The Technical IS Political

This article originally appeared in *Of Significance. . .* Vol. 3, No. 1 (2001), "Orderly Chaos: Effects of Changing Data Products." *Of Significance...* is a topical journal of the Association of Public Data Users. Reprinted with permission.

JIM JACOBS
AND
KARRIE PETERSON

Abstract. In the realm of government information, technical decisions about data format, access software and public distribution methods are inherently political decisions. They affect what kind of data can be accessed, how, by whom, and for how long into the future it will be available. To evaluate and respond appropriately to policy changes by government producers of data, technical issues must also be looked at in the light of social values shared by the data-using community.

How does a newly decked-out data product fare with regard to open access? Privacy of individuals? Documentation that allows the data to be correctly cited, tested for reliability, re-used in the future? Social and political concerns also come into play when the flexibility offered by distributing raw data is balanced against locking the data into a "user-friendly" software, and when products traditionally produced by the federal government are privatized. As private industry pushes harder for information to become a commodity—something that can be sold for profit—it is important for data users to push back with a strong philosophy of information as a social good, and to evaluate data products and access in light of their value to society, rather than on strictly narrow technical grounds.

The advent of the information infrastructure and the enormous opportunities it offers for citizens to access information has led to both the expected increase in access and, paradoxically, to some situations of sharply diminished access.[1]

> [W]e dare not think that, just because we can do good things with technology, it therefore threatens no radical and unconsidered change.... What looks like freedom in the short term may constrain us in the long term. Much the same applies to computers. We have to look beyond the immediate activities we are inclined to praise or curse. We have to grasp the underlying forces—human as well as technological—through which society is being reshaped.[2]

How could it be, given grand technological advances and tremendous cumulative experience in collecting, presenting and preserving data, that we sometimes have less access to data than before? Progress in technology does not automatically translate into more and better access to public data. While technology can make many things possible, it is social and political decisions that make them happen—or not.

Political decisions affect how technology is used, what kinds of data are gathered, methods employed in collection and analysis, which data products are created and how they function, and what data are made publicly available.

It is unquestionable that public data is easier to locate and use than it was even just a few years ago. There are new audiences for data because of this development. There are increasing expectations that the government should use technology democratically, by increasing access to civic data and services. Technologies are, meanwhile, advancing rapidly and becoming extremely complex—often with implications that are beyond the everyday understanding of average citizens. And there is increasing awareness that because data is socially valuable, there is money to be made in packaging and selling it.

Jim Jacobs has been Data Services Librarian at University of California San Diego since 1985 where he provides data services for faculty and students. Since 1990 he has co-taught the ICPSR summer workshop on providing social science data services. He has been on the APDU Board of Directors and the Administrative Committee of the International Association for Social Science Information Service & Technology.

Karrie Peterson is a recent graduate of the School of Library and Information Science at the University of Pittsburgh, and she worked briefly in the university library system there in government documents and GIS. Currently she is the depository librarian for U.S. government publications at the University of California, San Diego and regularly teaches class sessions in Statistical Literacy to new students.

Poor quality data products or problematic access can result from lack of technical expertise—it is simply very difficult for data publishers to know everything they need to know to produce the best product for any given audience. Traditionally users in the data community are consulted and do provide public feedback in this type of situation, and that relationship will no doubt continue. But this is not the type of problem we focus on in this paper.

Our attention here will be given to new problems and issues in the provision of public data as they arise from the collision of political and social agendas with new technologies. We focus primarily on public data, and by that we mean data collected, published and disseminated by governments. We believe that the data community has an important advocacy role to play in ensuring that public access to data is enhanced and not diminished by technological advances. At the same time, we believe that it is necessary to first recognize underlying political issues. There is little likelihood that a better data product or service will result from lengthy technical critiques when the flaws are a consequence of political tradeoffs or special interest agendas.

Thanks to considerable professional expertise, the data community—composed as we are of producers, distributors, archivists, academic researchers, policy setters, citizen activists, and so on—is uniquely poised to be a watchdog in areas where society could end up on the losing end of the stick when it comes to data access. In examining a few recent trends connected with new technology, we discuss where and how intervention and advocacy might be called for.

Costs

The political and the technological collide most visibly in the area of costs. Technological solutions to data dissemination and access are expensive and, when a government decides how to allocate its budget, it is nearly impossible to avoid political considerations. These may be as simple as having to decide among many worthy projects ("should we cut back on milk for school children or increase the data budget?"), or they may directly express an information policy ("Let's outsource our data distribution!"). The added expense of new technology adds new competition to limited budgets. Oddly, it is "cost savings" that are often cited as a reason to create new data products.[3] These cost savings, as you might imagine, are largely illusory, but the negative effects they can have on data access are real. Often, "cost savings" are really "cost shifting."

