AC 2007-155: THE DATA DILEMMA

Amy Stout, Massachusetts Institute of Technology

Anne Graham, Massachusetts Institute of Technology

The Data Dilemma

> There's a famous allegory about a map of the world that grows in detail until every point in reality has its counterpoint on paper; the twist being that such a map is at once ideally accurate and entirely useless, since it's the same size as the thing it's meant to represent [1].

Introduction

The proliferation of scientific data is inspiring a paradigm shift in the way we manage information. Scientists frequently use other scientists' data for their experiments [2], taking a step out of the traditional process known as the scientific method [3]. As data is rapidly produced and shared, the results of experiments are practically becoming disseminated as they are collected, speeding up a process that used to take longer [4]. With such a wealth of data available, information retrieval has become a critical component of scientific research. Tools like metadata, sophisticated databases and search engines are desperately trying to keep pace with the changing world [5]. Furthermore, there are social and legal issues to consider. What data can be shared and disseminated? Who owns data? What about "facts" that have been extracted from years of experimentation or using patented devices? Traditionally, so-called facts have not been copyrightable, resulting in laws that become blurred [6].

Another issue affecting data management is how to handle data as an object. Librarians are used to the book/journal model [7]. Open Access, a movement that started in the early

1990s in an effort to make published articles freely available to the public, is now extending its reach to data [8]. As part of a task force at MIT, librarians interviewed researchers to get their perspectives on data, with the goal of gathering ideas on how to assist the researchers. In addition, librarians are submitting a data set to MIT's institutional repository, DSpace, in an effort to investigate the technical challenges presented by data storage. This experience will provide insight into the technical and social issues librarians can address with expertise. As librarians become more skilled with data management, they will be able to better advise and assist scientists, opening up new collaborations between librarians and their academic communities.

Growth of Scientific Data

The proliferation of scientific data is overwhelming the research community [9]. Colossal improvements in analog instruments as well as the development of virtual instrumentation [10] contribute to an environment where data is being collected around the clock, every day of the year, all over the world [11].

While the agricultural revolution increased the production of food, and the industrial revolution increased the production of material goods, the information age has increased the production of data [12]. A UC Berkeley study conducted in 2001 found that the world produces around 250 megabytes of unique information per year for every man, woman and child on the planet [13].

The exponential growth of data is overwhelming the availability and accessibility of storage, even taking into account the growth of computer power predicted by Moore's Law, which theorized that the density of an integrated circuit would double every 18 to 24 months [14]. Not only is data storage critical, but the tools needed to access the data are essential. "Years ago…scientists could study the [data] output manually…they now need computers and sophisticated algorithms to wade through it all [15]."

Issues surrounding data storage

In part, the information revolution is a dream that has come true. Never has so much potential knowledge been so accessible. Never has so much data made itself available for analysis and understanding. However, data production is increasing, not at a consistent rate, but exponentially, outstripping our abilities to cope with it. The data storage curve is now rocketing upward at a rate of 800 percent per year. According to Usama Fayyad, "It makes Moore's law look like a flat line," he says. "Companies are collecting so much data they're overwhelmed [16]." There is literally no place to put all of this data.

> "Sometime next year…the largest scientific instrument ever built will come to life... [It] is expected to spew out some 15 million gigabytes a year—on average, that's more than enough to fill six standard DVDs every minute. Storing and analyzing the mountain of data, it turns out, is a task that no supercomputer in the world can handle [17]."

It may be intuitively obvious that data is not like a journal article or a book. And yet most of the data storage and dissemination models treat data like it is. Books and journals (and journal articles) are discrete items with specific publication dates. (Although with digital preprints and institutional repositories, this may change.) When they are finished they are finished – they are not, like data, frequently added to and manipulated. Publications are not created using rare, proprietary software which can keep data from being reusable. DSpace (http://dspace.mit.edu) is an example of an institutional repository developed by MIT's Digital Library Research Group and Hewlett Packard in the early 2000s. A presentation by Pablo Boczkowski of MIT's Sloan School of Management outlines the popularity of DSpace with librarians, and its dismal unpopularity with academics. He points out that DSpace does not mesh well with current academic practices [18]. For example, Scientists are likely to want the final versions of their articles in circulation, and not the preprint sent to the publisher. Once an item is submitted to DSpace, it is in circulation until it is forcibly suppressed. In addition, submission to databases like DSpace requires time and effort – many researchers are simply too busy to assign metadata and choose licenses for their work. And, of course, there is the issue of copyright. While a researcher may have the right to post a preprint (or even a final draft) of a paper in DSpace, wading through a publishing agreement may prove too time-consuming.

Though databases like DSpace are theoretically able to accommodate data sets, they fail to capture the slippery nature of data. For example, when is a data set complete? What kind of metadata structure should be applied to it? (Data Documentation Initiative -- DDI

[http://www.icpsr.umich.edu/DDI/] is a current favorite in the library world.) The construction of metadata is time-consuming and labor-intensive. Most, if not all, researchers are too busy to annotate their data. In addition, it is hard to imagine data being stored in perpetuity if it is stored in a proprietary format. Years from now, computer-aided design (CAD) software will be dramatically different -- each company manufactures products based on different algorithms, making it difficult to access data from different software packages or even different versions.

