INFORMATION GULAGS, INTELLECTUAL STRAIGHTJACKETS, AND MEMORY HOLES: THREE PRINCIPLES TO GUIDE THE PRESERVATION OF SCIENTIFIC DATA

*Paul F. Uhlir**

National Research Council, Washington, DC, USA
Email: **puhlir@nas.edu**

I would like to focus on three problem areas that are alluded to in the title of this essay. They all have to do with the social, rather than the technical, systems in research that undermine data preservation functions or the subsequent exploitation of the data for the benefit of research and society.

I refer to the first of these problems as "information gulags." By this, I mean the incarceration of large numbers of data resources in dark repositories, to be manipulated and viewed only by their masters. Such inaccessible warehouses of information, which have also been referred to as data cemeteries, unfortunately tend to be the norm at both the individual and institutional levels in public research throughout the world. Many researchers, particularly in disciplines in which the research is done by single scientists or small teams, keep the data they generate as closely held and jealously guarded private resources.

Such secrecy may be warranted in research areas that have inherent commercial potential or that may have other restrictions based on the protection of individuals' privacy or national security concerns. However, in large swaths of government or publicly funded research, there are no legitimate reasons to keep the data under lock and key, especially once the researcher has analyzed the data and published the results. Similarly, public research institutions that manage data centers or archives of accumulated data have a duty to make those holdings available as openly as possible for common use by the community, subject only to well-established, legitimate exceptions. This brings me to the first principle:

The value of data increases with their use.

Technocrats are fond of measuring value in terms of the cost of the instruments or the research that they fund. The data that are produced have traditionally been considered as by-products of the research, ancillary to the original investments or the final outputs of the research enterprise. Such a view is based on the uses of data under the old print paradigm, prior to the rise of digital networks. The vastly increased potential of digital data in the networked environment is now well known, however.

Based on the expanded capabilities of electronic data, it is essential to change the obsolete mind-set of data as a by-product of the research process that can be thrown away after the project is completed. Many types of scientific data should now be viewed as a fundamental infrastructural component of the modern research system, whose value is only increased with broad access and reuse by others. In particular, the growing field of knowledge discovery from automated search, extraction, and integration of data from diverse sources online is only made possible through open access.

To do this correctly, it is necessary for the data managers to take a user perspective and to optimize the data for dissemination and reuse as well as for long-term preservation value. This therefore also means that the data must be sufficiently organized and documented to be of broader utility, a non-trivial consideration.

*** * ***

Life plus 70 years. If this sounds like a long time, it is. In the United States, such terms are typically given to murderers, so that there is no possibility of release. But life plus 70 years is also the current term of copyright protection, which grants an exclusive property right to the author (now usually the publishing intermediary). Works

* The views expressed in this essay are those of the author and not necessarily those of the National Research Council.

for hire in corporations are protected in countries like my own for terms of 95 years, regardless of the life of the author—or the corporation.

It wasn't always like this. When copyright was first instituted in the United States, it was only for a period of 14 years. Such a monopoly right was viewed as a necessary evil that was conditionally granted to promote the progress of science and only for a limited period of time. The right also did not attach automatically; instead, the author or publisher had to file for it.

In what is one of the great paradoxes of the digital age, however, as the pace of innovation has quickened, copyright and other intellectual property rights have become longer, broader, and stronger. This now leads to absurd results. You may remember using Multimate word processing software and Lotus 1-2-3 spreadsheets. This was a mere 20 years ago, but that is eons in the digital age. I am sure you will be delighted when this software enters the public domain sometime around the end of this century, unless of course the term of protection is extended further still, which seems quite likely given legislative trends. Of course, the creators of software programs such as Multimate and Lotus 1-2-3 (and many others) do not need to have exclusive monopoly rights for many decades in software that becomes obsolete long before the copyright expires.

At the same time, copyright offers only so-called "thin" protection to databases because the contents of many factual collections lack the requisite creativity and originality to qualify for copyright protection. Thus, additional legal protections have been devised. Many databases and other forms of information are now licensed, rather than sold, on terms that are more restrictive to users that the exclusive rights under copyright law. Restrictive licenses can override the exceptions and limitations that are granted to the user under copyright law. The restrictions can be enforced automatically with a variety of technical protection measures. And the restrictions can be imposed in perpetuity because the license is a private agreement among the contracting parties that is not limited by public legislation.

Moreover, since the late 1990s, a new type of statutory exclusive property right has been implemented in the European Union and some other countries, which protects investments in the non-copyrightable portions of databases as well. This database protection law has many characteristics that are unduly restrictive on users of factual information, especially those serving the public interest. The law imposes an unprecedented absolute exclusive property right in anything more than an insubstantial amount of data. It provides for a potentially perpetual term of protection, as long as the database continues to be updated. The law applies to government and government-funded databases. There is no required public-interest limitation for activities such as public research, education, and library access and use. And there are excessive penalties for infringement, which is triggered whenever the user extracts more than a quantitatively or qualitatively insubstantial part (whatever that means) of the database and is not dependent upon proof of any economic harm.