For instance, cost can be shifted from the data producer to the user of statistical tables as when printed statistical publications are replaced with computer-based products or online access. The users bear the new cost in purchasing more and better computers and in actually printing when a paper copy is required. This is not a trivial problem. A recent study found that, although 58% of American households have Internet access, "a digital divide remains or has expanded slightly in some cases, even while Internet access and computer ownership are rising rapidly for almost all groups."[4] Thus, while a shift to internet access for statistical tables is a good thing and brings data closer to many, the very technology that enables this change can, at the same time, increase costs for access to individuals or to libraries that serve those individuals.

A recent GAO report makes this even more apparent.[5] In a survey of Internet users in the U.S., the GAO found that few homes have fast, broadband (DSL or cable) access. The survey found that "[t]he conventional telephone line was still the most common method of transport to the Internet" and that about 88 percent of respondents use this method. Further, the report found that the "survey results support the perception that access to and use of the Internet are influenced by a person's race, education, and income level." As we evaluate government web sites such as the Census Bureau's American FactFinder that have large pages, interfaces that require the most current browser and require the page to reload with each choice a user makes, and data pages as large as one megabyte or more with an estimated time to display of more than 3 minutes in some cases,[6] we cannot help but think that the advances of Internet access are not reaching most Americans.

When the funds are not available to libraries or individuals to upgrade their Internet access to the high end that agency sites seem to prefer, the data will simply not be available to those users.

Other more subtle forms of cost shifting can occur when we replace traditional distribution of print materials with new data products. The government can save money (taxpayers' money, of course) by replacing paper and ink with CD-ROMs or even Internet access, but the user who is best served by a simple table in a book or periodical may actually find using a computer-based product more time consuming and more confusing. The cost shifting is from the government budget to the individual's time and perseverance. Sometimes, reluctance to adequately test data products—an expensive and labor-intensive process—results in defective products being released that can never be used. In addition, unless the data product is designed carefully, information present in

the printed publication, such as footnotes and sources and formulas, may be unavailable or difficult to use in the computer product. The user may not be able, for instance, to ensure that the table found in the computer product is the newer version of one previously printed in a book. Again, if the individual does not have the time to spend with the new product, or if the new product has sacrificed some detail, the new product has created a net loss rather than gain.

These are examples of how a political decision, cost savings, can drive a technical decision and result in new costs and new barriers to access to some users. The decision to move to a new technology does not, of course, *require* "cost savings" or cost-shifting. But the implementation of the technical decision is political and the results are real.

Another use of "cost savings" is as a rationale or excuse for making a change in priorities such as an explicit shift away from data dissemination.[7] Although, we have seen less of this so far, we anticipate that we will see more agencies shying away from data distribution as other issues politicize the process even more. We will examine in later sections how the issues of privacy and privatization will create thorny problems with political implications.

Costs can also be shifted to the data analyst. When raw data, publicly distributed, is replaced with an easy to use product that will produce simple tables but not supply the underlying raw data, the analyst will have to seek the raw data from more expensive commercial sources. There is no technical reason why a simplified, easier-to-use data product must*preclude* providing public access to raw data, but it sometimes happens and, with increasing costs and competing priorities, may occur more often in the future. The data community must insist that the government continue to provide free or at-cost distribution of raw data to researchers, data archives, and digital libraries as an essential complement to new easy-to-use data products, or we will suffer either increased costs or a loss of our primary research materials or both.

The cost of preservation of data and long term access to data may also be overlooked when cost savings drive new data product development. Preservation issues are examined in a later section.

The cost shift can also be from one data producer/distributor to another, or from one budget to another. When an agency decides to make data available on the web, for instance, it may save money initially (on printing costs, on cost of answering phone inquiries, even on FOIA responses). But new, perhaps unanticipated, costs of the new technology soon appear. In a fascinating presentation at the 1999 Annual Symposium of the American Society of Access Professionals, an EPA official described with great enthusiasm the way EPA is able to use its web site to make community information available. Then she noted that there was an unanticipated and unfunded cost of three million dollars for this activity.[8] This gives us an uneasy feeling. While we appreciate the enthusiasm of the EPA and the ease of use of data at their site, we cannot help but wonder, once the real cost of the new data product (web access to EPA data) is addressed, will the cost be met? If not, what access will we have?

Privacy

As we all know, the guarantee of confidentiality to survey respondents is a foundation for survey research. If we cannot provide that guarantee, accuracy and reliability of surveys will decrease drastically. Government agencies such as the Census Bureau have a long and admirable history of protecting the privacy of individuals and businesses who report personal or proprietary information to the government.[9] While we assume that this principle will continue to be widely respected by government data producers, we see many new problems emerging that will complicate and endanger it.

New technologies, particularly those that provide easier access to data, are making it harder to ensure privacy. Networked technologies for collecting, storing, and disseminating public data make it more difficult to protect confidential information from accidental or inadvertent exposure.[10] In addition, data about individuals can now be integrated, compiled, and linked between agencies and across programs in ways never intended when the information was originally collected. Public agencies can easily collect information about users of agency web sites and can do so without the consent of the person. These risks are not new, but they become more severe with each new technological advance.