Issues surrounding data dissemination

Yet this data is valuable to scientific research. Or is it? How useful is data if it can't be accessed and disseminated? Not only must the data be made available, it must be described in such a way that its contents and structure are apparent to the user. Currently, laboratory data is frequently stored on researcher PCs, lab servers or communal repositories (in the form of reports) like arXiv (http://arxiv.org/). Institutional repositories like DSpace (http://www.dspace.org) and Eprints (http://www.eprints.org) are better suited to traditional, discrete publications than data sets. Metadata descriptors may help a user understand the contents of a data set, but researchers frequently don't have time to go through the labor-intensive process of adding meaningful descriptive information. The next-generation Internet, known as the Semantic Web, promises to accelerate the process of scientific data dissemination. Using catchy concepts like the Resource Description Framework (RDF) and its elegant triple-stores (http://www.w3.org/2001/sw/), the Semantic Web classifies data and allows it to be exchanged freely over the Internet.

However, technical challenges are not the only issues. Some researchers can't, for privacy reasons, share their data. Medical data is an example of information that must remain private. Some researchers simply won't share their data, for a variety of (usually understandable) reasons. "Scientists often closely guard their data – the numbers can hold keys to competitive secrets or proprietary information, or reveal the embarrassments of experiments poorly executed [19]."

There are also legal issues involved with the dissemination of data. Traditionally, data, or facts, have been freely available, untouched by the constraints of copyright or patent. But what about data that has been generated via an intricate, complicated process that is difficult to master? What is a fact, after all? Microbiologist Craig Venter, for example, tried unsuccessfully to patent gene fragments [20]. All of these issues show that volume of data is not the only issue. The data itself is inspiring a change in the way we understand information storage and dissemination.

A changing model

Formerly, scientific research was conducted on the model of: hypothesis, experiment, results [21]. Today the collection of data and the conducting of research may be done separately. A scientist may very well never collect data while spending his/her time analyzing data collected by others. Likewise, another researcher may spend his/her time only collecting data and never analyzing it [22].

One aspect of the data revolution that is inspiring a paradigm shift is the notion of concurrent publishing. The time lag between the collection and publishing of data is rapidly diminishing. In some cases, the instruments themselves publish data in repositories that are up-to-the-minute ready for human use [23]. A simple example is weather data from Doppler radar instruments.

In 1945, as director of the Office of Scientific Research and Development of the U.S. government, Vannevar Bush wrote an essay, "As We May Think." Frequently cited as a predictor of hypertext, "As We May Think" announces a world in which scientists collect data omnivorously, store it, and link it seamlessly with the rest of human knowledge [24]. Perhaps fanciful in nature, the internet along with an explosion of data hint at a world not far from the one suggested by Bush.

Open access to data

The Open Access movement which began in the early 1990s promotes the free dissemination of scientific research to the public via the Internet. Based on the potential of the Internet, the rights of taxpayers to access taxpayer-funded research and the exorbitant prices of scientific journals, the Open Access movement was born. While initially targeted towards journal articles, Open Access has evolved to include data. Great discoveries have been made using data gathered by other scientists. Beth Chen, a researcher at Cold Spring Harbor, discovered new neural synapses in a roundworm solely using data assembled by another researcher [25]. The data collection in this case was a

labor-intensive process in itself. Access to information is clearly a critical part of contemporary science.

Research at MIT

In late 2005, a group of MIT librarians formed a task force to investigate data and the role that librarians could play in archiving and disseminating scientific data at MIT. The task force polled faculty members about their collection, storage, and dissemination of data. Altogether 15 faculty members responded.

Some themes that appeared based on this research were:

- There are many technical barriers to data storage and organization.
- Metadata is rarely assigned, and is done in a haphazard way.
- Some data is not shareable, due to privacy considerations.
- Loss of control over data is an issue.
- Time and effort are needed to manage data effectively.
- Researchers would like libraries to provide access to commercial databases.
- Researchers need to get to earlier versions of data sets.
- Researchers are curious about the flexibility of DSpace with respect to data storage.
- There is a perceived need for a centralized rather than distributed model of data storage.
- Researchers need to evaluate what is worth archiving.
- Researchers have faith in the longevity of current formats.

A second part of the MIT Libraries' task force's experiment was to load a data set into DSpace, MIT's institutional repository. The goal of this activity was to experience firsthand the successes and challenges of managing data in DSpace. Although the researchers who responded to the survey indicated little interest in DSpace, it is equipped, theoretically, to handle data submissions. We decided to try submitting data before appointing ourselves experts on DSpace data submission.

