Taking count: A computational analysis of data resources on academic LibGuides in the U.S.

Cody Hennesy[1], Alicia Kubas[2], and Jenny McBurney[3]

Abstract

The LibGuides platform is a ubiquitous tool in academic libraries and is commonly used by librarians to compile and share lists of recommended social science numerical data resources with users. This study leverages the machine-accessible nature of the LibGuides platform to collect links to data and statistical resources from over 10,000 LibGuide pages at 123 R1 research institutions in the United States. After substantial data cleaning and normalization, an analysis of the most common resources on those guides provides a unique window into the data repositories, libraries, archives, statistical data platforms, and other machine-readable data sources that are most popular on academic library guides. Results show that freely available resources from U.S. government agencies are among the most common to be included on data and statistical resources guides across institutions. Resources requiring paid licenses or memberships for full access, such as Statistical Insight (ProQuest), Social Explorer, and ICPSR are linked to most frequently overall, regardless of the percentage of institutions that include them. Findings also suggest that libraries are more likely to share traditional licensed statistical resources (e.g., Cambridge's *Historical Statistics of the United States*) and collections of simple charts and graphs (e.g., Statista) than more robust and complex microdata resources (e.g., IPUMS).

Keywords

data reference, LibGuides, data librarian, web scraping

Introduction

Over the past several decades, as user expectations for data support have grown, academic libraries have increasingly served as hubs for data and statistical resource discovery and assistance. At the same time, the tools that librarians use to share online resources with their users have shifted significantly, with the LibGuides platform becoming a ubiquitous resource discovery platform in academic libraries. Library guides focused on numerical data and statistics resources have proliferated at a rapid pace, alongside more traditional disciplinary library guides, and currently represent a significant platform for researchers to navigate the sometimes-vexing world of data resource discovery. Indeed, finding and utilizing datasets in the social sciences can prove challenging for researchers who are not yet familiar with how data are organized, with the varieties of data types such as "published statistics, microdata, macrodata, survey data, longitudinal data, geospatial data," or with where datasets on particular topics can be found (Rice and Southall, p. 52, 2016). Along similar lines, data resources can be difficult for even information professionals to understand, as Bauder points out: "trying to find data can be a frustrating experience" for librarians who are more often familiar with bibliographic discovery tools (p. 11, 2014). The directories of data resources shared on LibGuides, then, are not only intended as guides for researchers who are new to data topics but also help librarians themselves keep track of the large number of agencies, commercial vendors, nonprofits, and repositories that provide access to data and statistical resources. While several books and many articles have been published in library and information science outlets to help library workers acquaint themselves with key data resources,

the LibGuides platform itself has become a compelling primary source for understanding the real-life recommendations of data librarians.

This study leverages the widespread presence of online library data guides to compile lists of the most common freely available and licensed data and statistics resources shared by academic libraries. To gather resources from library guides, the scope of institutions included was first limited to a subset of universities and colleges in the United States designated as R1 institutions, described as 'Doctoral Universities - Very high research activity' by the 2018 *Carnegie Classification of Institutions of Higher Education* (American Council of Education, 2022). Libraries at R1 institutions—where doctoral students and faculty often require access to social science data—consistently devote resources to data research support, and frequently use LibGuides to organize access to data resources. Due to the popularity of Springshare's LibGuides platform in U.S. academic libraries it was possible to automate the collection of data resources from the relevant guides. In 2021, for example, 91% of 799 academic libraries surveyed, including 95% of 132 Doctoral (R1) institutions, used LibGuides (Neuhaus, et al.). This ubiquity makes the platform an excellent primary source for macro-analyses of the kinds of resources that librarians share for specific disciplines or topical research areas. The shared HTML structure across the platform provides consistent access points for systematic downloads across institutional boundaries, creating opportunities for quick and thorough data collection.

Ultimately, resources from 10,448 guide pages related to data and statistics were collected from 123 of 131 R1 institutions (93.89%). Metadata related to 186,952 non-unique links were initially collected from these guides and, after substantial cleaning and normalization, 64,131 unique resources were analyzed to compile lists of the most common data resources across several categories. The final compilation of top resources leveraged both URLs and normalized link names to identify the most common data resources at this specific subset of academic libraries in the U.S. This is the first study of its kind to examine and compile data resources from real-world library guides as a path to explore the resources that data librarians find most essential to share with their users.

Literature review

As the available resources for data and statistics have grown and evolved over the last few decades, so have the strategies that librarians use for discovering these resources and highlighting them for users and other librarians in the field. The rise of the Internet and the ability to provide online data access in more accessible files and formats vastly increased the variety and availability of data and statistical resources (Kellam and Peter, 2011). In the 1980s and 1990s, the focus shifted from solely relying on print data and statistical resources to incorporating more online websites and resources. This ushered in more library trade publications where librarians focused on website lists and recommended sites for statistical resources, such as *Link-Up* and *ONLINE*, both magazines from Information Today (Berinstein, 1998; O'Leary, 2000).