In the political arena, privacy and access are often portrayed as two irreconcilable options: to allow data access and give up privacy, or deny data access as the expense for maintaining privacy. These can be highly charged political questions, but the data community knows that privacy can be protected without denying access to data. We have an important role to play in keeping the public and policy makers informed. Politicians and technocrats will find it easier to argue the hot-button issues than grapple with the technical details necessary to provide both access and privacy protection. Here we examine some of the issues and some recent developments.

It is technology that makes it easier today than ever before to reuse and "re-purpose" data but it is

the decision to do so that politicizes the technology and the data. For example, genetic testing to determine the presence of congenital disease is mandatory in some states because it is socially beneficial when used to catch and begin treatment early on in a child's life. Could events transpire that result in insurance companies or employers gaining access to those databases, legally or otherwise?[11] Personal data on health is fast becoming a valuable commodity. As we write this, in fact, a lobbying effort described as "one of the largest ever undertaken to block implementation of a federal health care regulation" is underway.[12] Hospitals, HMOs, insurers, and pharmaceutical companies are lobbying to scale back patient privacy regulations. If business sees encouraging signs in cases like this one in the field of healthcare, they may also seek to gain access to personal information from public data in a similar way—by attacking privacy policies. Protecting privacy and maintaining continued access to anonymized data may require informed political action and advocacy by the data community.

A second privacy issue that government has not yet dealt with adequately is the collection by government agencies of personal information about individuals as those individuals communicate with the agency through computer networks and agency web sites. The technology of the web currently enables and encourages collection of information about users of web sites.[13] Although these technologies are not inherently privacy invading, they do enable and simplify the collection of personal information.

One recent study of 1,813 state and federal government websites found that only seven percent of government websites have a privacy policy[14] and General Accounting Office has reported that only three percent of federal governmental websites surveyed implemented elements of all four of the Federal Trade Commission's fair information principles for Internet privacy.[15] Although there are nascent guidelines for federal web sites, regulations and enforceable policies are not generally in place. After all the publicity about the FBI's "Carnivore" (now called "DCS1000") software,[16] it may not be easy to convince the public that their privacy will be ensured even when policies are in place. Controversy about this kind of privacy issue poses a problem for collection of data that is useful to the government in providing services, designing better web sites, allocating staff and resources, and so on.

A third privacy issue is that new technology enables governments to take personal information that has been collected legitimately and exploit it for purposes that were not intended at the time of the data collection. The "linking" of data from different sources is a prime example of this phenomenon.[17] There are forces driving government toward the re-utilization and re-purposing of personal data already gathered. For governments to become more efficient they will need to be better about exchanging data between agencies and among programs. This will be a natural outcome of the move to "e-government."[18]

Governments want the utility of systems and databases that are compatible and allow for efficient storage and interchange of information. There will be many advantages to the public (fewer forms to fill out, consistency of information across agencies, streamlined applications and services) when this is done well, but the danger of misuse of the data is real.[19] The result of linking databases and other stores of information across programs and agencies can have serious privacy implications.

Consider these developments. At least one proposal for explicit authority to link databases emerged last year when the Congressional Budget Office sought to link data records on the same set of people from the Internal Revenue Service, two Census Bureau surveys, and the Social Security Administration in order to do sophisticated economic modeling.[20] Recently, the Environmental Protection Agency, in order to enforce the Clean Water Act, sought data collected from dairy farmers by the Natural Resources Conservation Service of the Department of Agriculture. This prompted one commentator to say, "Suddenly a beneficial service was being forced into the role of unwilling informant. Clearly, any disclosures by NRCS would undermine its credibility with its farmer clients."[21]

As technological advances make it easier to merge, link, and reuse data gathered for administrative purposes or from survey research, the tension between government efficiency and the confidentiality of personal information will increase and be difficult to resolve. This will influence both data collection and access. The GAO report cited above is a good start to addressing the problem of privacy and data linking formally, but the study explicitly excludes data linking projects that "are intended to result in actions toward data subjects (e.g., federal compliance audits)."[22] Although it will be difficult enough to address the issues of linking data collections for the purpose of compiling aggregated statistics, the more difficult issue will be the one the GAO report avoids: that of using linked data for enforcing laws and regulations. And, whether such uses are prohibited or not, the public perception of such uses may have a strong effect on our ability to gather raw data. Finally, once such linked collections become more common, it will be politically as well as technically easier to use

them for purposes other than simply compiling aggregate statistics.

Recent developments have shown that, even when the government is prohibited from compiling dossiers on individuals, agencies can and do purchase such information from private sector companies. This demonstrates still another difficulty; if agencies rely on the private sector for information about individuals, there is no public accountability for the accuracy of the collection, aggregation, dissemination, or use of such data. Use of such (inaccurate) data has resulted in individuals being fired from their jobs and purged from voter rolls.[23]

The data community is in a unique position of understanding both the technical and the ethical issues related to privacy and so can make a positive difference in the debate. One hopes that there will always be opportunities for public input when technology deployment decisions could result in privacy violations. But when the data community is not invited into the discussion, we may be faced with ethical questions about whistleblowing, finding either ourselves or our colleagues in that role because we have a professional understanding of the implications of technical decisions that may be too complex to catch the attention of the media or the public.

