DART: Building the new collaborative e-research infrastructure. *Proceedings of Educause Australasia 2007*, Melbourne, April. Retrieved from http://www.caudit.edu.au/educauseaustralasia07/authors_papers/Treloar-183.pdf

Young N., Ioannidis J., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Med 5*(10): e201 doi:10.1371/journal.pmed.0050201.

[i] A curve that follows the power law is characterized as beginning relatively high on the y-axis with the value of y dropping very quickly but levels off before approaching a y of 0. This is characterized in the equation, $f(x)=ax^k+o(x^k)$, where k defines the steepness of the curve and o defines a constant increase in y. (Simon,1955).

[ii] There are differences between when funds are awarded and when funds are distributed so the figures given here are approximate. Grants of less than $500 have been excluded since these smaller amounts are generally corrections and accounting adjustments rather than actual research awards.

[iii] "Hard data" is used here in common sense usage as in "hard facts" that can not be refuted.

[iv] http://www.fedora.info/

[v] http://www.dspace.org/

[vi] http://www.nceas.ucsb.edu/

[vii] http://www.nescent.org/index.php

[viii] http://www.iplantcollaborative.org/

[ix] http://www.icpsr.umich.edu/

[x] http://www.arl.org/

[xi] http://www.gbif.org/

[xii] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

[xiii] http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf07565

[xiv] http://code.google.com/apis/gdata/overview.html

[xv] http://www.nextbio.com/

[xvi] http://www.imls.gov/

[xvii] http://www.lis.uiuc.edu/programs/ms/data_curation.html

[xviii] http://ils.unc.edu/digccurr/

[xix] http://www.sir.arizona.edu/program/digin/index.html

[xx] http://www.dcc.ac.uk/

[xxi] http://www.dcc.ac.uk/events/idea-2008-edinburgh/

[xxii] http://ils.unc.edu/digccurr/aboutII.html

[xxiii] http://www.dcc.ac.uk/events/dcc-2008/

[xxiv] http://www.nbii.gov/

[xxv] http://knb.ecoinformatics.org/software/eml/

[xxvi] http://www.geneontology.org/

[xxvii] http://www.plantontology.org/

[xxviii] http://www.obofoundry.org/

[xxix] http://www.openarchives.org/pmh/

[xxx] http://wiki.tdwg.org/twiki/bin/view/DarwinCore/

[xxxi] http://wiki.tdwg.org/twiki/bin/view/ABCD/

[xxxii] http://wiki.tdwg.org/TAPIR/