In the current landscape, there are many potential avenues for finding data and a proliferation of sources and formats to search or consult along the way. Data can encompass both quantitative and qualitative data as well as nonnumeric sources like textual data, audio files, and other corpora (Johnson, 2019). As for more traditional numerical data and statistical information, sources span government surveys and databases, international agencies, non-profits, private organizations, researchers and academic journal articles, data repositories, and more (Johnson, 2019; Kellam & Peter, 2011; Bauder, 2014; Geraci et al., 2012). In particular, government data from local to

international levels is becoming more commonly shared via open data portals (Huck, 2020). Librarians have compiled data guides and composed specific resource reviews to help others navigate the complex world of data reference. Multiple book-length guides exist on the topic: Kellam and Peter compiled a comprehensive list of basic sources in 2011, covering paid, free, and hybrid resources, while Bauder published a reference guide to freely available, online data sources in 2014. More recently, Johnson published a practical guide for data librarians in 2019, which includes a lengthy appendix of freely available data sources.

In addition to these more exhaustive lists of data sources, resource reviews for specific resources have been published in *CHOICE*, the *Charleston Advisor*, and other venues throughout the 2000s and continue today (O'Leary, 2000; Carroll, 2001; Stark and LaGuardia, 2003; Feldmann, 2011; Geck, 2013; Jakub, 2014; Verma, 2014; Werner, 2014; Geck, 2015; Rodriguez, 2016; Geck, 2020). Data resources are also sometimes published as part of larger lists of recommended reference resources (Etkin and Coutts, 2015). In addition, recommendations for data sources in particular disciplines or subject areas have been published in a variety of venues on business, education, public health, and art topics (e.g., Boslaugh, 2007; Smith, 2008; BRASS Education Committee, 2012; McNulty, 2013). *IASSIST Quarterly* published a double issue in 2009/2010 focused on discipline-specific data with articles providing overviews and lists of sources on various topics, including the American Community Survey, microdata on developing countries, and sources for international labor data, among other topics (Bordelon).

Librarians often rely on library research guides to help organize the complex landscape of data and statistics resources across different disciplines, and surface both subscription and free online resources for researchers' use. The most widely used platform for library guides is Springshare's LibGuides, which is used by over 5,700 institutions across 105 countries (Springshare, 2022). As early as 2011, Kellam and Peter noted that LibGuides are a key resource for many university libraries, and that "the LibGuides 'community search' function can be a quick way to find a librarian-created research guide to locating statistics on a particular geography or topic, such as health, business, or finance" (p. 100). Since LibGuides are so widely used, the existing library literature includes many best practices for designing and organizing research guides and addresses how librarians have attempted to optimize LibGuides for sharing data resources specifically (Hoffman, 2015; Wheatley et al., 2020). Guides, in general, have been used for data collection management purposes—identifying existing collection gaps, for example—and tracking statistics on frequently used resources (Rice and Southall, 2016). Because many data resources are born digital and available online, highlighting resources on a platform like LibGuides is essential for access and use by researchers (ibid.).

However, library research guides also present significant challenges, especially with regards to keeping resources up to date and the proliferation of broken links. Librarians describing their methods for guide maintenance note that this work can be quite tedious and is often a low priority when more pressing work arises (Ornat et al., 2021). However, because LibGuides are considered an important portal for researcher discovery, many librarians have made efforts to address these entropic forces, holding events such as LibGuides parties to encourage better guide maintenance (ibid.).

Varying institutional contexts, combined with the data needs of specific institutional users, leads to a wide variety of data service models. Some data reference services include dedicated data librarians with different areas of data expertise; others have a single designated data librarian working in conjunction with other subject librarians; while others have no official data librarians and rely on their

subject librarians to gain data expertise in their subject areas (Bordelon, 2009/2010; Geraci et al., 2012; Rice and Southall, 2016; Foster et al., 2019). Additionally, reference service points may receive data-related questions regardless of who is currently staffing the desk. Research guides for data and statistics can be essential for library staff who are helping researchers answer data-related questions. Data and statistics LibGuides are frequently used as data directories by both library patrons and library staff.