Privatization

> [Consider] a society in which commercial goals are achieved efficiently with electronic technology, but in the process, free access to information as a social commitment goes by the wayside. Only data that has commercial value will be collected and retrieved. [24]

The struggle to keep public information publicly available is not a new one.[25] The key players here are the public, who would like free access personally and professionally to everything that has been produced at taxpayer expense; information industry businesses who would like to get exclusive control over portions of free government data and sell it; and the government, often caught in between laws ordaining free access and the impulse to fatten the agency's budget by selling data or privatizing its collection and dissemination.

In the sense we use the term 'privatization,' we mean it as a generally negative occurrence when data that used to be collected, published, and distributed for free or at marginal cost by the government are replaced by commercial, for-profit products. (There is also a different, symbiotic relationship between the information industry and government, in which businesses complement rather than replace government data products by acquiring government data, enhancing and improving it, and then selling it. The value that the private sector typically brings to government information ranges from convenience of packaging and distribution to superior indexing and organization.)

Technology heightens the struggle between public access and privatization in a subtle way as new audiences for data emerge wanting and expecting advanced data products. Better data-access products along with a widespread availability of desktop computers and software have created these new larger audiences for data than existed in the past.

This fundamentally changes what it means to provide public access to public data. At one time, as recently as ten or fifteen years ago, government distribution of raw data to data archives and researchers was considered adequate, but today a new audience for data expects to have access at their desktops. People in search of quantitative information, who used to rely on statistical material tabulated by the government, are becoming more attuned to using data via simple user-friendly software that enables some analysis. This group of newly interested users includes small business, academic researchers, students, public interest organizations, professional associations, community groups and so on. Traditionally, the government went to great effort and expense to provide these sorts of users with tabulated statistical publications, and more recently with CD-ROM-based data and accompanying software. As online access replaces print statistical publications and other formats, it's clear that this category of data users will continue to have access to public data they can actually use only if "front-end" software is provided. Will that software be provided free, as a government responsibility to share data it collects at public expense and uses in conducting government affairs and setting policy? Or will these data users "pay twice" for their data by having to get usable chunks of it from the commercial sector?

The information industry has been quick to claim turf in answering that question, maintaining that provision of "services" and "adding value" to data should be the role of the private sector and that the government must not compete with commercial enterprises. A recent study commissioned by the Computer & Communications Industry Association (CCIA) said, "The government should exercise caution in adding specialized value to public data and information." In a summary of findings, the report specifies proper roles for government as providing public data, improving the efficiency of governmental services, and supporting basic research. Beyond that,

the report sees only "yellow light" and "red light" areas that governments should enter with caution or avoid entirely; these are the areas that the information industry wants to limit to commercial exploitation. Governments, the report says, should not only avoid adding specialized value to public data, but should also avoid providing on-line services, avoid markets where private-sector firms are active, avoid maximizing revenues and avoid taking action that would reduce competition. The report concludes that "…existing norms for government provision of goods and services need to be updated for the digital age."[26]

In short, the report suggests that governments should collect data and distribute raw data, but should not create data products that make the data easier to use. It is quite probable that the next decade will decide whether the government continues to view public data as a social good for which they have a responsibility to provide broad, convenient public access, or the information industry prevails in portraying data as a commodity that dwells in the zone of for-profit enterprise. Certainly the data community can have an influence in the outcome.

The potential for data loss is also a consequence of overt decisions allowing privatization in the collection and publishing of what used to be public data. As noted above, governments are responding to this same need to update the "norms for government provision of goods and services" by moving increasingly toward a new model of digital government, but these are expensive services. Many local governments have already coped with the expense by contracting with private firms to manage their data and the contractors are demanding exclusive rights to the distribution of these government databases.[27] A report on this by the newspaper industry was bluntly titled, "Government for Sale."[28] As governments attempt to provide better services digitally, these conflicts between ensuring public access to public data and privatizing public data will increase.

Privatization of data creates new problems. For instance, when governments collect data, they can be held accountable for content and methodology, but commercial data vendors cannot. The recent controversy over the content of Consumer Price Index[29] demonstrates how data can have political content and implications and should warn us of another danger of turning over to the private sector the collection and production of what should be public data. Also, when print publications are privatized, a customer buys and owns the publication. But in the electronic realm, privatization more often means that "access" will be leased rather than data purchased and this raises a whole host of questions for data users about long-term access when providing the data is no longer profitable to the company. These issues include copyright and fair use, and the likelihood that data will be attached to some proprietary software.

