Analyzing Long-Term Usage Trends of Electronic Government Information

Celina McDonald, University of Maryland Libraries, College Park, MD 20742

Introduction

As academic and research libraries undertake large-scale collection realignment projects in order to reduce their collections' physical footprint, government documents collections are often seen as ideal candidates for shrinking. In addition to meeting basic withdrawal criteria such as currency, uniqueness, and usage, many federal publications both current and historic are available electronically from no-fee and subscription sources. Furthermore, maintaining tangible government documents collections requires a significant financial and labor investment.

Unfortunately, institutions' efforts have largely been stymied by incomplete holdings and usage data for their tangible materials. When libraries transitioned to OPACs in the 1970s and 1980s, government documents were not universally incorporated into the systems. Although great strides have been made in subsequent years, many libraries continue to have significant portions of their collections uncataloged. A byproduct of this oversight is that it limits the quantity and quality of usage data that can be gathered on these materials. Because cataloging these materials and acquiring sufficient data demands a significant resource investment, many libraries undertake collection weeding projects without sufficient data.

The purpose of this study was to harvest, analyze, and identify ways electronic usage data of government publications could be used to make collections decisions. Previous studies have explored similar themes; however, limitations on the depth and breadth of usage data hindered efforts to develop a full understanding of library users' government information use.

Methods

The study was conducted at the University of Maryland (UMD), a large research institution in College Park, Maryland. The UMD is the regional federal depository library for Maryland, Delaware, and the District of Columbia. In addition to its tangible materials, the University Libraries provide online access to government information through federal and vendor resources.

For this study, vendor database usage statistics were harvested from usage reports from 2007-2016. The reports accessed contained data harvested from COUNTER and non-COUNTER compliant sources and had been aggregated for internal use. Because of changes in metadata naming and data collection practices over time, a thorough review of the data was required to eliminate duplication and misleading data. For example:

- ProQuest session data did not correspond to a specific database, but rather combined session data for all products. Deduplicating this data was necessary to avoid an artificially high usage count.
- Ebsco statistics were not COUNTER compliant. Usage for all consortial users was combined into a single count, making it impossible to isolate usage from the University of Maryland.
- Vendors and databases, such as the Digital National Security Archives (Chadwyck-Healey), were acquired by other companies (ProQuest), which could lead to discontinuous data if not carefully tracked..
- UMD Libraries acquired and discontinued databases, which resulted in data gaps, even when users had uninterrupted access (e.g., AGRICOLA, which was available through FirstSearch for 2007-2010 and Ebsco from 2010 to present.)

This study also included usage data for electronic federal publications from the Government Publishing Office (GPO). GPO produces Persistent Uniform Resource Locator (PURL) reports for federal depository libraries. Data harvested from the GPO includes:

- Detailed reports for the most recent 12 months, and
- Annual general access data.

The detailed reports are accessible through the FDLP PURL Usage Reporting Tool. Libraries can request data for specific hostnames and IP addresses. The raw data was then reviewed to ensure only UMD-College Park data was included in the study. Reports for 2007-2015 were similarly reviewed; however the data is for basic usage. Each row of data (for a database or a specific title) was categorized by its publishing agency.

Decennial PURL Usage

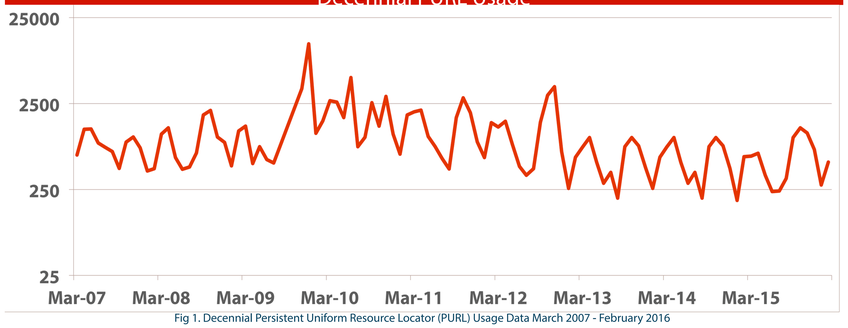

Fig 1. Decennial Persistent Uniform Resource Locator (PURL) Usage Data March 2007 - February 2016

12-Month Detailed PURL Usage Data

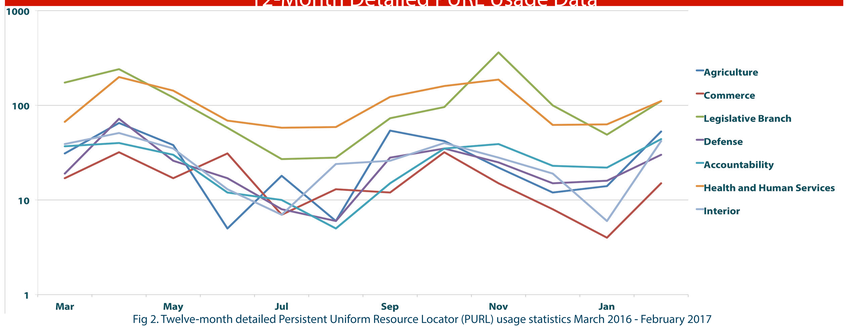

Fig 2. Twelve-month detailed Persistent Uniform Resource Locator (PURL) usage statistics March 2016 - February 2017