Library and information science scholars have frequently looked to LibGuides as primary sources for identifying resources that are key to particular fields, often using manual counts of resources from dozens of disciplinary guides to embark on content analyses. Of these studies—looking at guides in theater (Furay, 2018), theology (Van Dyk, 2015), electrical engineering (Osorio, 2014), geology (Dougherty, 2013), physics (McCormick, 2020) and nursing (Stankus and Parker, 2012)—anywhere from 37 to 100 guides have been examined at a time, often from a sample taken from a larger collection of relevant guides. Along similar lines, common resources have been compiled from topical guides devoted to 3D printing (Horton, 2017) and those aimed at physician assistants (Johnson and Johnson, 2017). Many of these studies break resources down by format, reporting on the most common databases, books, ebook collections, journals, and websites that were listed on the subject guides. Of those looking at databases, research studies have compiled and analyzed lists ranging from 72 to 143 unique databases. While this study does not discriminate by resource format, by leveraging computational methods it significantly expands upon the range of both the number of guides (10,448 guide pages) and number of unique resources (64,131) under study. It is also the first to examine the presence of data and statistical resources on academic LibGuides.

Methodology

Data and metadata from data and statistical resource links that are shared on LibGuides were processed in three broad steps: data collection, cleaning, and analysis. Data collection and analysis were performed using Python in a JupyterLab computing environment, while portions of the cleaning and normalization processes were performed using both Python and OpenRefine. While data scientists joke that 90% of data research projects consist of data cleaning, with only 10% of the work involving analysis, this project, unfortunately, reflected an even more lopsided balance. A significant amount of the labor for this study involved the collection, cleaning, and normalization of inconsistently structured data. The analysis itself consisted of a comparatively straightforward compilation of sums, averages, and percentages to show the most common data and statistics resources on academic library guides.

Data collection

Data were ultimately collected from 123 Carnegie R1 institutions using a list of LibGuides URLs previously compiled by Hennesy and Adams in a study of organizational practices for LibGuides at academic libraries (2021). While 124 of 131 R1 institutions used the LibGuides platform, one of those institutions—Columbia University—was dropped from the study since the LibGuides search interface on the site did not function such that guides related to data and statistics resources could be accurately identified. The final dataset under analysis reflected data and statistics guides from 123 institutions[4].

Data and metadata were scraped from LibGuides using the *requests, Beautiful Soup*, and *Selenium* Python packages, following a determination that the research constituted a fair use (Reitz, 2015; Richardson, 2015; Selenium Developers, 2021). The data was collected in two general steps on November 18 and December 8, 2021. First, URLs for all guide pages that related to data or statistics were extracted. No distinction was made between various guide types (e.g., course, subject, or topic guides), and URLs were compiled at the level of individual pages, also often referred to as 'tabs.' For an Economics guide with a tab called *Data resources*, for example, the URL for the *Data resources* page/tab was collected. Second, the resource links from every guide page collected in the previous step were compiled. Throughout both web scraping steps, several 'pre-cleaning' processes were also undertaken to increase the relevance of the compiled data.

In the first phase of data collection, the terms *data* and *statistic*[5] were systematically submitted as keyword searches of the LibGuides platform at each institution, using Selenium to page through the full set of search results. Text string matches for the terms *data* or *statistic* were compared against the title of each search result, which represented a single page/tab from a guide. When a match was found, the URL and title for the guide page (and parent guide) was collected and stored in a *Pandas* dataframe (Reback et al., 2022). An initial set of 22,514 guide pages was refined in several ways. First, the results were deduplicated to eliminate 7,029 identical guide pages that were retrieved by searches for both terms. Second, the authors scanned the initial results to identify keywords in guide and page titles to identify and remove guides that were unlikely to provide directories of data or statistical resources. The authors identified text strings from the initial search results that indicated guides focused on topics such as software tools, data visualization, data management, data sharing, data publishing, data analysis, text mining, citations, reproducibility, workshops, tutorials, and more. Matching on those text strings eliminated an additional 4,233 guide pages. This likely eliminated some relevant resources from the final dataset, but it was an essential step to improve the overall signal to noise ratio. Finally, 118 guide pages including the term *mathematics* in the title were removed (unless they also included the term *data*), due to the common appearance of subject guides dedicated to Mathematics and Statistics, which primarily focused on bibliographic resources.

The second phase of data collection involved scraping content links from the 11,134 guide pages compiled in the previous step. Links from the page header and footer, guide navigational links, and links in librarian profile boxes were skipped, while the collection focused solely on links shared in the primary content boxes of each guide. Each link was ultimately represented as a row in a Pandas dataframe with columns for the resource URL (e.g., https://libguides.asu.edu/ipoll) and name (e.g., 'Roper Center For Public Opinion Research (iPoll)'), along with the URL from which the resource was collected, and LibGuides' site id for the institution.

