GOVERNMENTS AND THE DIGITAL RECORD: THE HISTORIAN'S PERSPECTIVE

Report on a Panel Discussion on Government Information and Societal Memory
Convened by the American Historical Association, January, 2014

PREPARED FOR THE CENTER FOR RESEARCH LIBRARIES
GLOBAL RESOURCES COLLECTIONS FORUM

BY BERNARD F. REILLY, JR.

March 27, 2014

CENTER FOR RESEARCH LIBRARIES
GLOBAL RESOURCES COLLECTIONS FORUM
April 24-25, 2014

Contents

Executive Summary

The growing embrace of digital technologies for records and communications by government agencies U.S. and worldwide is producing a wealth of source materials for historians. At the same time this migration to the digital environment has the potential to make the historian's task more difficult in the future. A panel of historians convened recently by the Research Division of the American Historical Association identified several factors that pose a threat to the long-term survival and integrity of the electronic evidence of the workings and activities of governments. The factors include the sheer volume and variety of digital documents, new obstacles to the release of records, the decentralization of government records management and publishing, and the growing role of third parties in the government information "supply chain."

The panel discussion provided a sense of working historians' concerns about long-term integrity and accessibility of electronic government documents and data. An understanding of those concerns was to form the basis for further exploration by the Center for Research Libraries of what role CRL in particular, and research libraries in general, should play in addressing issues of archiving and integrity in digital evidence.

The discussions suggested measures that research libraries, working in tandem with the historical community, can take to help ensure the integrity of the digital public record and its usefulness to historians. These include: promoting greater understanding of the digital information life cycle; fostering awareness of emerging scholarly uses of electronic government information; and advocacy for greater centralized coordination of government information management technology.

I. Background and Purpose of the Meeting

On January 4, 2014 the American Historical Association's Research Division convened a panel to provide CRL the perspective of working historians on an important issue: the survival and integrity of electronic government records and born-digital government documents and data. This perspective was to inform discussions at a two-day research libraries forum to be held by CRL in April 2014. Attendees at the April forum will ponder the role of research libraries in preserving government information in an environment where official records and communications are predominantly electronic.

This exploration was an outcome of the 2013 conference "The Global Dimensions of Scholarship and Research Libraries: A Forum on the Future" (http://www.crl.edu/focus/winter-2014). That conference examined how well and to what extent U.S. research libraries are supporting scholarly access to source materials for international studies. One of the findings of the Global Dimensions Forum was that the radical changes in the information supply chain brought about by the Web and digital media require a fundamental rethinking of library strategies to acquire and preserve primary sources. Those

changes are especially apparent, and particularly acute, in the realm of government records. In the past, libraries collected and preserved government publications in paper and in microform, and on tangible media like CD-ROM. Many served the function of repositories outside the sphere of governmental control. A network of designated federal depository libraries formed the core of this activity In the U.S. Now most government information is disseminated via the Web, and most government agency records are maintained in electronic form – often in "the cloud." Therefore new ways must be found to ensure that the internal communications and public documentation of the workings of governments, and the important data gathered by governments, continue to be available to researchers in the future.

The attending historians were encouraged to talk about the impact the government's shift to electronic records and digital publishing has had, or is likely to have, on their work and on the work of others in modern historical studies. As background for that discussion CRL provided notes on a few significant "events" illustrating realities of the new supply chain and life cycle of government information. Those notes are appended to this document.

Present at the meeting were the following:

- Randall Packard, William H. Welch Professor of the History of Medicine, Johns Hopkins University
- Kristin Mann, Professor of History, Emory University
- Ann McGrath, Professor of History and Director of the Australian Center for Indigenous History, Australian National University
- Matthew Connelly, Professor of History, Columbia University
- Joyce E. Chaplin, James Duncan Phillips Professor of Early American History, Harvard University
- Debbie Ann Doyle, Coordinator, Committees and Meetings, American Historical Association
- Bernard Reilly, President, Center for Research Libraries

The author also held a follow-up discussion with Antoinette Burton, Professor of History, University of Illinois, by telephone.