How common is privatization? We have seen the privatization of print publications for some time. (Notable examples of government publications that became private publications include *U.S. Industrial Outlook* and *The Journal of the National Cancer Institute.*[30]) Privatization of public data is less common, but not unknown. For instance, in 1995, the U.S. Bureau of Economic Affairs transferred the production as well as the distribution of the "Business Cycle Indicators" to the Conference Board.[31] What was once freely distributed to depository libraries is now a commercial product. The "National Economic, Social, & Environmental Data bank" and other data-CD-ROMs and statistical publications that were once distributed free to depository libraries are now available at the government fee-based web site "Stat-USA," with only limited free access.[32] (Although this is not "privatization," since the government still provides the data, it is an example of the government treating information as a commodity and selling access rather than providing it free or at-cost.)

How likely is it that privatization of data will continue and expand? As the collection and distribution of data become more politically charged, because of issues of privacy and expense, the "easy-out" for many public data producers will be to wash their hands and turn over the problems to the private sector. When this happens, the symbiotic relationship is destroyed and the balance is upset. As noted above, Herb Schiller, an expert in communications as they affect society, succinctly warned that when this trend reaches its zenith and the profit motive is the force behind collection, publishing and preservation of data, " [o]nly data that has commercial value will be collected and retrieved."

Responding politically as these issues arise will be a challenge in the data community, as our members come from both the public and the private sector. More discussion and debate among ourselves is quite likely as this philosophical issue plays out.

Sensitive—Not Classified

> The FBI has also requested computerized check-out records from technical and science libraries and has asked private information providers, including Mead Data Central and Charles E. Simon Co., to help monitor use of their databases. Although public and university libraries do not have

> classified information, the FBI has justified its interest in library use by a version of the "information mosaic" theory: that discrete and benign pieces of information can be put together to present a danger to national security and therefore need to be controlled.[33]

As odd as it sounds, public data can become less accessible if it is easier to use. This was exemplified in the mid-1980s when a series of decisions endangered public availability of public data. In 1986 National Security Advisor to President Reagan, John Poindexter, signed a memorandum that gave federal departments "broad new powers to limit release of government data and created a new 'sensitive' classification to restrict access to national security related information." Under the authority of National Security Decision Directive 145 of September 1984, the memorandum was attempting to protect data that, although not classified as secret, might be of use to foreign powers. At the time, the government was concerned that even commercial databases of news stories, such as Mead's Nexis, could provide "sensitive material" of use to foreign powers. These policies were driven by the conclusion that the power of being able to access public information through electronic databases made acquiring and compiling information so easy that non-classified materials became "sensitive" and should therefore not be allowed to be in those public databases.[34]

Although this theory of information control largely disappeared for almost a decade, it is resurfacing now that data products make information easily available to more people.

Some controversies over the suppression of public data raise the speculation that, especially in the regulatory arena, government officials may be bowing to pressure from industry not to make certain kinds of data so very widely accessible. Some of us will remember, for instance, the Chrysler minivan flap in 1995 when the government declined to issue a recall order but also refused to release the investigation data that included videotapes of latch failures in crash tests — "data" that would surely have appeared across the nation on network news. Vociferous public protest changed their minds about public access to public data four days later.[35] And in 1997, the Federal Aviation Administration announced that it would disseminate airline safety data on the Internet but that data on maintenance violations and engine trouble would not be included because FAA officials felt that "certain FAA data was prone to misinterpretation."[36]

Just this year, the Environmental Protection Agency has issued draft rules and a call for comments implementing the *Chemical Safety Information, Site Security and Fuels Regulatory Relief Act of 1999*.[37] The law amended the *Clean Air Act* to keep chemical industry risk management plans with chemical releases worst-case scenario data off the Internet. The EPA's *Federal Register* announcement said that "concerns were raised that potential Internet distribution of [portions of these plans] would pose law enforcement and national security risks."[38] Environmental activists argued, of course, that the far greater social benefit of access to data about local environmental hazards outweighed the putative national security issues.[39]

A final current example of the connection between the technology of data access and the availability of public data involves a medical malpractice database, established in 1990 under the *Health Care Quality Improvement Act*. Legislation was introduced last year, and similar efforts will likely be repeated, to give free public access to this data so that people can inform themselves about their doctors. Opponents of opening up the database to the public have claimed that it would "raise more questions than it would settle." The ability to get this kind of data on to the average person's desktop gives rise both to the impetus to do it, and the fear of widespread access.[40]

If this trend continues, as we expect it will, we will see more of these cases of withholding public data for the alleged protection of the public. As technology makes it easier to use data, these issues of "sensitive" data will come up repeatedly and they will endanger data collection and access unless we recognize the politics behind many of these claims and call public attention to them.

Permanent Access: Who Gets Control?

In the print world we enjoyed semi-automatic preservation because everyone had to get their own copy and because paper lasts a pretty long time resulting in lots of copies on a stable medium. It has never been as easy to preserve computerized data for the long term. As has been pointed out on the pages of this publication,[41] mistakes have been made in preserving digital data and, consequently, we have lost data. We have a much better awareness of the need for and the difficulty of preserving data today. Although, for the moment, the procedures and standards that can ensure preservation and permanent public access to public data do not yet exist, two things are clear. First, we can preserve today's data adequately until the time that longer-term solutions exist. Second, digital libraries and data archives have good tools now and will have better tools in the future for conversion, migration, and preservation of data.