Using data generated by the MIT Libraries' Userneeds (2006) and the Journal Use (2005) Studies, the task force plans to deposit library research data in DSpace. Depositing the Journal Use Study data should be easy because the data is available in a flat file. Depositing the Userneeds Study data is more challenging because it is stored in a relational database. Learning to cope with complicated data structures will pave the way for future exploration.

The Userneeds Study data is and will be stored in FileMaker Pro, but what happens when FMPro becomes obsolete? The data needs to be also stored in a flexible file (or set of files) that can be reconstructed using the database software programs of the future. Task force members have created a flexible file of the Userneeds Study data and are now developing a metadata schema using DDI (Data Documentation Initiative) that will provide meaningful structure to the data.

As managers of traditional collections, librarians are poised to take on the data world. Knowledge of information storage, preservation and dissemination will carry them, and their researchers, into the next decades. While librarians are not, as such, equipped to advise on data collection, they are skilled data managers, able to provide leadership in this area. As Scott Carlson writes, "Librarians will have to step forward to define, categorize and archive the voluminous and detailed streams of data generated in experiments [26]."Questions and ideas librarians can explore with their communities include:

- Why store, preserve and share your data?
- What happens to research data when researchers leave the institution?
- Open Access, Creative Commons (http://creativecommons.org) and copyright
- Where to store your data
- Technical issues surrounding data storage
- Preservation issues
- How to access data created by other researchers
- Metadata creation, schema and the Semantic Web (http://www.w3.org/2001/sw/)

Conclusion

The explosion of scientific data creates new challenges for scientists, data managers and librarians to archive and disseminate information. So much data is being generated that some scientists have removed themselves from the traditional scientific method by using only data generated by others' experiments. Social, technical and legal issues hinder data

creation and management, not to mention the nature of data itself. Not like a book or a journal, traditional methods of managing materials fall short with respect to data. Open Access, a movement that originally addressed publications like books and journal articles, has evolved to include data. Librarians at MIT conducted an experiment in 2005-2006 that underscored the issues of management and dissemination. The knowledge gathered from the results of this experiment informed them of how to strengthen librarian services to academic communities. In its fledgling stages, data storage and dissemination are ripe to mature in the upcoming years. Such an evolution is crucial to science. After all, as renowned physiologist Claude Bernard said, "Art is I; science is we [27] ."

[1] Lewis J. Memory Overload. Wired. February, 2003.

[2] Carlson S. Lost in a Sea of Science Data. Information Technology. June 23, 2006.

[3] Bacon F. Novum Organum. 1620.

[4] Burger F, Reich S, Wagner R. Information Technology as the Glue for a Concurrent Publishing Environment. FAW - Research Institute for Applied Knowledge Processing. 2006.

[5] Schmidt C. Data Explosion: Bringing Order to Chaos with Bioinformatics. Environmental Health Perspectives 3 (6) May 2003.

[6] Shreeve J. Craig Venter's Epic Voyage to Define the Origin of Species. Wired. August 2004.

[7] Boczkowski P. The Reinvention of Librarianship in the Development of DSpace. Presentation at MIT, 2005.

[8] Budapest Open Access Initiative. 2002. http://www.soros.org/openaccess (accessed 01.05.2007).

[9] Gagliardi F, Grey F. Old World, New Grid. IEEE Spectrum. July 2006.

[10] National Instruments. http://www.ni.com (accessed 01.05.2007).

[11] Microsoft Corporation. Towards 2020 Science. 2006.

[12] Alesso P, Smith C. Thinking on the Web. (John Wiley & Sons: New Jersey). 2006.

[13] Lyman P, Varian H. How Much Storage is Enough? ACM Queue 1 (4) June 2003.

[14] Moore G. Cramming More Components onto Integrated Circuits. Electronics 38 (8) April 19, 1965.

[15] Schmidt C. Data Explosion: Bringing Order to Chaos with Bioinformatics. Environmental Health Perspectives 3 (6) May 2003.

[16] Lewis J. Memory Overload. Wired. February, 2003.

[17] Carlson S. Lost in a Sea of Science Data. Information Technology. June 23, 2006.

[18] Boczkowski P. Cultural and Political Dynamics in the Adoption of DSpace. Presentation at MIT, 2005.

[19] Gagliardi F, Grey F. Old World, New Grid. IEEE Spectrum. July 2006.

[20] Shreeve J. Craig Venter's Epic Voyage to Define the Origin of Species. Wired. August 2004.

[21] Bacon F. Novum Organum. 1620.

[22] Hede K. There's Gold in those Archives. HHMI Bulletin. May 2006.

[23] schraefel, m. Presentation at MIT. November 17, 2006.

[24] Bush V. As We May Think. The Atlantic Monthly. July 1945.

[25] Hede K. There's Gold in those Archives. HHMI Bulletin. May 2006.

[26] Carlson S. Lost in a Sea of Science Data. Information Technology. June 23, 2006.

[27] Bulletin of the New York Academy of Medicine, vol. 4, 1928.