II. Focus of the Discussion

Because of the makeup of the group, much of the discussion focused on the records and documents of the U.S. federal government, as opposed to foreign or state and local government information. The discussion covered two very different types of historical evidence:

1) Data and documents officially published or exposed to open, electronic networks by government agencies in their day to day work, such as executive orders, published reports and statistics, statutes, and reports and proceedings of legislative bodies and committees

2) Internal documents and communications of government agencies, such as emails, cables, and memoranda, and other files designated as permanent records .

More particularly, the discussion was limited specifically to "born-digital" materials, rather than paper records and documents that have been digitized.

Records of the U.S. Department of State were a particular focus of the discussion. Those records represent a special case, for a number of reasons. The Central Foreign Policy files and diplomatic correspondence are particularly rich sources for historians. The Department of State, moreover, was an early adopter of electronic records and communications systems, and has evolved through multiple generations of electronic records. And because there is a clear and longstanding statutory requirement that a comprehensive record of American diplomacy be maintained, oversight of the Department's archiving practices has been a matter of public record. Finally, the records management protocols of the State Department are exceptionally sophisticated, as they affect the confidentiality of sensitive diplomatic communications and matters of national security.

Despite this focus, the discussions yielded findings and perspectives that for the most part apply to government-produced records and data in general, both domestic and foreign. The same challenges are being encountered, or are likely to be encountered, by historians seeking to gain access to materials from governments abroad, albeit on a different scale. The present report is not a narrative account of the discussion itself, but rather the author's summary of the concerns raised or suggested in the discussion, and of the points of consensus on how libraries might help address those concerns.

III. Issues Identified

Even prior the digital era, historians encountered obstacles in their efforts to access the records and internal communications of governments. Official reluctance to declassify records and the underfunding of the agencies tasked with managing retired agency files were chronic realities even when documents were all paper-based. (Attendees observed, in fact, that a relatively small percentage of U.S. government paper records are available either in microform or digital formats.) The digital era, however, has introduced new factors that hamper the efforts of governments to effectively preserve their records. Government agencies' widespread adoption of digital media and networks for reporting, communication, documentation, and other core functions, and their embrace of direct web publishing, while generating an unprecedented wealth of documentation, pose new threats to the permanence and integrity of that documentation. Given the ways in which electronic records, data and documents are created and managed by some government agencies, many attendees wondered whether those records will be usable for scholarly purposes in the future.

For historians five factors are matters of particular concern:

1. The sheer number of digital documents being created
2. The extreme variety and technical complexity of those documents
3. New obstacles to the release of records
4. Decentralization of government publishing and records management
5. The growing role of third parties in the government information "supply chain."

Apart from these factors, two developments that coincided with the advent of electronic records were viewed as having a bearing on the survival and accessibility of government records and data. First, the reduction in the size of the public sector since the 1980s has reduced the capabilities of government agencies tasked with caring for official archives. This trend accelerated with the austerity measures implemented following the worldwide financial crisis of 2007-2008. Second, the recent tendency to outsource government functions, including records management and publication, to the private sector, has relegated the provision of key services to providers less accountable to the public and less friendly to the historical community.

1. The Sheer Number of Digital Documents Being Created

The web and digital media have enabled government agencies to become publishers, since official documents and communications can be reproduced and distributed virtually without limit or cost. Combined with pressures for government transparency, this has led to an unprecedented abundance of digital records and a deluge of government-produced data. As a consequence, the retirement and declassification of records by many government agencies cannot keep pace with the flood of materials, and the GPO's attempts to gather and preserve non-current agency documents and publications fall well short of being comprehensive.