But, in addition to the technical aspects of

preservation, there is also the issue of long term access. The federal government provides for the *preservation* of data in the National Archives, but NARA does not (at least at this point) provide broad public online access to data. We will focus here on the long term access issues.

Along with the shift from print to electronics, and with the shift from the distribution of data to the "access" of data on government servers, there has been a shift in who has de facto responsibility for providing permanent public access. Once, libraries and data archives kept print materials and data for long term public access; now, these new government policies of data access are shifting much of that responsibility to the government. The old model of wide distribution and many collections ensured long-term access through many different institutions, with many different funding sources, serving many different constituencies. The new model puts the burden of long term public access on the shoulders of either the agency producing the data and offering access to it through their web site, or the Government Printing Office which collects some data for its Electronic Collection and is engaged in hammering out long term preservation practices.

Not only is the concentration from "many collections" to a single collection (e.g., each data set held only in an agency "collection" or at GPO) problematic, there is also the problem that agencies do not have policies, or funding, or even a congressional charge to provide permanent public access. Agencies will need to seek funding for this in addition to seeking new funding for short-term access. Even if funding is adequate for many years, long term access cannot be ensured because agencies are subject to political pressures, future budget cuts, even dissolution.

When the government is in sole control of access to data, short-term and long-term access are both endangered. Data posted to an agency web site can be easily yanked for political reasons and experience shows that without significant protest, litigation, or both we may not see the data again. In 1998 for instance, after the House Commerce Committee posted a 104-page memo describing how the lawyers for a tobacco company suppressed research on the health hazards of smoking over a thirty-year period, the tobacco company objected to it being made public. The committee removed the memo from its web site and withheld another 400 documents after the company said they contained trade secrets.[42]

A recent incident is a shocking and overt example of two of the dangers to long term preservation we are describing. Shortly before the inauguration of President Bush, the Clinton/Bush transition team sent a memo to all federal agencies instructing them to remove from their web sites any information that related to Clinton administration policies. The removal and alteration of documents that were "published" on the Arctic National Wildlife Refuge web site presents a troubling development when government policy casts the web site as the location of the only official copy of those documents. The first danger exemplified here is that, without notice and for overtly political policy reasons, the removals and alterations changed the historical record and removed from access information that was considered "published." Although an agency may claim the right to make such changes, we believe that it cannot claim that the web site provides "permanent access" and simultaneously treat the web site as a constantly changing bulletin board subject to policy changes. The second problem is that, although the agency admits to removing "anything that would be counter to new policy," it claims to have left scientific information untouched. But, even the agency admits that, in deciding what to keep and what to remove or change, "[t]he line between opinion and science became fuzzy." Indeed, the line between what is science and what is policy and what is policy-approved science can be very thin. When data are treated the way these documents have been treated, what guarantee will we have of future access to reliable, accurate data?[43]

Long-term access is also at risk for other reasons. In tight budget times, how likely are agencies to assign a high priority to permanent public access to important data that are used by only a few scholars? Should we leave it up to next year's federal budget whether or not we will have access to data collected ten years ago? Can we assume that the successive administrations will continue to fund access to data that place certain policies in an embarrassing or unflattering light?

An alarming aspect of this shift in who controls the long-term life of government data is that we are talking about huge amounts of valuable information being placed in direct government control. If systems fail, or funding isn't made available for "end of life cycle" needs such as migration and refreshing, or an agency is dismantled and data is not given over to another institution, huge amounts of public information will be lost.

Politically, data users will have to insist that government follow meaningful preservation practices. New data products must take into account long-term access. The best practice we currently have for ensuring data preservation is through wide distribution of raw data to data archives and digital

libraries. Online access to government data on government servers is not sufficient. Data archives have more than two decades of experience in keeping data accessible. The tools for conversion, migration, and preservation will be as common among the emerging digital libraries as online public access catalogs are to book-based libraries. While we celebrate "how much is instantly available," we will also have to carefully examine and question long term access policies and criticize shortsighted politicians who think only of the next election, and whose funding concerns don't reach years ahead to the needs of future generations of researchers.

Blur

> In the future world of government information, the links between content and services will become more tightly coupled and more complex.... Will the government eventually develop services that filter and customize information for individuals...? Technical solutions and policy issues are closely intertwined in the Internet environment.[44]

We are already seeing a lot of what Lippincott and Cheverie describe above. Public data is available in all shapes and forms from sites such as "FedStats" (www.fedstats.gov), CDC Wonder (wonder.cdc.gov), Ferret (ferret.bls.census.gov), and the "Statistical Briefing Rooms" (www.whitehouse.gov/news/fsbr.html). The dynamic nature of web sites such as these also allow users to get just the information they need quickly, and without having to acquire all the data or a complete publication.

Where once we would have seen statistical publications and data products, today we see statistical services and data services. As we saw above, administrative data and survey data may merge or be linked to provide new data services. This blur between products and services creates new access problems even as it improves access. "Blur" will be the context for many of the problems we have described above.

