LEVIATHAN

Libraries and Government Information
in the Age of Big Data

In This Issue

Research libraries have long collected and preserved the documents and publications of domestic and foreign governments, and have acquired in microform the archives of many of those regimes. In this way, they ensure that information and documentary evidence vital to scholarship remains available for the long term.

The digital environment brings new complexities and uncertainties to this task. The immense oceans of information that government agencies produce and collect today, and the inherent fluidity and impermanence of digital media, require a fundamental rethinking of longstanding preservation practices.

At the same time, those very technologies are creating tremendous new possibilities for information access and management. Governments worldwide are committing to "open data," and a growing legion of activists and entrepreneurs are now embracing government transparency as a cause, and even a business opportunity.

Clearly, for CRL and the North American research libraries concerned about the continued integrity and accessibility of information, the times call for a new playbook. The 2014 Global Resources Forum *Leviathan: Libraries and Government Information in the Age of Big Data,* convened by CRL on April 24–25, explored these new realities and examined the complexities and opportunities they present. Forum speakers identified ways in which "e-government" imperils the long-term integrity and accessibility of information. The discussions yielded five broad strategies that research libraries, individually and collectively, can adopt to address those threats.

The *Leviathan* presentations and discussions are now archived online at: www. crl.edu/leviathan. This issue of *FOCUS* offers a synopsis of the challenges and strategies identified. The *Leviathan* conversations will form the foundation for CRL planning and priorities in the field of government information for the coming years. While the activities called for may not resemble those one traditionally associates with CRL, we believe that they are suited to the challenges we face.

—*Bernard F. Reilly, Jr.*
President

www.crl.edu

Will the Challenge of "Big Data" Mean the End of History as We Know It?

A Report on the 2014 Global Resources Collections Forum

Session I: Libraries and the Records of Governments
**Thursday, April 24, 1:00–5:00 P.M.**

In the past, CRL and member libraries acquired and made available to scholars the records of key U.S., Canadian, and British government agencies. Those records included census and immigration records, State Department communications, British Foreign Office files, and the papers of various presidents and ministers. CRL is a repository of the transcripts of the 1960s–70s secret military tribunals in Brazil and the incriminating files of the Khmer Rouge's Santebal police in Cambodia. The Cooperative Africana Microform Project, operating under the CRL umbrella, has preserved in microform the records of the French colonial regime in Senegal.

Today most government records are created and managed in a multitude of digital systems and platforms, many of them in "the cloud." With the availability of digital media, moreover, the volume of the records produced has mushroomed. (The Bush White House alone generated over 200 million emails and 11 million digital photographs.)

Both the Government Accounting Office and the Congressional Research Service have raised concerns about the potential loss of U.S. government electronic records and data. This *Leviathan* session explored the policy and political challenges the digital government ecosystem presents for libraries and the implications of this new reality for the stewardship of important historical evidence.

Thomas S. Blanton, Executive Director of the National Security Archive at George Washington University, offered a keynote on "Information, Transparency and Government Records in the Digital Age: A Public Policy Perspective." Blanton spoke about the issues surrounding access to government records today, from the encryption of government agency emails to delays in the declassification of politically sensitive records. The National Security Archive efforts to secure declassification of U.S. State Department and CIA records through FOIA requests and legal action provided a number of case studies in archival disclosure. Blanton also noted the striking disproportionality between the resources devoted to government data, and those allocated to its preservation, as illustrated by the Utah data storage facility planned by the National Security Agency and the facilities of NARA.

Columbia University Historian Matthew J. Connelly provided a historian's perspective on the barriers and obstacles scholars face today in obtaining access to electronic U.S. government records and internal communications. Speaking as one doing
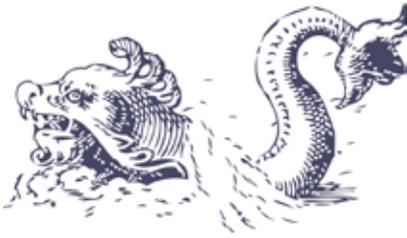
Why "Leviathan"?