2. The Extreme Variety and Technical Complexity of Electronic Records and Documents

Digital technology has fundamentally altered the nature and life cycle of government records and documents. Materials in digital formats, like text documents, email, still images, video, and databases, have diverse and highly complicated life cycles. Government agencies employ an array of proprietary and open source digital tools to create, manage and maintain their records and communications today. Understanding how documents are produced and managed at any given point in time is crucial to the ability of historians to authenticate and interpret those materials at a later date, and to researchers' ability to trace and document provenance and chain of custody. Historians will want to know, for example, how the State Department cable system worked during the Benghazi crisis, or what was involved in the deletion of email messages from White House servers during the George W. Bush presidency (and how some of those deleted messages were recovered).

One attendee pointed out, the large and growing disparity between the sophistication of the technologies used by government agencies for producing and distributing digital content, and the limited technological resources and capabilities of the archives and libraries expected to preserve them. One attendee observed that this

disparity is already apparent in many libraries' inability to manage digital multimedia content, like video, audio, and geospatial data.

3. New Obstacles to the Release of Records

Attendees believed that the age-old conflict between the interests of historians and those of government agencies has become only more problematic considering the ease with which electronic records can be altered, encrypted, or destroyed. Expanding oversight and regulatory apparatus at the national, state and local levels, moreover, is impeding the declassification, disclosure and exposure of government data and records to an unprecedented degree. This opacity is driven by concerns for individual privacy, diplomatic relations, and national security. It is manifest in a growing array of legislation and regulation in the U.S., from HIPAA to the USA Patriot Act. Attendees believed that these restrictions are likely to impact some fields in history more than others, such as U.S. diplomatic history, and the history of medicine and public health. Indeed, the case has been made by the U.S. government that the Internet and the availability of capabilities for mining electronic records have increased the potential for harm through inappropriate disclosure.

4. Decentralization of Government Publishing and Records Management

In the analog era both the GPO and National Archives and Records Administration (NARA) exerted some influence, and even control, over U.S. federal government agencies' publishing and records management activities. (In the past, up to 50% of U.S. government documents passed through the GPO.) Today, it is not clear whether either NARA or GPO has the authority, or even a clear mandate, to impose on agencies a degree of standardization conducive to adequate archiving. Government agencies today routinely self-publish, and much of the responsibility for decision-making on the systems used by agencies to manage their internal records is vested with the agencies themselves. The guidance NARA provides to agencies on the use of email, social media platforms, and cloud services, is largely advisory.

The resulting lack of uniformity in document creation and record management practice among agencies is likely to complicate or even thwart the task of archiving those materials. The plethora of tools and platforms used to produce, organize and store documents(proprietary email and record management software, social media applications, etc.), fosters a problematic lack of uniformity that is likely to drive up the cost of preservation and archiving. On this issue the panel echoed the observations of a July 2013 Congressional Research Service report, *Retaining and Preserving Federal Records in a Digital Environment: Background and Issues for Congress:*

> The variety of electronic platforms used to create federal records, however, may complicate the technologies needed to capture and retain them. It is also unclear whether the devices and applications that agencies currently use to create and retain records will be viable in perpetuity—making access to federal records over time increasingly complicated, costly, and potentially impossible.

What uniformity of practice does exist today among U.S. agencies seems to be largely the result of national security concerns. The Department of Defense has created widely accepted standards for the certification of database platforms. And the science and social science academic communities help shape the norms for government handling of information, such as ethical protocols on privacy and human subject research. But matters important to historians, such as provenance, authorship, version control, and the integrity of content over time, do not figure prominently among the priorities of many government agencies. It was also perceived that the "curatorial" expertise necessary to represent historians' concerns and properly inform agency and NARA decisions on these matters is not in place.

5. The Growing Role of Third Parties in the Government Information "Supply Chain"

Because, government records and historical data are becoming more difficult to access and use, third-party organizations increasingly intervene today to provide alternative channels for public access. The role of commercial publishers like ProQuest, LexisNexis, and Ancestry.com in making government information available is growing. And new actors have lately emerged. Since 2008 WikiLeaks has released to the web millions of pages of classified, internal records of the Departments of State and Defense; and more recently Edward Snowden exposed the clandestine data-gathering practices of the National Security Agency by disclosing, with the assistance of the publishers of *The Washington Post* and *The Guardian*, thousands of pages of agency documents. And over the past several years the Internet Archive has captured millions of public documents that have since been removed from the web by their agencies.