Decennial Detailed Vendor Database Usage Data

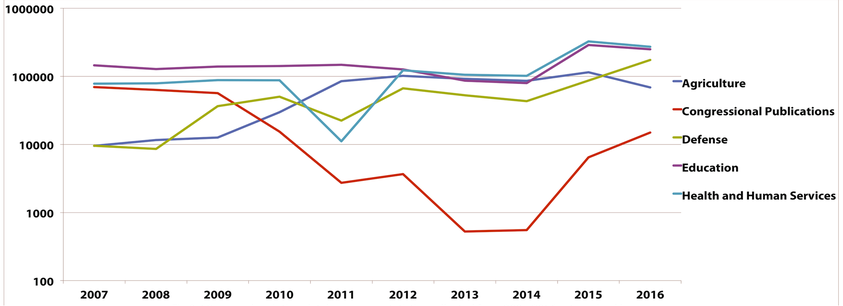

Fig 3. 10-year subscription database usage statistics collected from 2007-2016 categorized by agency

Results

Analyzing the data presented several challenges. The most significant issue was the non-COUNTER compliant usage data from Ebsco, which provides access to content from the Departments of Agriculture, Defense, Education, and Health and Human Services. Because an accurate representation of UMD-CP usage is not available, ideally this data would be excluded from the analysis. Unfortunately, that was not possible for several reasons:

- Some content was only or predominantly available through Ebsco. For example, MEDLINE (a known high-usage database at UMD) was only available from Ebsco in 2016. After much consideration, the Ebsco data was included in the analysis because it was decided that including the data would provide a more accurate picture of usage than would excluding it.
- Detailed reports accessed through the FDLP PURL Usage Reporting Tool showed that the most frequently used content came from the Departments of Agriculture, Defense, Education, and Health and Human Services. Strengthening this conclusion was the similar trajectories followed by FDLP and vendor data.
- Another issue with the vendor data was duplication. Database usage reports included data points for searches, total searches, sessions, total sessions, views, visits, full-text article retrieval attempts and successes. To avoid duplication and data gaps, session, visit, views, and full-text retrieval data were omitted from the study. Instead, searches were the metric used to measure database usage.

Data from both GPO and the vendor usage reports included significant gaps. In particular, GPO servers regularly experienced outages from 2007-2009. Notably, from August - October 2009, no data statistics were available due to frequent service interruptions. 2009 data was further complicated by a suspicious usage spike in December. Long-term trends across multiple databases and years point to a reliable decrease in usage in December, but in 2009 there was a fourfold increase of usage from November to December. The most likely explanation for the spike was determined to be a single individual engaging in massive, possibly automated, downloading of content over a short period of time.

Analysis of the detailed PURL usage data was facilitated by the inclusion of Superintendent of Documents (SuDoc) numbers, which organizes government publications by agency rather than than subject. Each SuDoc number starts with a letter that corresponds to a federal agency or department.

This allowed an easy method of assigning categories to most retrievals, giving a detailed view of usage trends for different types of publications corresponding to the arrangement of the physical collection. Over 330 hits did not have SuDoc numbers and were therefore not included in the study.

Conclusion

Data from vendors and the GPO can help libraries gain insight into users' information needs.

- Despite flaws in the data sample, both sets of data indicated users' reliance on content published by the Legislative branch and Departments of Agriculture, Defense, Education, and Health and Human Services. This information will be used to make recommendations for transfer/withdrawal of tangible materials, which are rarely used and occupy a lot of space.
- Although categorizing materials by agency makes it difficult to analyze data for subject matter, it will facilitate collections projects in the future because materials are already organized in this manner.
- Analyzing the PURL data was easier than anticipated and a more in-depth analysis should be undertaken in the future. Recently researchers in the medical fields used text-mining to analyze literature in their field to identify information gaps. The existing data tables can be analyzed using more advanced statistical analysis programs such as R, SPSS, and SAS to gain more insight into users' subject matter needs.

Use of vendor data will require more experimentation and exploration. Due to in variations in metric naming practices from publisher to publisher and year-to-year, it was difficult to develop a clear picture of database usage. Although it is clear that some agencies are used more than others, it is not possible to see how great the difference is. Newer tools from Ebsco and more detailed reports from COUNTER-compliant vendors should provide greater insight in the future.

Despite difficulties analyzing the data, undertaking this type of project at least every five years can prove beneficial because in addition to showing areas where materials can be withdrawn or transferred to off-site storage, it can be used to support electronic database purchasing decisions. For example, content gaps recently identified in UMD's electronic congressional publications collection should be filled as soon as possible in order to ensure users' needs are being adequately met.