The Forum title refers to the English political philosopher Thomas Hobbes's 1651 treatise, *Leviathan or The Matter, Forme and Power of a Common Wealth Ecclesiasticall and Civil.* The treatise argued the need for "strong, undivided government" as a check on the disorder and brutality inherent in human society. Hobbes used in his title the biblical term for sea monster, referenced in the Old Testament and pictured on many maps from Hobbes's era to signify uncharted territory.

research on U.S diplomatic history, Connelly described the large and growing backlog in the declassification and processing of State Department records, and the "asymmetry" between the volume of records created by well-resourced federal agencies like the Defense, CIA, and State Departments, and the sparse funding and human resources available to NARA. On the other hand, Connelly noted, computer technologies offer new possibilities for automating and thus accelerating declassification and processing of such records. He cited recent work to develop tools to surmount those challenges.

The issues and challenges of preserving electronic government records constituted the topic of presentations and comments by a panel of representatives from the national archiving agencies: **Paul Wester, Jr., Chief Records Officer for the United States Government,** and **William A. Mayer, Executive for Research Services,** both of the **National Archives and Records Administration,** and **Cecilia Muir, Chief Operating Officer,** and **Paul Wagner, Director General and Chief Information Officer,** of **Library and Archives Canada.** The discussion covered the structural, political, and technical challenges both organizations are encountering in their efforts to fulfill their mandates to preserve the permanent records of government agencies. Challenges include inadequate funding in an era of government cutbacks; limited influence of the archives over agency adoption and deployment of records creation and management technologies; and the overwhelming volume, variety, and complexity of records being produced by agencies.

Session 2: Libraries and the Information of Governments
Friday, April 25, 9:00 a.m–12:00 noon

In the print era, many libraries served (and today continue to serve) as depositories of tangible government publications, bearing responsibility for providing local access to government documents and information in print and microform. Most academic libraries also acquired or subscribed to commercially produced microform sets and databases of government documents and information.

Today, the public information landscape is being transformed: by government agencies' widespread adoption of social media, their embrace of "cloud" services for storage and management of records and information, and Open Government mandates that make oceans of agency data freely available on the web. This *Leviathan* session examined the implications for libraries of this new government information "supply chain."

John C. Bracken, Director of Journalism and Media Innovation, the John and James L. Knight Foundation, gave a keynote on Open Government and Big Data, describing new creative and productive uses of data released by local and federal governments. Bracken showed how many of these uses, funded by Knight Foundation, are driving technological innovation and producing significant public benefits. Such initiatives are made possible by new government economic and transparency efforts at the federal, state, and local level that expose open, structured data, long collected by governments but until the advent of the Internet never widely disseminated.

James A. Jacobs, Technical Advisor for CRL's Certification Advisory Panel, reprised the salient points in his study, commissioned by CRL, of the current state of archiving born-digital government content. It was clear from Jacobs' report that so much information is published by U.S. federal government agencies that the scope of that information is difficult, if not impossible, to define. Jacobs pointed out that, in fact, there is not even consensus on the appropriate unit of measurement with

which to reckon the dimensions of web-based agency output, vastly complicating the task of archiving that content. Jacobs' study found that despite the efforts of the National Archives and Records Administration, GPO, the Library of Congress, Internet Archive, and others, only a small fraction of U.S. government agency content has been usefully archived to date.

R. Eric Petersen, Senior Analyst in American National Government, in the Congressional Research Service, spoke about the "The Digital Future of FDsys and the Federal Depository Library Program" from a public policy perspective. He noted that organizations in the commercial sector are investing large sums to maintain statistical data and "historical" content, and that such investments are on a scale that dwarfs the resources available to GPO and most government archiving bodies.