What is not clear, however, is how well these third party efforts will ultimately serve the interests of historians. The published partnership agreements between NARA and its digitization service providers indicate that those arrangements impose limits, albeit temporary, on NARA's use of the digitized content, and that the rich proprietary metadata the vendors produce are not made available to NARA by the vendors. Entities that maintain large online repositories, like WikiLeaks and the Internet Archive, have no obligation, either to the federal government or to the historical community, to maintain those resources for the long term.

IV. Conclusions: An Agenda for Research Libraries

The purpose of the panel discussion was to provide CRL a sense of working historians' concerns about long-term integrity and accessibility of electronic government documents and data. An understanding of those concerns is to form the basis for further exploration, at the April 24-25 2014 CRL forum *Leviathan: Research Libraries and Government Information in the Age of Big Data*, of what CRL in particular and research libraries in general, should play in addressing the issues of archiving and integrity of digital evidence.

The discussions suggested that certain new measures could be taken by libraries that would help ensure the integrity of the digital public record and its usefulness to historians. Attendees agreed, however, that measures requiring significant new federal government funding are unlikely to be embraced in the present economic and political environment. Attendees also favored solutions that would draw upon and benefit multiple constituencies, such as historians and social scientists, who have a shared stake in the long-term integrity of evidence.

Specifically, the discussions suggested three areas in which libraries and the historical community might work together to achieve that end. While the discussion centered on the records of the U.S. federal government, the prescribed measures could apply to library strategies for dealing with foreign government archives and information.

1. Promote Greater Understanding of the Digital Information Life Cycle

Government agency adoption of digital technologies and communications has fundamentally altered the nature and life cycle of government information. This change requires of historians a new understanding of how official records and documents are created, managed and distributed. This new understanding will be essential to the ability of scholars to authenticate and interpret those documents and to their ability to trace and document provenance and chain of custody in the future. Attendees believed that training historians in the electronic documents life cycle should be a larger component of digital humanities programs. Existing digital humanities programs focus largely on tools and techniques for presenting research findings and sources, and on the visualization of information and knowledge. More attention must be paid to "improving the underlying data" and authenticating digital content. Training historians on the implications of encryption technologies, cloud services, and the like would be useful. It was suggested that the AHA might provide fellowships or even maintain resident expertise on these subjects.
Librarians could contribute by monitoring and mapping the new life cycle of government information and by evaluating the key systems, tools and platforms used by government agencies to produce and manage records. It was observed that information about these systems and platforms can be gleaned from published sources like the Federal Register. Mapping the life cycle of State Department records would provide a good test case for this activity..

Libraries might also document and assess the performance of third parties that preserve and distribute government records and content. Scrutinizing the terms of the arrangements made by governments with service providers like

ProQuest and Ancestry.com, for example, could provide information useful for judging provenance, authenticity, and chain of custody of electronic government documents distributed through those and other third-party platforms. Some research libraries also possess the requisite curatorial and technical expertise to assess the organizational imperatives and capabilities of entities like WikiLeaks and the Internet Archive, which will have a bearing on those entities' responsiveness to the needs of historians over time. The same expertise could also be brought to bear on evaluating agencies' content management platforms and practices, identifying where the risk of loss or corruption of evidence exists. Moreover, since libraries constitute the major "market" for commercial service providers like ProQuest, their leverage might be used to improve the outcomes of these "outsourcing" arrangements.

Finally, libraries could also help develop and promote new technologies to accelerate the release of government records. New technologies and approaches need to be deployed to expedite and automate the processing and declassification of agency records, and to make those materials available on a timely basis. Genealogical organizations have had some success using crowdsourcing for annotating and indexing digital historical content, and citizens' initiatives have experimented with the use of natural language processing software to automatically create finding aids.