Data cleaning

Data cleaning steps were taken to both reduce the number of irrelevant resource links in the final dataset and to normalize the names used to reflect those resources, so that they could be counted accurately. Initial steps to drop irrelevant rows from the dataset were conducted in Python, followed by a series of filtering, faceting, and clustering efforts in OpenRefine, with the aim of renaming resources consistently across the dataset. Beginning with a set of 227,639 resource links, 4,146 resources were dropped because either the URL or name of the resource was empty (common for image links). Another 1,119 resources were removed when a resource URL matched a guide URL that was already present in the dataset to exclude links to other LibGuides from the analysis. The resource

names were then converted to lowercase and stripped of extra whitespace to enable names to be clustered regardless of small differences in text strings. Resources were exported to a Google sheet where the authors manually explored the data, seeking to identify keywords that could be used to filter out common irrelevant resources. Regular expression matches for specific text strings were used to remove 24,110 links with terms related to specific software tools, and resources related to data visualization, data management, data sharing, data publishing, data analysis, text mining, qualitative data analysis, citation tools, workshops, and tutorials. Regular expressions were also constructed to identify substrings from resource URLs that identified irrelevant resources, excluding 10,594 more links. URL substrings were identified, for example, to remove links that did not point to other websites (e.g., URLs that started with *mailto:*), root domains that were irrelevant to the project (e.g., *libapps*, *youtube*, and *creativecommons.org*), and frequently occurring links from specific institutions (e.g., links to library services). An additional 536 resources were removed that had URLs that did not begin with *http* or had malformed URL schemes, and 182 resource names that had no alpha-numeric characters (usually punctuation marks) were removed. By dropping resources from the dataset as outlined above, the number of guide pages ultimately reflected in the final analysis was also reduced from 11,134 to 10,448.

At this stage, Python was used for an initial round of resource name normalization, before exporting the data to OpenRefine. URL schemes and common proxy prefixes were split from resource URLs to create a list of simplified URLs. The resource names for the 1,000 most common simplified URLs in the dataset were re-assigned using a Python dictionary which replaced the name as extracted from the LibGuide with a controlled term for the resource. A resource with the string *dataplanet.sagepub.com* in the URL, for example, was assigned the name *data planet (sage)*, and resources pointing to *cdc.gov/nchs* were renamed *nchs (u.s. national center for health statistics)*.

Next, the dataset was exported to a CSV file and imported into OpenRefine for further normalization of the resource names. Similar resource names were identified and merged using the 'Cluster & Edit' feature, with the default 'key collision' method and 'fingerprint' keying function selected. Each cluster was examined and similar resource titles such as *global health (ipums)* and *ipums global health* were merged under a single name. While this unsupervised method had a similar outcome as the normalization of names using regular expressions in Python, it had the added benefit of identifying similar titles that the authors had not specifically looked for.

Next, text facets in OpenRefine were created of the most common URL domains in the dataset. Resources for the 150 most common domains were selected one at a time, faceted by their simplified URLs, and then renamed *en masse*. By creating a facet for resources with a domain of *www.census.gov*, for example, resources such as *www.census.gov/acs/www* and *www.census.gov/programs-surveys/acs* could be simultaneously renamed *u.s. census: american community survey*. Throughout this process a list of common proxied resources was compiled, including resources such as *social explorer* and *passport (euromonitor international)*. Resource names were filtered using a variety of terms common to each specific resource or platform, and then renamed using the preferred form. Filters for the terms *euromonitor*, *passport*, and *gmid*, for example, were each used to find possible matches for Euromonitor International's Passport platform, all of which were then renamed *passport (euromonitor international)*. Finally, the 'Cluster & Edit' feature was repeated at the end of the normalization process to correct for resources that had been imprecisely renamed during cleaning. Throughout the normalization process in OpenRefine far too

many judgments about specific resources were made to document in full here. Each decision was guided by the intention of "extract[ing] the useful signal from the noise," with regards to how librarians named links to data and statistics resources on their LibGuides (Au, 2020). Ultimately, 16.81% of the unique resource names in the dataset were normalized, reducing the number of unique resource names from 77,094 (following the initial data cleaning in Python) to 64,131.

Data analysis

A CSV of the normalized data was exported from OpenRefine and merged with the original dataframe in Pandas, representing 186,952 resources from 10,448 unique guide pages[6]. Common Pandas functions were used to count, group, and sort values in the dataset in a variety of combinations. Descriptive statistics were generated to understand the scope of the data, looking at the number of unique resource names before and after normalization, the number of guides and resources included in the final dataset, and the average, maximum, and minimum number of resources included from each institution. Tables were generated by sorting the data by the most common resource names, domains, top-level domains, and simplified URLs. The percentage of institutions that included each of the 500 most common resources was calculated and sorted to show the resources that were most commonly found across institutions (see Table 1).