When taken together, the statutory intellectual property laws and restrictive private licenses impose intellectual straightjackets on users of facts that were historically in the public domain. The laws and contracts have been optimized for multinational content intermediaries who have in many cases usurped the rights of the original authors and have merely cloaked themselves in the mantle of the authors' rights. While the hyper-protectionist intellectual property laws favor the interests of large multinationals over the consumer and small business owner, they are especially pernicious when applied to government and publicly-funded research data. This leads us to the second principle:

Public (and, to a large extent, publicly funded) "…information wants to be free…"[1]

What do I mean by that? Everyone knows that there is no free lunch. However, in economic terms, information has qualities that are known as a public good. A pure public good has two characteristics. One is that the good is not depletable. That is, it cannot be diminished by use. Unlike a physical object, information is non-depletable, and to the contrary, as my first principle asserts, the value of data increases with use.

[1] The full quote, attributed to Stewart Brand (1984), was: "On the one hand information wants to be expensive, because it's so valuable. The right information in the right place just changes your life. On the other hand, information wants to be free, because the cost of getting it out is getting lower and lower all the time. So you have these two fighting against each other." See: http://en.wikipedia.org/wiki/Information_wants_to_be_free.

The second characteristic of a public good is that it is non-excludable. This means that it is not possible to capture exclusively the benefits of that good by keeping it from others. Information can be excludable with difficulty. It is completely excludable if treated as a secret, either as a national security or trade secret, or for protecting personal privacy. However, once the secrecy or privacy is breached, it is no longer excludable and can be communicated more broadly. The cat's out of the bag. Information that is subject to various statutory intellectual property rights and licensing restrictions also is intended to be excludable, but at the same time such commercial proprietary information is typically made available to the public. Although this is done on an individual customer basis, strict excludability becomes difficult to control and enforce.

In the case of the government provider, the rationale for the control and enforcement of exclusionary policies for non-sensitive data and information is much less compelling than in the private sector and can be viewed as contrary to the public interest. This is especially true in the context of a government function such as basic research, which also has its own public-good and public-interest characteristics. Although research can be partly excludable as well, it too is non-depletable. The rationale for artificial excludability of research or other data produced and maintained by government entities is either weak or non-existent, especially when balanced against the positive benefits that can be derived from free and open dissemination online.

Government funded data and information, such as that produced through publicly-subsidized research, lies somewhere in between the government and private sector. It is non-depletable and partly excludable, and the presumption for its excludability will vary, depending upon its characteristic as a national security, privacy, or potential commercial good.

The public good and public interest nature of government generated and government funded scientific data makes the presumption of free and open access on digital networks the most effective and efficient approach. It maximizes the returns on the public investment in generating those data. The value generated to science, society, and the economy from free and open access to public data will in most cases outweigh a restrictive, proprietary approach to disseminating such information. Conversely, the artificial exclusion of potential users, either through the charging of high rents for access or through the imposition of controls on reuse, will work directly against both the first and second principles. Thus the maintenance of information gulags and the application of excessive intellectual straightjackets to public research data are likely to have many negative effects, as I have described in earlier articles and just summarize here.

First are higher research costs. Many factual databases cannot or should not be independently recreated, either because they contain observations of unique phenomena or historical information or cost a great deal to generate. Moreover, databases with a monopoly status that are maintained on a closed proprietary basis will tend to result in higher, anti-competitive pricing. Managing government or publicly funded databases on a restrictive, proprietary basis also adds substantial administrative overhead on both ends to make each transaction, further taxing the public research system. This is particularly obnoxious when public institutions license data at high costs and restrictions to other public institutions.

Second, there are many lost opportunity costs. Perhaps not as obvious, there is much less data-intensive research possible if the publicly-funded data with reuse potential are not preserved or made easily available. This results in significant lost opportunity costs that are certain to occur but are difficult to measure.

Third are the barriers to innovation. The production of downstream copyrightable or patentable intellectual goods by both the public and private sectors depends to a large extent on access to the free flow of upstream public factual data and information. The overprotection or unavailability of public databases leads to what economists refer to as "deadweight social costs" resulting from sub-optimal use, thereby taxing the innovation system and slowing scientific progress.

Fourth there is less effective cooperation, education, and training. A failure to preserve research data and to make them easily available or erecting barriers that are too high necessarily results in less effective interdisciplinary, inter-sectoral, and international cooperation.

