It Is Not Just a Document: Using Government Data in Teaching and Research

Catherine Johnson
ProQuest, catherine.johnson@proquest.com

Marianne Ryan
Northwestern University, marianne-ryan@northwestern.edu

Melissa Oakes
ProQuest, melissa.oakes@proquest.com

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

Part of the [Library and Information Science Commons](#)

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Catherine Johnson, Marianne Ryan, and Melissa Oakes, "It Is Not Just a Document: Using Government Data in Teaching and Research" (2013). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284315271>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

It Is Not Just a Document: Using Government Data in Teaching and Research

Catherine Johnson, Publisher, ProQuest

Marianne Ryan, Associate University Librarian for Public Services, Northwestern University

Melissa Oakes, Sales Manager, ProQuest

Abstract

Government documents have long been perceived as valuable resources containing unique information content. But documents are also sources of deep, rich numeric and textual data that are available nowhere else. Identifying the specific sources that contain these data, tapping it, and manipulating it can be confusing, at best. But once discovered, the unique data within government publications can contribute to research and teaching in a wide variety of academic disciplines.

By working together, librarians and vendors can suggest ways to derive greater value from government information resources, explore new ways of thinking about text and data patterns, and assess some of the challenges and opportunities facing faculty, researchers, and library and information professionals as the expectations and possibilities regarding use of such digital content changes.

New Ways of Thinking about Government Documents

Government information resources have always been perceived as a valuable and unique content, though not necessarily universally appreciated—in part because of the difficulty of use. Users and expert librarians would spend hours poring through shelves of materials and mysterious print indexes to try to identify and locate the exact piece of information desired—or something that would fit the need.

Back in the day, the language of “documents” was pretty straightforward. But even then, the data aspect of government publications was enormously important—it is just that it was mined manually. The Federal Government has always been the world’s largest producer of statistical information, and documents have always been sources of deep, rich data (before it was called data) that is available nowhere else. In fact, government information is the original big data. When a government document was thought of as “just a document,” librarians and researchers used a vocabulary to describe these elements that included words such as report, statement, vote, list, statistic, graph, table, guide, and map. Although this language is still valid, a new terminology is emerging that reflects changing research needs. This new vocabulary includes

words such as data, pattern, text mining, importing/exporting, visualizing, layering, geo-referencing, intersections, and manipulating.

As information has increasingly become available electronically, there has been increased awareness of the richness of the content and greater accessibility to it. Since the mid-1990s, significant amounts of government content have been born digital, with now only a small percentage being issued in print. And as new tools have been developed to provide more robust access to that information and the ability to extract and manipulate the data contained in it, demand and expectations have increased, bringing new challenges. In addition to format changes that have affected user expectations, other factors have contributed. One is a trend toward greater interdisciplinarity in education. Gone are the days when a majority of students and faculty focus on isolated areas of study. Academia today reflects a global society, an entrepreneurial environment, and a problem-solving world. Much research focuses on the blending of disciplines (across sciences, social sciences, and humanities), which requires identifying patterns and trends in data across multiple areas. Government information resources, the quintessential interdisciplinary content, are ripe for meeting these blended expectations.

Another factor has been the changes in users themselves and in the generational differences in their information-seeking behaviors. A good example of this is described in a chart on the ALA web site (<http://www.ala.org/rusa/sites/ala.org.rusa/files/content/sections/rss/rssection/rrsscomm/virtualreferencecommittee/an07infosekgen.pdf>). As it shows, researchers born before World War II are accustomed to a top-down flow of information and stable learning environments; they often prefer materials that are organized and summarized. Extrapolating this model to the government documents world, this type of researcher might be characterized comfortable with documents arranged by SuDoc number. Members of the Boomer Generation, on the other hand, came of age in an era of more interactivity and ease of use. Because they began their research career working with microform versions of government publications, it is not surprising that these researchers are familiar with using supplementary finding aids. Generation X-ers are highly independent and generally more comfortable with change. When working with government information resources, they expect to be able to do keyword searching and cut and paste. Today's researchers, who began their research experience online, are more media savvy multitaskers who are at ease with technology and collaborative work. When they use government content in their research, they are apt to think of projects that involve data, images, and the manipulation of content in a more integrative way.

