The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Recycling Information: Science Through Data Mining

Michael Lesk,
Chair, Department of Library and Information Science,
Rutgers University

January 2008

Summary

An article considering the changes afoot in the world of Science and how the exponentially increasing amounts of recorded data are affecting the way in which scientists now work, for example with data mining. Changes in the way that resources become obsolete are also discussed and how more value must be placed on the work of professionals in digital curation.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

My first chemistry professor, the late Frank Westheimer, once told a student “a month in the laboratory can save an hour in the library”. This balance is shifting more and more to the library, or nowadays the data center, as it becomes cheaper and faster to store and save data, and ever more expensive to pay for laboratory assistants. This is going to be a disruptive change to science as a profession, and we have not faced up to it. We need to give greater recognition to those who preserve and exploit data resources; we have not figured out how to do that either.

Today the traditional scientific method is changing, thanks to data mining and automatic sensors. Instead of the traditional sequence “think of hypothesis, design experiment to test it, run experiment, analyze results” we now have “think of hypothesis, look up data that relates to it, and evaluate it.” Nowadays in many fields of science, data collection is done on a massive scale by automated methods, with large-scale databases to hold the results. The astronomers have some 40 terabytes, the seismologists have 60 terabytes, and the human genome data is 80 terabytes. The climatologists at NCAR have 4 petabytes. (I can remember being asked to hold my disk space to 200 kilobytes. That was a long time ago).

As we get larger and larger data resources, and more and more intelligent data mining software, it becomes easier to make discoveries by going through existing data, rather than by collecting new data. Researchers in topics ranging from genomics to epidemiology publish papers based on data mining, not on new experiments. But what happens to questions of promotion and tenure, if running experiments is the traditionally valued skill but is no longer the best way to make progress? Today, you get promoted in a museum by collecting more material. Exploring the basement is a lower-priority activity. This prestige system will be changing, as researchers publish faster by data mining than by exploration. But that change is going to induce stresses. We need to recognize the increased role of data, and the importance of rewarding people who save it and know how to analyze it.

Houston, We Have a Problem

Not everything is easy in the new world. You can look at a published paper, or even a traditional laboratory notebook, and expect to be able to read it and understand it. If it is falling apart with age, you can see that with your own eyes - and there will be decades between when you can see that it needs preservation and when you can no longer read it at all. If you’re handed a disk drive or a tape cartridge, you have to find a suitable machine and software program to read it, and you cannot even tell until you try whether it is still readable. Nor will it be obvious from the outside what it is. As examples of how bad the situation is, NASA temporarily lost its audio recording of the first Moon landing, eventually finding it in a box labeled “bad tape”. Even worse, NASA lost its original copy of the video recording of the first Moon landing, and that has not been seen in 30 years.

Nor do scientists accept the same responsibilities for letting other people see data that are traditional for publications. To quote a 1995 National Academies study, “A large amount of valuable scientific data gathered with federal funds is never archived or made accessible to anyone other than the original investigators, many of whom are not government employees. In many instances, the organizations and individuals that receive government contracts or grants for scientific investigation are under no obligation to retain the data collected, or to place them in an accessible archive at the conclusion of the project. Thus, data sets that commonly are gathered at great expense and effort are not broadly available and ultimately may be lost."

The same study explained the social reason for this: "A general problem prevalent among all scientific disciplines is the low priority attached to data management and preservation by most agencies. Experience indicates that new research Year projects tend to get much more attention than the handling of data from old ones, even though the payoff from optimal utilization of existing data may be greater."

I once read through the biographies of the departmental curators of the Smithsonian's National Museum of Natural History. Every one ran a collection, ranging from 17,000 items (meteorites) to 40 million items (paleobiology). Every one was out there collecting somewhere in the world. Each biography gives good reasons for continued collection, and certainly in the face of environmental deterioration I'm not suggesting that collection should stop. But as long as promotion comes mostly from collecting, and not from care and study of what we have, we'll continue to find that resources once gathered are not treated well.

What Can We Do?

Traditionally, libraries, museums, and archives were the organizations that kept things around. Can we just ask them to keep the data files? Unfortunately, although they have people who understand preserving things for the future, they are all under budget stress today, and they generally lack the technical skills to deal with the computer files from multiple scientific specialties. Furthermore, the constantly cheaper cost of storage means that we tend to keep more and more, so that human time to look at it becomes ever more expensive. Some years ago Bill Arms suggested that we would be dividing all our material into three piles: (1) stuff of such obvious importance that we will find the time to study it and figure out how to convert it to standard formats and preserve it; (2) stuff clearly so trivial that we're prepared to discard it; and (3) stuff we can't afford to study and which we will leave as it is, hoping that our successors have better automatic tools to decipher it. Bill suggested that 90% of the stuff would be in the last category. The good news is that as storage becomes cheaper, it becomes less of a problem to save it, so we don't have to worry that our successor will erase the disks so that they can be re-used.

Justifying the human time to study data objects isn't easy. As mentioned before, researchers are rewarded for collecting new data, not preserving the results of their past work. The British Library wrote in its strategic plan some years ago of its goals of access and preservation: preservation was done for future users, and access for current users. Only current users, however, vote on the library budget. Unless data preservation also helps with current access, it will be hard to support. Fortunately, many of the important steps, such as conversion to standard format and creation of useful metadata, help with both preservation and access.

The Social Milieu of Digital Preservation

As mentioned above, the most important step is to develop the idea of data curation as a profession, and to somehow recognize its practitioners and their work. There is much in common across different scientific databases, and the skills needed to handle them should be shared. Each scientific community should not have to relearn issues of database design, digital forensics, statistical quality assurance, and visualization.

A particularly interesting problem is the length of time that data must be held privately before it can be shown to the public. Usually traditional archives did not take material until it was going to be generally available. For example, historically, the Public Record Office in the UK (now known as The National Archives) accepted government records 50 years after they were created. Nowadays that is no longer workable: imagine giving somebody today a pile of 1950s-era steel-based magnetic tape, or a box of 80-column punched cards. The archive has to take the material as it is created, and if it can't be shown to anyone else, preserve it in secret. It's hard to fund that kind of operation.

Even in scholarship, we have no agreement on the length of time somebody should have private use of data. Protein chemists have agreed that as soon as you claim to have measured a structure, you have to deposit it in the Protein Data Bank. Astronomers have agreed on one year of restricted use. The Dead Sea Scrolls were kept secret for 40 years. And yet molecular biology is of potentially immense commercial value, while it is hard to think of anything with less financial potential than either cosmology or theological history. Archivists who can't make their work available are unlikely to attract funding or recognition.

Perhaps most important, we have to somehow create a new profession with rewards commensurate with the true stature of the field. Part of this will be conferences and journals that let practitioners get academic credentials, part will be organizations that recognize permanent responsibility for data storage and exploitation, and part will be having a few important people talk about the significance of the area. An essential step will be to view data deposit and preservation as the responsibility of any scientist who collects data. Biomedicine is the area in which this discussion is furthest advanced, with strong arguments that all detailed clinical data should be available for public use.

Conclusions

Perhaps the simplest and most telling development, however, would be increased government support for data curation. It was impressive how many top scientists in other areas moved into the "digital library" community when it had funding; we need to preserve that momentum and build on it.