Confidentiality and privacy issues, for instance, will loom over government interactive web sites. Users will be concerned over the confidentiality of the information they send to the government as they, say, file taxes or register their car. Will their data be intercepted on the way to the government? And once confidential data are collected through public web servers, keeping the information confidential will be more difficult as we have seen by the example of commercial sites whose files of credit card information have been stolen.[45] Once collected, will the government use the information only for the purpose gathered? And, as you use data, will government data-servers track what data you use and draw conclusions about you? Will the public react to invasions of privacy (whether real or imagined) by being less cooperative with agencies collecting administrative and survey data?

When the government provides new access to data by providing services, we have seen that the question of privatization and the roles of government and the private sector quickly arise. As "distribution of data" becomes "access to administrative services," confidentiality may preclude release of data that would once have been available. As an agency concentrates on its administrative role and uses database technology to drive its service, the database will inevitably contain much more information than can be released to the public. The static data file that, in the recent past, would have been constructed and distributed will, in this near future we are describing, be replaced by a dynamic database. As public "access" to the data (or some of the data) is provided by the government service, why should the agency go to the trouble of creating and distributing a static data file? Think of a relational database into which companies report chemical spills and employee injuries. This database might have confidential information it and cannot be made public in its everyday dynamic form. At what point will the government extract the data that is clearly in the public sphere and make it available?

As services replace "products," who will save the data that was once packaged as a product?

Politics is Politics

The idea that quality of data might depend upon our political advocacy is no doubt distasteful to many readers. Paying our respects to professional ethics is one thing, but wading into the messy world of politics is quite another! After all, we are concerned with facts, data, and the search for the truth. This is a far cry from trying to ferret out political motivations, hidden agendas, special interest influences, and so on.

It seems that even our professional ethics will need to be reevaluated as we evolve into an online world. Once ethics might have involved pretty simple notions such as not breaking the confidentiality of records. As data distribution, data access, and data products become increasingly complex to the point where methods of collection, presentation, and preservation are beyond the grasp of people outside the field – and data is still a "social good" that is the

basis for policy decisions affecting our society–will our professional responsibilities include being a bridge between this complex world and our fellow citizens?

Notes

[1] *The digital dilemma: intellectual property in the information age*. National Research Council. Washington, D.C.: National Academy Press, c2000.

[2] Talbott, Stephen L. "The Machine's Hidden Agenda." www.oreilly.com/~stevet/meditations/agenda.html

[3] California Legislative Analyst's Office. *"E-Government" in California: Providing Services to Citizens Through the Internet*. (January 24, 2001) http://www.lao.ca.gov/2001/012401_egovernment.pdf

[4] *Falling Through the Net: Toward Digital Inclusion A Report on Americans' Access to Technology Tools*. October 2000. National Telecommunications and Information Administration. Page xvi. http://www.ntia.doc.gov/ntiahome/digitaldivide/index.html

[5] General Accounting Office. *Characteristics and Choices of Internet Users*. GAO-01-345 (February 2001)

[6] "Browser Note" on American FactFinder http://factfinder.census.gov/. Accessed April 17,2001.

[7] See, for instance, Samuelson, Robert J. "Out of Print." *Newsweek* 126, no.11 (11 Sept. 1995): 59. In this article, Samuelson quotes Martha Farnsworth Riche, director of the Census Bureau, "'If someone else can do it, let's shift it to the outside,' she says. 'We've had a hiring freeze since at least 1992, and those [printed] reports take an enormous amount of time from professionals.' They need to concentrate on doing surveys of 'an economy and population that are changing dramatically. Our statistics have fallen behind' Only Census can collect much of this data, she says. Let academics and analysts prepare reports."

[8] "The information office of the future." *Government Information Insider* VIII: 4 (Fall 1999): p.41.

[9] See, for instance, the protection of confidentiality of respondents as embodied in the Privacy Act of 1974 (5 U.S.C. § 552a) and Title 13 ("Census") of the *United States Code*.

[10] In its sixth annual *Computer Crime and Security Survey*, Computer Security Institute and the San Francisco Federal Bureau of Investigation's Computer Intrusion Squad found that "Eighty-five percent of respondents (primarily large corporations and government agencies) detected computer security breaches within the last twelve months." CSI. San Francisco, March 2001. http://www.gocsi.com/prelea_000321.htm

[11] Petersen, Carolyn. "The Danger Within." *Managed Healthcare* October 1998, Vol. 8(10): 20-24.

[12] Rubin, Alissa J. "Lobbyists Go Full Tilt in Bid to Ease Patient Privacy Rules." *Los Angeles Times*, (March 24, 2001) p.1.

[13] Most web server software by default collects into log files a variety of information about every click on every page of a web site. One common way for web sites to maintain "state" in an otherwise "stateless" communication protocol is to save "cookies" on the browser's local client machine. In addition to adding functionality to web sites, this practice allows web site designers to know quite a bit about a person and their browsing behavior.