A panel discussion on "New Models of Access" then explored the role of third-party aggregators and publishers in organizing and enhancing government-produced records and information. **Susan Bokern, Vice President for Information Solutions, ProQuest,** described the work that ProQuest does to identify, gather, and authenticate the records of U.S. federal legislation, congressional hearings, and statistics, as well as the editorial value that ProQuest and other publishers add to the "raw" materials. **Robert Lee, Director of Online Publishing and Strategic Partnerships, East View Information Service,** spoke about the special difficulties of obtaining publication and data from oppressive regimes abroad, such as China and Russia. Lee elaborated on the political considerations that affect this part of the information supply chain. **Robert Dessau, CEO of voxgov,** described a new model for aggregating government information, one that involves real-time harvesting and annotating of public policy-related materials from media and government websites to produce value-added resources for policy research.

Session 3: New Models of Stewardship: An Agenda for CRL and North American Research Libraries

Friday, April 25, 1:00–3:00 P.M.

In *Leviathan* Session 3, three library leaders offered their perspectives on the appropriate role of research libraries in this radically transformed government information landscape. The presenters discussed what stewardship means in an age of digital government and big data. All agreed on the importance of defining this role in relation to the parts played by government itself, other civil society actors, and the commercial sector. The discussion suggested the outlines of a new strategic framework for cooperative action and new multi-year strategic priorities for CRL.

Mary Case, University Librarian, University of Illinois at Chicago, spoke about the tension between the interests of scholars—and thus the libraries that represent them—and the imperatives of privacy, national security and other matters that drive government agencies and the archiving branches of government. The scholarly desire for greater disclosure of government records and information can run counter to government's natural aversion to the risks that disclosure can create. Case also argued that digital preservation challenges require new models and approaches, different from the depository systems and "custodial model" of the print era.

Brent Roe, Executive Director, Canadian Association of Research Libraries, spoke about the successes and limits of collaborative efforts by Canadian libraries to date in this new environment, and the pressing need to deal with the legacy print corpus. He also argued the merits of a distributed approach to harvesting and archiving digital government information.

Responding to these comments, **Ingrid Parent, University Librarian, University of British Columbia and former Assistant Deputy Minister, Library and Archives Canada,** observed that access to persistent government information is a global issue, and that serving as "mechanisms of good governance" should be a priority of libraries everywhere. Parent encouraged the research library community to engage in what David Johnson called "the diplomacy of knowledge" and to work at the international level to advocate new partnerships and "drive new connections." Parent suggested that IFLA's new PERSIST initiative, undertaken in concert with UNESCO, would be an appropriate platform for library action on this front. ❖

A New Strategic Framework for North American Research Libraries

In a summary of the two days' discussions, CRL president Bernard Reilly listed five factors presenters and attendees identified that loom as particularly serious threats to the long-term integrity and accessibility of e-government information. These threats mirrored some of the concerns raised by the panel of historians convened by the American Historical Association Research Division in advance of the *Leviathan Forum*. (See the report "Governments and the Digital Record: the Historian's Perspective.")

1. **Scale / "The Leviathan":** In the age of Big Data, the sheer number of records, documents, and datasets being produced by governments is immense and is growing exponentially. Moreover, the variety and complexity of the digital information being produced—as the applications and platforms adopted by government agencies continue to multiply and evolve—is increasing as well. The result is large and growing backlogs of unprocessed records in NARA and sporadic inclusion of agency documents in FDsys.

2. **"Known Unknowns":** The universe of government information is growing and changing so rapidly that the current scope of that universe is essentially immeasurable. This, combined with the tendency of governments to withhold from public view records and communications they consider sensitive, enormously complicates the task of preservation.

3. **Asymmetry:** A large and growing disparity exists between the amount of information produced by governments, and the resources allocated by the public sector to preserve that information. Even at the national level in the U.S. and Canada, federal agencies historically tasked with archiving are ill-equipped to meet the new challenges. The problem is not simply one of mismatched resources, but of limits on the authority those agencies have over government publishing and records management.

4. **"The Cloud":** In the government information ecosystem, organizations that bring robust digital technologies and capabilities to the table now play a larger role than in the past. Such organizations include:

   - Cloud service providers and other suppliers of technology platforms used by government agencies to create and manage records and publish information

   - "Fourth Estate" organizations that aggregate, analyze, and interpret government records and data, and make them available to citizens and researchers

   - Private-sector, market-oriented aggregators and other third parties that

locate, organize, reformat, and distribute public sector data and documents to users in the academic, government and business worlds.