To provide a better picture of the freely available websites reflected in the dataset, a separate count by root domain was generated (see Table 2). Root domains exclude both URL paths and subdomains, allowing for the compilation of resources such as *www.census.gov*, *data.census.gov/cedsci*, and *factfinder.census.gov* under the shared root domain of *census.gov*. An unfiltered list of the most common root domains, however, also includes many URLs for library catalogs (*exlibrisgroup.com*), single sign-on servers (*openathens.net*), proxies, and link resolvers, often appearing under the root domain for an academic institution (e.g., *harvard.edu*, *upenn.edu*). Because this kind of root domain obscures the resource they point to, they were excluded from Table 2, essentially dropping licensed resources from this view of the data. For example, *harvard.edu* had a total of 4,480 results, but was excluded because the two most common Harvard full domains were *nrs.harvard.edu* (1,120 results) and *id.lib.harvard.edu* (1,069 results), neither of which help us identify links to data or statistics resources. Additionally, while Harvard Dataverse is an important resource, the *dataverse.harvard.edu* domain had only 478 results, while other top Harvard domains pointed primarily to library guides (382 results) and the library catalog, HOLLIS (123 results). On the other hand, *umich.edu* remained on the top root domain list since the most common University of Michigan domain in the dataset was for *icpsr.umich.edu* (2,222 results), a relevant data resource.

Finally, to better understand the role of paid resources as sources of data and statistics on R1 LibGuides, a spreadsheet of the 500 most common resources in the dataset was annotated manually to note whether a resource was *paid*, *free*, or *hybrid*. The *hybrid* tag was applied to resources that included any content that was not freely available, even when most of the content was free and public (e.g., *ICPSR*). The *free* tag was used for both open resources, as well as those that required the creation of a free account to download data. The twenty most common paid and hybrid resources were then extracted (see Table 3).

Results

At least one guide page with *data* or *statistic* in the title of a tab or the guide was found for all 123 institutions in the study. A mean of 84.94 guide pages were included from each institution, with a

maximum of 336 pages from a single institution, a minimum of one from another, and a standard deviation of 64.15. Statistics reflecting the number of unique resources across institutions, unfortunately, are not accurate reflections of the number of actual data and statistical resources due to a significant amount of noise that remained present in the resource lists following cleaning and normalization. For example, many links that appeared in small quantities, such as non-data resources (e.g., *duo authentication)*, local services or spaces (e.g., *borchert map library)*, and vaguely named resources (e.g., *history* or *publications*) were not removed during cleaning and so are included in the summary statistics for resources overall. We can still make founded observations from the top results, however, since specific 'noisy' resources are too uncommon to show up in the most popular 500 results. It's harder to make claims based on the entire dataset due to this long tail of noisy results, however. Bearing that in mind, there was a mean of 791.96 unique resources found per institution, with a maximum of 3,284 and a minimum of 21 unique resources found at specific institutions. The standard deviation of unique resources per institution was 638.44, suggesting a normal range of 153.52 to 1,430.40 data and statistical resources per institution.

Table 1 lists the twenty data and statistics resources that were most common on LibGuides across R1 institutions, while an online supplementary appendix[7] provides a sortable list of the 200 most common resources by overall count. The two resources that appeared most frequently across institutions were ICPSR and data.gov, both of which were included on guides from 94.31% of institutions. ICPSR was also the resource with the most overall links (1,834) across the entire dataset. Of the top twenty data and statistics resources that were most common across institutions, fifteen were links to free and public resources, ten of which were to U.S. government websites (all following references to "top" or "common" resources in this paragraph refer to their frequency across institutions). The most common U.S. government resource, data.gov, compiles open data from a variety of government agencies. The U.S. Census Bureau was the source of three of the top ten resources, while most of the other top U.S. government sites represented agencies dedicated specifically to statistics and analysis: the National Center of Health Statistics, National Center for Education Statistics, Bureau of Labor Statistics, Bureau of Justice Statistics, and Bureau of Economic Analysis. Freely available resources among the twenty most common that were not from U.S. government sites were from the United Nations (UNData), the European Commission (Eurostat), the World Health Organization (the WHO Global Health Observatory), the World Bank (World Bank Data), and the International Monetary Fund (IMF Data). Five of the top twenty resources were for platforms on which at least some data requires a paid subscription or membership: ICPSR[8], Roper Center iPoll, Social Explorer, *Statistical Abstracts of the United States* (ProQuest), and *Historical Statistics of the United States* (Cambridge).

When sorting the same list by the overall number of links each resource had across guides, however, seven of the top ten resources were for platforms from which at least some data required paid access: ICPSR (1,834 links), ProQuest's Statistical Insight (1,532 links), Social Explorer (1,430 links), Sage's Data Planet (1,318 links), Statista (1,104 links), Roper Center iPoll (1,090), and ProQuest's *Statistical Abstracts of the United States* (1,058 links). While Statistical Insight was the second most common resource by link count (1,532), it was present on guides from only 61.79% of the institutions. Notably, all links to the eighth most common data resource across institutions, U.S. Census Bureau's American FactFinder, were dead at the time of data collection, since the site was decommissioned in March 2020 (United States Census Bureau, 2021).