[14] West, Darrell M. "Assessing E-Government: The Internet, Democracy, and Service Delivery by State and Federal Governments." September, 2000. www.insidepolitics.org/egovtreport00.html

[15] General Accounting Office. "Federal Agencies' Fair Information Practices." September 11, 2000. GAO/AIMD-00-296R

[16] Sinrod, Eric J. "By Any Other Name." *Computerworld*., (March 19, 2001)

[17] General Accounting Office. *Record Linkage and Privacy : Issue in creating new federal research and statistical information*. GAO Report: GAO-01-126SP (April 2001).

[18] For more on e-government, see, for instance, the National Science Foundation's support of digital government initiatives (www.nsf.gov/od/lpa/news/press/00/pr0031.htm and www.diggov.org/)

[19] See, for instance, Doeppers, Carole M. "Information Gathering by the Public and Private Sectors." May 2000. www.aclu-wi.org/issues/data-privacy/demand-for-data.pdf

[20] Cohn, D'Vera. "CBO Request For Census, Tax Records Raises Fears; Privacy Concerns Cited; Change in Law Is Sought." *The Washington Post* (October 13, 2000): p. A37. Although, According to the GAO report sited above, the CBO sought the linked dataset, stripped of personal identifiers, the incident and responses to it are indicators of the highly charged nature of data linking.

[21] Gellman, Robert. "Collecting data? Beware who else wants it." *Government Computer News*, v19 n22 (August 7, 2000):31.

[22] GAO. *Record Linkage and Privac*y p.11

[23] Simpson. Glenn R. "FBI turns to private sector for data: ChoicePoint turns a profit by selling personal information" *Wall Street Journal* (April 13, 2001).

[24] Schiller, Herbert I. "Public information goes corporate." *Library Journal*, (October 1991): 42-45.

[25] See, for instance, Dowling , S.A. "Information Access: Public Goods or Private Goods" *Social Science Computer Review* 12 (3): 333-350 (Fall 1994); McMullen, Susan. "US government information: selected current issues in public access vs. private competition" *Journal of Government Information* 27 (2000) 581-593.

[26] Stiglitz, Joseph E., Orszag, Peter R., and Orszag, Jonathan M. *The Role of Government in a Digital Age*. Commissioned by the Computer & Communications Industry Association. October 2000.

[27] Childs, Kelvin. "The price of public information." *Editor & Publisher* v130, n42 (October 18, 1977):13.

[28] *Government for Sale*. National Newspaper Association (1997).

[29] Moulton, Brent R. "Bias in the Consumer Price Index: What Is the Evidence?" Bureau of Labor Statistics, Office of Prices and Living Conditions, Working Paper 294 (October 1996)

[30] Prophet, Katherine. "Threats to Public Access to Federal Government Publications in Canada and the United States" *Government Information in Canada/Information gouvernementale au Canada*, Number/Numéro 18 (August 1999).

[31] *Survey of Current Business* (November/ December 1995) p. C1.

[32] http://www.stat-usa.gov/. Depository libraries are given confidential passwords so they can allow access to Stat-USA by two simultaneous in-house users.

[33] *Human Rights Watch*. "Electrifying Speech." July 1992. http://www.cpsr.org/cpsr/free_speech/ electrifying_speech_human_rights_watch.txt.

[34] Schrage, Michael. "U.S. Limits Access to Information Related to National Security." *The Washington Post*, November 13, 1986, p.A1.

[35] Willman, David. "U.S. Agrees to Disclose Chrysler Minivan Crash Test Data." *Los Angeles Times*, August 29, 1995. Part A: Page 12.

[36] Mintz, John. "FAA to release data on safety of airlines." *The Washington Post*, 30 January 1997, p.D1.

[37] Public Law 106-40.

[38] *Federal Register*, January 17, 2001 (Volume 66, Number 11) Page 4021-4024

[39] Hess, Glenn. "Concerns Over Right-to-Know Data Prompts Further Debate in Congress." *Chemical Market Reporter* 255:25 (1999) p. 1. *See also*: Gellman, Robert. "Don't Succumb to a Net-Driven Secrecy Panic." *Government Computer News* 18:9 (1999).

[40] Convey, Mary Christine. "Controversy Heats Up over Malpractice Database Access." *National Underwriter* 104:44 (2000) p. 9+.

[41] *Of Significance...* Vol 2, No. 2, (2000) "Preservation of Public Data".

[42] Weinstein, Henry. "Tobacco memo pulled from House Web Site." *The Washington Post*, 27 April, 1998. page A4.

[43] Benner, Jeffrey. "Oil and Websites Don't Mix," *Wired News* (March 23, 2001) http://www.wired.com/ news/politics/0,1283,42536,00.html.

[44] Lippincott, Joan K. and Joan F. Cheverie. "The 'Blur' of federal information and services: Implications for university libraries." *Journal of Government Information* 26:1 (1999) p.25-31.

[45] Harrison, Ann. "Credit-card numbers stolen via known security hole." *Computerworld*, March 10, 2000.