Finally, there is the potential for widening of the gap between OECD nations and developing countries and between the richer and poorer institutions or researchers within countries. Developing countries are particularly disadvantaged by a lack of availability or high barriers to access. Unnecessary access and reuse barriers to publicly funded research data, and especially to data produced and held directly by the government, thus result in diminished returns on the social and scientific capital investments in public research and in the inefficient distribution of benefits from those investments, even as the improving technological capabilities offer ever greater opportunities to increase such returns.

The statutory hyper-protection of digital information, including that of public data, can be addressed in several ways. One is by properly balancing the existing or prospective intellectual property rights. This means pushing back on the demands of the high protectionists from OECD countries who seek to enhance their existing control of information markets. A second related approach is to carve out broad public sector and public interest exceptions from the intellectual property laws that otherwise protect the rights and interests of the private sector. A third approach is to adopt Creative Commons "common use" licenses that use IP laws to enforce only "some rights reserved," rather than the full statutory rights. This private-law approach has been successfully adopted for promoting open access and reuse of publicly-funded scientific articles throughout the world and is increasingly being considered for government information in those jurisdictions that copyright public information. Nonetheless, it should be noted that the application of Creative Commons licenses to databases and data products that are not fully copyrighted remains uncertain and needs to be examined more closely.

*** * ***

We now turn to the third policy issue in digital preservation, that is, the need to avoid "memory holes." This problem can be both intentional and unintentional but is one that has the most severe consequences to preservation. Unlike the first two problems outlined above, its negative effects can be complete and irreversible.

The term "memory hole" comes from the novel *1984* by George Orwell. You may recall that in the totalitarian dictatorship that Orwell depicted, the state exercised complete control over all information and modified it at will to suit its own nefarious purposes. As Orwell wrote, "Who controls the past controls the future. Who controls the present controls the past." The chief protagonist in the novel, Winston Smith, an apparatchik in the "Ministry of Truth," would delete information from the official record that was deemed inconvenient to the state by jettisoning it into a vacuum tube aptly called the memory hole. This would change the historical record, allowing the state propagandists to manufacture their own version of the "truth" for public consumption.

Although such manipulation of information is most pervasive in dictatorial regimes, whether at the extreme left or right of the political spectrum, all governments engage in such practices to various extents. They do this through techniques such as the spurious categorization as top secret or confidential of politically embarrassing official information, the revision or falsification of data, or the intentional destruction of public records. Of course, the difference between a democratic and totalitarian regime is that in a democracy, sooner or later such practices tend to get publicly exposed and discussed, whereas in a dictatorship they only become publicized at great personal risk to the messenger.

But there is another aspect of the memory hole syndrome that merits even greater attention and vigilance. That is the inadvertent loss of data. As we all know, electronic bits are ephemeral. And we are all aware of some of the horror stories that abound.

Take the case of NASA, the National Aeronautics and Space Administration in the United States. That agency is generally regarded as a paragon of technological achievement. It successfully sent man to the moon at the dawn of the digital age. It developed robots to explore the heavens and invented all manner of gadgets. So, of course, NASA has preserved the record of all its great achievements. Wrong. It turns out that the agency did not preserve the data from its very first mission, Explorer 1. It does not have many of the original tapes from the human exploration of the moon. It has lost much of the early Landsat data, which otherwise form such a valuable longitudinal record of our planet's environmental changes. The list goes on.

My purpose is not to pick on NASA. I like NASA. Many of my most admirable professional colleagues are there. The agency now does a commendable job of preserving data from its space missions, but it has not always been so.

The point I wish to make is that if one of the most technologically advanced organizations on Earth has forever lost a great deal of highly valuable, irreplaceable data from just a few decades ago, what has been going on elsewhere?

NASA's losses are not unique—far from it. The problem is that the losses are hard to measure. They are highly distributed, silent, and invisible. Despite the great uncertainties of their exact nature and extent, we can be sure that the losses have immense, and sometimes tragic, consequences. According to the UNESCO Digital Heritage Program, the "enormous trove of digital information produced today in practically all areas of human activity and designed to be accessed on computers may well be lost unless specific techniques and policies are developed to conserve it."

This brings me to the third and final principle:

"Digital resources will not survive or remain accessible by accident"[2]

This means that lifecycle planning should begin before each data collection project. Fortunately, the digital preservation imperative is now quite well recognized, even if it is not comprehensively addressed. But it is an especially acute problem in the developing world, where the competition for scarce public resources makes it a particularly vexing issue.

Individual researchers and public data centers and archives thus have a great responsibility. They need to preserve and make available as freely as possible the relevant factual information that is increasingly being generated in digital form about all aspects of their human and natural environment. So as they move forward, they should avoid the traps of information gulags, intellectual straightjackets, and memory holes.

---

[2] Quote from presentation by Bernard Smith, Commission of the European Communities, at the ICSTI/ICSU/CODATA Digital Preservation Workshop, Paris, France (2002).

(Article history: Received 17 August 2010, Accepted 8 September 2010, Available online 23 September 2010)