Just as teaching, learning, and research are increasingly team-based and dynamic, support for them must be, too—between content developers, managers, and libraries. So what are libraries doing to support? One example can be seen in a recent restructuring at Northwestern University Library where, until recently, a large depository library was located in a stand-alone department that handled all government information (including international), social sciences data, and maps. In response to changing patterns of information-seeking behavior, feedback from constituents, and the evolution of new university and library strategic plans, that department was transformed. Government information is now part

of a unit within the Research and Information Services Department that includes geospatial, business, and economics information and data services. The other unit in the department has support for general information services and global and international studies, emphasizing the importance of having those areas in tandem. Positions are being recast to provide needed expertise. Many other libraries are taking similar measures to change the structures and staffing that support cross-disciplinary data services.

Looking for Patterns

As new ways of thinking about government information emerge, librarians are being asked types of questions by researchers that differ from the questions of yesterday. Vendors are being asked to provide different tools. Many research projects that involve a new way of thinking can be completed with existing desktop tools, but the work is often extremely laborious and time consuming.

If we think about government documents in the traditional document-by-document way, individual government publications can be found by searching the full text or metadata using a database such as ProQuest Congressional. A researcher interested in historic documents about Native American treaties might use the keyword "Cornplanter" to retrieve such things as speeches by the Seneca Chief Cornplanter who signed a treaty with the United States on January 9, 1789. Additional searches for Cornplanter within the U.S. Serial Set produce a map from 1890 showing the Cornplanter Reserve in Warren County, Pennsylvania, and a House report from 1908 describing the efforts of Cornplanter's heirs to gain standing in court to attempt to recover property near Marietta, Ohio, purportedly granted to Chief Cornplanter in his lifetime.

An example of a new way of thinking about the Seneca documents would be to devise a comparison based on the number of documents referencing Native American treaty signers in the entire corpus of Congressional information. The number of documents referencing Chief Cornplanter compared with references to the Cherokee Chief Nenetoyah can be plotted on a

graph to show that Congressional interest in Nenetooyah waned over time, while interest in Cornplanter spiked at specific points. For example, spikes of interest occur during the period in which Cornplanter's heirs were pursuing their land claims, as well as a period prior to the 1965 creation of the Kinzua Dam in Warren County, Pennsylvania, when the building of the dam was contested because it would (and eventually did) submerge lands promised to Cornplanter and his heirs by George Washington.

Many historic and current documents refer to specific geographic locations. Patterns based on references to geographic place names in Congressional documents over time might indicate, for example, whether references to Indians in their States of origin and the locations to which they removed remained relatively consistent, or whether a drop in references to the place of origin began to occur once Indian removal policies were put in place. Visualization of voting patterns is also instructive when ye and nay votes are plotted by political party and geographic regions. Using common desktop tools, researchers can plot by region a comparison of the ye and nay votes on the Indian Vaccination Act of 1832 with votes on the Swine Flu Vaccination Program of 1976. What would be even more interesting would be a comparison of votes on all Congressional bills related to vaccination programs over time by region, but without specialized tools this would be a lengthy project.

Comparative textual analysis of language can also be revealing. A casual characterization and comparison of language found in Member statements in the *Congressional Record* and its predecessors regarding vaccine programs shows that Members from different time periods made their arguments by using different types of language. A characterization of language used by Senator Alexander Buckner of Missouri in the April 16, 1832, *Register of Debates* in opposition to the Indian Vaccination Act of 1832 indicates that more than 25% of the nouns used were negative words intended to evoke images of fear, such as injury, enemy, ferocity, desolation, death, and cruelties. Although he makes the point that he is opposed

to providing vaccinations to the Indians while a proposal for a small appropriation for a hospital in his district had been denied, he uses relatively few words directly related to money or health.

By contrast, a May 26, 1976, statement made by Representative Henry Waxman regarding the Swine Flu Vaccination Program of 1976 includes a large number of words related to the science of health, such as scientific journals, evidence, clinical trials, proper dosage, data, and adverse reactions. Words related to money are rare, but the cost issue is touched upon briefly in phrases such as "cost of care" and "future earnings of those who died." Waxman's arguments, however, are consistently grounded in the language of science and health.