Table 1: Most common data and statistics resources, sorted by percentage of institutions including them

	Resource name	% of institutions	Overall count	Example URL[9]
1	icpsr[10]	94.31%	1834	www.icpsr.umich.edu
2	data.gov	94.31%	954	www.data.gov
3	u.s. census bureau	93.50%	956	www.census.gov
4	u.s. census: data	90.24%	1145	data.census.gov/cedsci
5	undata (united nations)	88.62%	1054	data.un.org
6	nchs (u.s. national center for health statistics)	86.99%	856	www.cdc.gov/nchs
7	nces (u.s. national center for education statistics)	86.99%	546	nces.ed.gov
8	u.s. census: american factfinder	83.74%	720	factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
9	bls (u.s. bureau of labor statistics)	82.93%	535	www.bls.gov
10	roper center: ipoll	80.49%	1090	proxy2.library.illinois.edu/login?url=ropercenter.cornell.edu
11	eurostat (european commission)	80.49%	501	ec.europa.eu/eurostat
12	social explorer	78.86%	1430	www.socialexplorer.com
13	bjs (u.s. bureau of justice statistics)	78.86%	437	www.bjs.gov
14	statistical abstracts of the united states (proquest)	78.05%	1058	www.libraries.rutgers.edu/indexes/statabus
15	who: global health observatory (world health organization)	78.05%	491	www.who.int/gho/en
16	world bank: data	76.42%	545	data.worldbank.org
17	historical statistics of the united states (cambridge)	74.80%	684	hsus.cambridge.org/HSUSWeb
18	imf data (international monetary fund)	74.80%	445	data.imf.org
19	bea (u.s. bureau of economic analysis)	74.80%	422	www.bea.gov
20	cdc: data and statistics	73.17%	361	www.cdc.gov/DataStatistics

The analysis above highlights specific webpages and services from major platforms and agency websites but does not provide as clear of a view of the overall popularity of freely available organization websites. Table 2 shows the most common URL root domains for resources in the dataset by total count, providing a broader perspective of the sources of data and statistical resources in the dataset, while excluding resources that require authentication. It is still the case that U.S. government sites predominate, representing six of the top ten root domains by count. Several websites that did not appear in the list of twenty most common resources across institutions (Table 1), however, appear here: usda.gov (1,853 results) and nih.gov (1,570 results). This makes sense as the most common full domains for each (nass.usda.gov and ncbi.nlm.nih.gov) reflect only 28.60% and 30.38%, respectively, of the domains associated with the root. In fact, the usda.gov root domain is associated with 69 different full domains, with a mean count of 26.86 per domain. This contrasts with the root domain of bls.gov, which is only associated with four full domains, with a mean count of 516.5 per domain. In other words, USDA and NIH are agencies with a wide variety of sub-agency pages and services that librarians refer to on their guides. While no single page from these agencies is highly represented, their offerings are important sources of data and statistics when considered holistically. By compiling these counts by root domain, we gain a better sense of the overall importance of parent organizations such as the U.S. Census Bureau, Centers for Disease Control and Prevention, Department of Education, as well as the World Bank, United Nations, and the World Health Organization as sources of data and statistics.

Table 2: Most common root domains (excluding proxies)

	Root domain	Count	Top full domain for root	Count of full domain (% of root domain)
1	census.gov	9,670	census.gov	7,656 (79.17%)
2	cdc.gov	6,072	cdc.gov	5,382 (88.64%)
3	umich.edu	3,691	icpsr.umich.edu	2,222 (60.20%)
4	ed.gov	3,122	nces.ed.gov	2,546 (81.55%)
5	worldbank.org	2,705	data.worldbank.org	1,120 (41.40%)
6	un.org	2,636	unstats.un.org	1,076 (40.82%)
7	bls.gov	2,066	bls.gov	1,928 (93.32%)
8	usda.gov	1,853	nass.usda.gov	530 (28.60%)
9	nih.gov	1,570	ncbi.nlm.nih.gov	477 (30.38%)
10	who.int	1,411	who.int	1,202 (85.19%)

Table 3 presents a subset of the twenty most common non-free resources, sorted by the number of times they were found across all guides. Non-free resources here include databases that are only

accessible to institutions with subscriptions, as well as 'hybrid' platforms that require payment for some subset of data access, though some hybrid platforms provide significant amounts of freely available data. Overall, 20.28% of the 500 most common resources (by overall count) were either hybrid or fully paid resources, while 79.72% of the resources were entirely free (see Table 4). The paid and hybrid resources represent a diverse group of platforms in terms of publishers or vendors: while two ProQuest databases (Statistical Insight and *Statistical Abstract for the United States*) appear in the top twenty across institutions, no other single data provider shows up more than once on the list of top paid/hybrid resources. The most common paid or hybrid resources by overall count, as noted previously, are ICPSR, Statistical Insight, Social Explorer, Data Planet (Sage), and Statista. A few databases that primarily collect literature resources (e.g., OECD iLibrary and EBSCO's Business Source Complete) also appear here. Paid resources, taken overall, have a significantly higher number of links across guides than free resources do. The mean number of links for a paid resource was 174.43, compared to 104.18 for a free one, and 318.00 for a hybrid resource (see Table 4).