2. Promote Awareness of Emerging Scholarly Uses of Electronic Government Information

The vast oceans of information being produced and distributed by government agencies demand that new tools and approaches be employed by the scholars who mine them. Scholars now employ a host of new technologies to cope with the ever larger bodies of text and data. Many historians are working directly with computer scientists to mine centuries' worth of government-produced census and statistical data, and to identify and track trends in large historical record groups. As the tide of digital information rises further, text and data-mining and other computer-assisted research may become as commonplace among historians as they are in economics, sociology and finance. Librarians have a vested interest in being aware of these new tools and methodologies, to serve historians adequately, and are in a position to foster cross-fertilization among the disciplines, and between the disciplines and the technologists.

3. Advocate for Greater Centralized Coordination of Government Information Management Technology

Research libraries and historians have a common interest in NARA and the GPO exercising greater influence and authority over the management of records and publications by government agencies. The degree of discretion now held by the agencies on key document and records management decisions has resulted (, and is likely to continue to result) in the loss of important historical evidence. Attendees believed that a strong case could be made for giving greater oversight authority to NARA and GPO on the basis of security, efficiency and effectiveness of the agencies, as well as the economic benefits of government transparency and well-structured government data. Research libraries and the AHA might work together with the American Bar Association, which also has a stake in the integrity of electronic records, to build support for that authority in the Congress and beyond.

This is an ambitious agenda, and one that librarians and historians will not be able to achieve alone. Social science and public policy researchers share a common interest with historians in providing for the effective management of evidence. The AHA and CRL might enlist support for these measures from major foundations like the John D. and Catherine T. MacArthur Foundation, the Carnegie Corporation of New York, and others that already recognize the societal value of government transparency and the integrity of the public record. More broadly, the news media, the courts, and the journalism profession also have a stake in the future integrity of evidence.

Indeed the survival and authenticity of the documentary evidence of government's actions and workings is important to civil societies worldwide. The historians' panel discussion in January 2014 helped begin to give shape to the new role research libraries must play in this endeavor.

Appendix: Key Events Illustrating the New Realities of the Government Information Supply Chain

U.S. Government Publications and Data

1. **Open Government Data:** On May 9, 2013 the Obama White House issued an Executive Order mandating the open Web publication of data gathered and produced by U.S. federal agencies. From the press release:

> The Obama Administration today took groundbreaking new steps to make information generated and stored by the Federal Government more open and accessible to innovators and the public, to fuel entrepreneurship and economic growth while increasing government transparency and efficiency.
>
> Today's actions—including an Executive Order signed by the President and an Open Data Policy released by the Office of Management and Budget and the Office of Science and Technology Policy—declare that information is a valuable national asset whose value is multiplied when it is made easily accessible to the public. The Executive Order requires that, going forward, data generated by the government be made available in open, machine-readable formats, while appropriately safeguarding privacy, confidentiality, and security.

2. **The Federal Digital System:** In 2008 the Government Printing Office put into place the Federal Digital System or "FDsys," a digital repository to provide "free online access to official publications from all three branches of the Federal Government." The purpose of the repository was to "guarantee long-term preservation and access to digital Government content." As of December 2013, FDsys still held no publications or documents from the following federal agencies: Department of Homeland Security, Department of the Interior, Department of State, Department of Education, Department of Agriculture, or Department of Defense.

It is also not clear whether GPO has any statutory jurisdiction over digital government publications. From a March 2012 Congressional Research Service report:

> The emergence of a predominantly digital [Federal Depository Library Program] may call the capacity of the statutory authorities GPO exercises into question. Whereas GPO is the central point of distribution for tangible, printed FDLP materials, its responsibilities . . . may be less explicitly specified, regarding its distribution of digital information. . . . The agency has archiving and permanent retention authorities for tangible materials, but those authorities do not envision digital creation and distribution of government publications. Digital distribution authorities provide for online access to publications, but are silent on GPO's retention and preservation responsibilities for digital information.