The opening statement of Representative Phil Gingrey of Georgia in a November 18, 2009, Congressional committee hearing on the adequacy of H1N1 vaccine supplies contains only five brief paragraphs, with four paragraphs containing language with money-related words: appropriated, billion, fiscal, dime, IOU, fiscal health, debt, and pocketbook. Gingrey makes a point similar to the point made by Senator Buckner in 1832 about the importance of not giving more to our "enemies" than is given to our citizens when he expresses concerns that prisoners in Guantanamo Bay might be vaccinated without ensuring that there is enough vaccine left to vaccinate elderly citizens.

More complex textual analysis might be used to illuminate the context surrounding all Member statements in the *Congressional Record* and committee hearings on vaccine programs over time.

Plotting the incidence of word usage historically indicates the popularity of specific words at different times. The incidence of health-related words such as antibiotic, antiseptic, vaccines, disinfect, leeches, and quarantine in congressional documents is obviously related to the status of scientific knowledge at any given time. Vaccination is a concern of Congress from the earliest years forward, but interest spikes following the testing of the Salk vaccine for polio in 1952. Examining the incidence of these words

in the context of scientific discoveries and historical events shows that the death of U.S. President James Garfield from sepsis following an assassination attempt occurred at a time when Joseph Lister was already making known his ideas regarding antiseptic surgery, but that Congressional (and therefore possibly popular) interest in the subject had not yet begun to emerge.

Patterns in witness affiliation testimony in Congressional hearings over time are also indicative. For example, witness affiliations of associations with names that contain the words smoking, tobacco, cigar, or cigarette can be characterized as organizations supporting labor unions, wholesalers and retailers, manufacturers, farmers, exporters, and proponents of antismoking. A researcher looking for just a single document might be interested in this early antismoking testimony of Charles M. Fillmore in a 1926 Congressional committee hearing:

The Chairman: What is your business?

Mr. Fillmore: I am general secretary of the No-Tobacco League of America.

The Chairman: This is getting to be a good, free country!

Mr. Fillmore: Sure, it is a free country.

The Chairman: Pretty soon we won't be allowed to smoke. I am getting pretty old.

Mr. Fillmore: Well, Senator, we are not after you. We are after the children. We hope to raise the new generation a little better.

The Chairman: You have got me!

Mr. Fillmore: We are not wasting time on fellows like you.

Mr. Chairman: I hope you will not operate on me, anyhow, until after I am dead.

All right. That is all in good part.

Mr. Fillmore: Sure. Sure, I am used to that. It is part of our business.

The above is an amusing exchange. But looking for patterns in affiliations provides an entirely different perspective, showing, for example, a sharp spike in farmer organization testimonies in mid-twentieth-century years and an uptick in antismoking testimony in the years after 1981 as awareness of the dangers of smoking increased.

Challenges: Today and Tomorrow

New ways of thinking about government documents suggests a future that has not yet arrived. Librarians are indeed getting asked different types of questions, and individual researchers are undertaking projects that they accomplish through innovative means of their own devising, but many librarians are still struggling to get users to understand the value of Congressional content. For example, Congressional documents include a wealth of information in science, technology, and engineering areas, but many STEM researchers would not immediately think of Congressional information as a possible source of value.

Still, text mining and analysis offer additional opportunities for research using government documents. A health researcher interested in a single document on antibiotic resistance might cite a single GAO report; a researcher using text mining or analysis techniques might examine language or numeric data in all government documents to assess public awareness of the relationship between the use of antibiotics in food animal industries and the increase in antibiotic resistance. A women's studies researcher interested in definition of rape and sentencing might cite a single Congressional Research Service (CRS) report; a researcher interested in historical patterns might examine all bills and laws related to rape over time. A political science researcher might look at the context of the word "terrorism" over time while a business researcher might be interested in comparing historic and current asset depreciation studies.

Although it seems clear that researchers today are thinking about government text and numeric data in ways that transcend the limitations of individual documents and that this trend is on the rise, the specifics of emerging user needs have yet to be

fully defined and understood. By working together, librarians and vendors can improve their understanding, develop appropriate strategic

approaches, and ensure that researchers of the future have the content and tools they need to carry out their research.