Table 3: Most common non-free (paid and hybrid) resources, sorted by overall count

	Resource name	% of institutions	Overall count
1	icpsr	94.31%	1834
2	statistical insight (proquest)	61.79%	1532
3	social explorer	78.86%	1430
4	data planet (sage)	50.41%	1318
5	statista	64.23%	1104
6	roper center: ipoll	80.49%	1090
7	statistical abstract of the united states (proquest)	78.85%	1058
8	oecd ilibrary	69.11%	851
9	historical statistics of the united states (cambridge)	74.80%	684
10	simplyanalytics	44.72%	672
11	policymap	43.09%	645
12	wharton research data services (wrds)	43.90%	470
13	polling the nations	45.53%	313
14	passport (euromonitor international)	34.96%	249
15	china data online	33.33%	209
16	business source complete (ebsco)	34.15%	191
17	data citation index (web of science/clarivate)	32.52%	177
18	economist intelligence unit (eiu)	21.14%	175
19	ibisworld	43.09%	168
20	europa world plus	26.83	161

Table 4: Paid, hybrid, and free resource types, of the top 500 resources

Resource type	% of top 500 resources	N resources in top 500	Number of links to resources overall	Mean number of links per resource overall
Paid	17.47%	87	15,175	174.43
Hybrid	2.81%	14	4,452	318.0
Paid & hybrid combined	20.28%	101	19,627	194.33
Free	79.72%	397	41,361	104.18

Discussion

Comparing the inclusion of freely available data and statistical resources to those that require some level of payment for user access highlights that librarians share links to paid and hybrid resources more often than they share free resources on their guides. Across all 123 R1 institutions in this analysis, librarians included links to paid resources (with a mean of 174.33 links per resource) and hybrid resources (a mean of 318.00[11]) on their guides far more often than they included free ones (a mean of 104.18). This phenomenon is especially evident when looking at popular paid resources such as ProQuest's Statistical Insight, which was the second most common by count, with 1,532 links, despite only being available at 61.79% of institutions. It makes sense for institutions with monetary investments in data and statistical resources to make concerted efforts to inform their users of their availability. But the fact that free resources are less commonly shared suggests that they may be underrepresented on resource guides. In this case, a data librarian's incentives for sharing a particular resource (e.g., improving the library's return on investment) is less than ideally aligned with users' needs for finding data and statistics. Another contributing factor is likely the fact that LibGuides' A-Z Database lists, which make it easy to include items across guides at a particular institution, usually focus on licensed resources instead of freely available websites. It is often logistically easier for a guide editor to include and maintain links to licensed resources from the A-Z list on their guides, regardless of their research value. Electronic resource librarians often manage these "database assets" for their LibGuides instance. The central management allows a change to a database URL or name, for example, to immediately propagate across all guides at the institution. In practice that level of management rarely extends to the kinds of publicly available websites that are most common on data and statistics guides, and instead focuses on subscription resources. For this reason, public data resources are inconsistently named and described across LibGuides, making it difficult to fully account for the data and statistics resources that are common across academic libraries. One potential area for improvement would be for electronic resource librarians to add the major free data and statistics resources identified by this study to their local A-Z lists.

A potentially broader opportunity for libraries on the LibGuides platform would be to take advantage of the ability for librarians to share guides and content links, not just institutionally, but across the entire community of other LibGuides institutions. This platform attribute, in theory, could support a

network of well-defined and managed links to popular resources that librarians could quickly plug into their own guides and then forget about, with ongoing maintenance managed centrally. In practice, due to technical and usability issues related to finding and maintaining resources across LibGuides instances, however, this would likely require significant development by Springshare to be implemented successfully. A lack of organization of links to free sites both within and across institutions contributes to the widespread presence of broken links to obsolete data and statistics resources. In December of 2021, for example, 83.74% of institutions were still sharing links to the U.S. Census Bureau's American FactFinder site, a site which was decommissioned in March 2020 (United States Census Bureau, 2021). American FactFinder was the eighth most common data resource shared across the institutions analyzed, and 720 links to the resource were still present on institutional guides 20 months after the link ceased to point to a functioning website. The continued inclusion of American FactFinder on data LibGuides after its relatively high-profile decommissioning suggests that maintenance and link rot are as problematic on data and statistics guides as previous research has found for the broader LibGuides context (Ornat et al., 2021).