3. **Foreign Broadcast Content:** Since the late 1990s the U.S. Open Source Center (OSC), a division of the Central Intelligence Agency, has continuously monitored broadcasts, web postings, newspapers, wire services, and other news sources from hundreds of countries outside the U.S. for national security purposes. English-language transcriptions of the content harvested by the OSC have been available to academic researchers in the *World News Connection, a* database distributed by a number of commercial publishers. (The database is the digital, successor product to the *Foreign Broadcast Information Service Daily Reports*, production of which was discontinued in 1996.)

In October 2013 the National Technical Information Service (NTIS), the government agency that brokers access to Open Source Center content for commercial publishers, informed distributors that, effective December 31, 2013, OSC would no longer make available newly harvested OSC content for distribution.

U.S. Government Agency Records

1. **Management of Agency Electronic Records:** according to testimony given to the Congress by officials of the Government Accountability Office in June 2010, federal records management "has received low priority within the federal government," and the creation of "[h]uge volumes of electronic information" posed a "major challenge" in agency record management. GAO officials noted that poor federal records management could leave the government "exposed to legal liabilities, and historical records of vital interest could be lost forever."

In May 2011, the National Archives and Records Administration (NARA) published a report on U.S. federal agencies' self-assessments of their recordkeeping, which found that "90% of agencies had a moderate to high risk of records mismanagement." The agency self-assessments found that 45% of agencies had records management programs with "moderate risk" and another 45% had records management programs with "high risk" of records mismanagement. According to NARA, in fiscal year 2011 "80% of [U.S. federal] agencies captured e-mail records by printing them out and filing them."

2. **The White House and Presidential Records:** The administration electronic records transferred by the George W. Bush White House to the National Archives in 2009 totaled over 72 terabytes of data in almost 400 million files. They included more than 200 million emails, more than 11 million digital photographs, 48,000 digital motion videos, more than 29 million records of entry by workers and visitors to the White House complex, records from 36 other computer systems or applications in the EOP, as well as vice presidential electronic records and classified electronic records. Prior to 2009, estimates indicate that between 5 million and 23 million emails of White House staff were either lost or destroyed by the White House.

3. **Social Media and Federal Agencies:** A June 2011 report by the Government Accountability Office found that 23 of 24 major federal agencies used FaceBook, Twitter, and YouTube as platforms for the public dissemination of agency information. A 2012 Congressional Research Service report called attention to the proliferation of the proprietary platforms employed by the agencies of the U.S. federal government for their records, and the accompanying dangers.

4. State "Erasure Laws": In an April 28, 2013 ***New York Times article***, editor Bill Keller wrote about the proliferating erasure laws adopted by state governments. "States feel greater pressure" Keller wrote, "to put public records offline." After a New York newspaper published names and addresses of local handgun permit-holders, the Legislature in Albany sharply limited access to that information." He also noted that Google's ***latest transparency report*** showed a sharp rise in requests from governments and courts to take down potentially damaging material.

Foreign Government Records and Information

1. **African Censuses:** Population censuses in Africa provide much of the information used to develop public policies and inform economic development. In the late 1960s, the United Nations Population Fund (UNFPA) set up the African Census Analysis Project (ACAP), which enabled some twenty African nations to take their first general census. Since then, computerized data storage technologies evolved at such a rapid pace that, with no measures having been taken to transfer the data to new storage media, much of the data collected in the 1970s and 1980s censuses were completely lost.

Only rare, perfunctory publications exist of the African censuses from this period, making it impossible to carry out any new evaluation of the information collected in the past.

2. **Proprietary Systems for Government Geospatial and Statistical Data:** In a growing number of instances, government agencies in Latin America, Asia and other regions rely upon proprietary applications and commercial platforms to manage and provide access to their data and publications. For example, the popular ArcGIS platform, a geographic information system application produced by the Redmond, California-based software company Esri, is used by government environmental agencies, geological surveys, and utilities throughout Latin America as the core repository architecture supporting public and back-end interfaces for the geospatial information those agencies create and collect. And the Bangladesh Bureau of Statistics posts its data openly on a highly unstable site hosted by Synesis IT, a commercial IT services provider based in Dhaka.