These organizations are an important part of the equation, as the channels through which much government information reaches users today. The content they serve is often inseparable from the tools they produce.

5. **"The Fog" (Non-Transparency):** Ironically, in an era of open government data, too little information is available about how government records and information are created and maintained. In particular, we need more data about:

- *Technology:* the myriad sophisticated systems and software that are now part of the critical infrastructure of governments. These include NARA's Electronic Records Archive, acknowledged to be problematic, as well as the kinds of email systems that resulted in the wholesale loss of White House and IRS messages. If such systems remain "black boxes," information vital to scholars about the production, provenance, and alteration of content will be lost or rendered unusable.

- *Finance:* The resources invested by the commercial aggregators and distributors of government content, and the returns such entities realize on those investments, are often undisclosed. Lack of such information thwarts library due diligence in evaluating resources for purchase and subscription.

- *Politics:* The actions of governments themselves often have a bearing on the integrity and accessibility of their records. All governments, to a greater or lesser degree, tend to resist public scrutiny of their activities. As Mary Case observed, the behavior of the agencies—and even that of publicly funded archives—can at times be in conflict with the mission of the research library and the interests of researchers.

A *"Leviathan"* Agenda

These are big challenges. To address them, research libraries will require a new "playbook." Existing mechanisms, such as the U.S. Federal Depository Library Program and the Government of Canada's Depository Services Program, were designed to overcome geographic obstacles to broad public access to documents that were tangible and available in limited quantities. The realities of today's "paperless government," global digital networks, and ubiquitous information pose different challenges and call for new approaches.

The broad strategies outlined below, based on the *Leviathan* discussions, are essential elements of a realistic action agenda for U.S. and Canadian research libraries. They constitute a new "playbook" for effective stewardship of government information on behalf of the scholars our libraries serve. As such, they will henceforth form the basis for CRL's own planning and priorities.

1. **Triage:** Given the scarcity of resources for preservation today, libraries must focus their efforts on what is known to be at risk and what is not likely to be adequately preserved by other actors, public sector or private. Because HathiTrust, ASERL, and other organizations have declared their intention to comprehensively archive all U.S. federal government documents, CRL will then focus its efforts on securing access for its community to other digital resources, i.e., materials in established CRL areas of strength, such as records relevant to international affairs, U.S. and its allies' diplomatic history, and American internal and national security; records and documents produced by governments in conflict zones, unstable areas, and in regions of U.S. and Canadian national strategic interest; and materials produced by corrupt and/or non-transparent governments abroad, which are likely to be lost if not independently harvested and archived.

2. **Drill Down:** Determining what is at risk requires a detailed understanding of how the digital government information lifecycle and supply chain work, including the workings of critical archiving infrastructure like NARA's Electronic Records Archive and GPO's FDsys. Too little is known about the cloud services upon which agencies depend, and the proprietary systems and platforms agencies use to distribute information. In the future, "application literacy" will be an essential part of the skill set of both librarians and researchers. CRL will undertake, as resources allow, the analysis, mapping, and documentation of the technologies and platforms used by NARA and Library and Archives Canada to preserve born-digital U.S. and Canadian federal government records. CRL will also seek the support of its constituents and other libraries in the Federal Depository Library Program to conduct an independent evaluation of the FDsys repository system, providing a gap analysis and baseline data that can inform planning for the systematic archiving of born-digital U.S. government agency publications.

3. **Differentiate:** Researchers in different fields require different things of government information: integrity and authenticity mean one thing to an economist and another to a historian. To serve constituencies using new analytical tools tailored to their respective disciplines, libraries will have to abandon the "one size fits all" regimes used to accommodate those constituencies in the print era. CRL will focus its ongoing analysis of research methodologies and practices on the users of electronic records and data in the focus areas identified above, and will publicize those methodologies through its Global Resources Forum.

4. **Collectivize:** Research libraries must actively engage as a community with key "suppliers" of government information, including NARA, GPO, and Library and Archives Canada; and must unite to gain leverage in dealings with key aggregators of government information, such as ProQuest, Bloomberg, and others. Given the scale of the resources behind those suppliers, CRL will work with NERL and other appropriate consortia to ensure that North American research libraries "speak with one voice" in negotiating not only affordable access to government information, but also provisions for long-term accessibility and integrity, quality of metadata, interoperability, and tools for analysis and mining of text and data.

5. **Act Up:** To obtain the resources and standing necessary to play a meaningful role in a realm as large as government information, research libraries will have to forge new partnerships. Organizations now abound that share with libraries an interest in public access to government information and documentation. The National Security Archive, for example, uses FOIA requests, litigation, and other means to compel the U.S. government to declassify and disclose records and information. Though not a library, the NSA has succeeded in making a wealth of critical historical evidence available to scholars and citizens alike.

Under a new effort led by UNESCO, the ICA, IFLA, LIBER, and other partners are working to encourage national governments and the technology industry to promote greater persistence and disclosure of government information. CRL support for the efforts of those open government data initiatives could well accelerate declassification and disclosure of government materials important to scholars. As a first step, as Ingrid Parent recommended, CRL will draft an "advocacy statement," based on the *Leviathan* strategic framework, to be brought to IFLA this summer in Lyon.

Look at http://3stages.org/crl and hold mting to discuss UO's progress on Data for Local Communities http://library.uoregon.edu/dc/dlc/ #CRL_Leviathan

—*Mark Watson*
@mrwatson44

#crl_leviathan Today in WaPo: Magistrate's revolt in releasing digital data http://wapo.st/1ifht1V - retain judges' emails on this!

— *Rachel Brekhus*
@brekhusr

Plz remember #CRL_Leviathan attendees, #fdlp first & foremost a collaborative network of libraries & librarians. That's what's needed now

—*Freegovinfo @freegovinfo*

"@poptheapp: The persistence of memories of the persistence of memory." https://gopop.co/36724 referenced in my #CRL_Leviathan talk

— *John S. Bracken @jsb*

This is the view from the other window at the conference site. #stunning #CRL_Leviathan pic.twitter.com/cpORBoiKgq

—*paulwester @paulwester*

John Bracken: how can we make the internet better? Knight News Challenge @knightfdn @JournalismLib http://www.newschallenge.org #CRL_Leviathan

—*Columbia CHRDR*
@HRDocumentation

Stewardship will, under certain circumstances, continue to involve libraries and archives taking—as they did in the paper era—wholesale custody of materials produced by governments. But this depository role has become less relevant. If we define preservation as protecting and empowering access for our constituents, then archiving becomes only one of several appropriate strategies. More often, however, libraries will have to actively monitor and scrutinize how government records and information are produced, managed, and distributed, whether by the government agencies themselves or by third parties, and identify the faults and deficiencies of those operations.

It has been at least a decade since paper documents ceased being the primary form in which most researchers access government data. The *Leviathan* presentations and discussions suggest that it is time for libraries to adopt a new playbook to support that access. ❖

Center for Research Libraries Staff Contacts
(800) 621-6044

President
Bernard F. Reilly x 334
breilly@crl.edu

Administrative Services Specialist
Yvonne Jefferson x 319
yjefferso@crl.edu

Member Liaison and Outreach
Services Director
Mary Wilke x 351
mwilke@crl.edu

Director of International Resources
James Simon x 324
jsimon@crl.edu

Director of Technical Services
Amy Wood x 327
awood@crl.edu

Director of Information Systems
Patricia Xia x 341
pxia@crl.edu

Head, Access Services
Kevin Wilks x 314
kwilks@crl.edu

Head, Stack Management
Bethany Bates x 339
bbates@crl.edu

Digital Program Manager
Virginia Kerr x 265
vkerr@crl.edu

Communications Specialist
Gwen Ihnat x 289
gihnat@crl.edu

Global Resources Program Contacts
(800) 621-6044

Director
James Simon x 324
jsimon@crl.edu

Project Coordinator—
Global Resources Network
Judy Alspach x 323
jalspach@crl.edu