Several data and statistical resources were less popular than the authors would have expected, especially compared to other popular resources that seem to be of relatively little research value. Free and high-value microdata from IPUMS, who provide well-maintained granular survey and census data from around the world, were not completely absent from the top resource lists, but links to their resources were not common. There were only 334 links across all guides to the most popular IPUMS resource collected, the National Historical Geographic Information System (NHGIS), which was found at 62.60% of institutions. The main IPUMS website (ipums.org) was only available from 50.41% of the institutions, with 249 links overall. This is striking compared to the popularity of ProQuest Statistical Insight, which was linked to 1,532 times, even though it primarily functions as an index to statistical resources that aren't available directly from the ProQuest interface. It's difficult to understand the enduring popularity of this resource, given the difficulty most users would have finding the sources that are cited without the help of a librarian. This seems to reflect a broader trend in which complex numerical datasets that require more sophisticated data tools or methods to analyze are included on guides less frequently than collections of more traditional tabular statistics (e.g., Cambridge's *Historical Statistics of the United States*) or simple charts and graphs (e.g., Statista). After ICPSR and Statistical Insight, the three most common paid/hybrid resources—Social Explorer, Data Planet (Sage), and Statista—provide platforms that present data in more accessible formats via dynamic data visualizations, maps, and/or tables. While in some cases the underlying data is exportable to CSV or JSON, these platforms provide outputs such as charts and graphs that pre-digest the data in ways that make data analysis outside of the platform unnecessary. The relative popularity of Social Explorer may also stem in part from the fact that it only recently (and quietly) transitioned from robust long-term free access to a more limited-term public access via trial accounts.

Free data resources that are popular for data science applications, such as Kaggle and the UCI Machine Learning Repository, are other notable absences on academic library guides. The Google-owned Kaggle provides a repository of user-generated datasets that are frequently used in online tutorials in data analytics and statistics and are often used in data science competitions for developing accurate machine learning models. Despite this, only 47 links to Kaggle were found on these guides, appearing at only 23.57% of the institutions studied. While Kaggle is a free commercial platform, its relative

absence here might be understood to reflect the relative paucity of guides focused on computer science and data science resources, or as part of a larger pattern in which "unruly" sites are rarely included on data guides. As Huck mentions, "the 'Wild West' of data discovery encompasses author websites and sharing platforms such as GitHub and Kaggle. It is considerably harder to find useful data through these websites, because your best search tool is a regular web search" (2020). A similar lack of visibility extends to most official data repositories: while ICPSR is a significant exception, the next most common data repository was Harvard Dataverse, found at 51.22% of institutions, with 234 direct links to the repository (and an additional 244 links to datasets within the repository). This trend likely reflects the fractured landscape in which user-created datasets are stored in any number of local institutional repositories, few of which are large enough to show up across a wide range of guides. Rather than link to specific repositories, in fact, many guides include re3data.org, a directory of data repositories, which was linked to 339 times at 66.67% of institutions. While it's likely that each institution includes links to their preferred local data repository, these fall under the radar when compiled across institutional boundaries.

Finally, it's worth noting that the most common data and statistics resources identified across LibGuides in 2021 are largely accounted for in previously published directories of data resources. All of the resources coded as freely available in *Table 1*, and all of the resources except for ncbi.nlm.nih.gov in *Table 2*, are included in Bauder's 2014 reference guide as either major or minor sources. While Bauder, whose book-length guide does not include any paid resources, often identifies resources at a greater level of specificity (pointing to a specific survey from a government agency, for example), a surprisingly high percentage of the most popular free web resources found on data LibGuides in 2021 were already present in this 2014 collection. While it seems that many of the demands and services related to data librarianship have evolved significantly over the previous decade, the general landscape of free data and statistical sources appears to be fairly stable.

Limitations and future directions

Despite the common shared infrastructure across LibGuides from different institutions, naming practices for links to data and statistical resources across guides are extremely inconsistent. The fact that the URLs for licensed resources are generally obscured behind a number of proxies and single sign-on services makes it difficult to accurately account for the presence of paywalled resources using their URLs alone, while the prominence of textual links that use terms that are irrelevant to the resource at hand (e.g., "here," "2017," "%") complicates the task of counting the presence of specific resources by name. The data cleaning processes implemented here adjust for those challenges by leveraging both strings from URLs and from named links to cluster and merge the same resources together. But given the messiness of the data and the size of the dataset, the long tail of resources shared on these guides was often un- or undercounted. Essentially, any resource that did not turn up on initial lists of the most common thousand or so URLs and resource names was likely to not be normalized, and therefore fell through the cracks in the overall analysis. Further study would be required to highlight data and statistical resources that are less commonly shared but have high potential value for academic library audiences.

While iterating through the time-consuming series of data cleaning steps detailed above, it occurred to the authors how little consistency there was in how libraries name and describe resources on LibGuides. Given the rich history of consortial library collaborative projects to create shared cataloging

frameworks and controlled vocabularies it's striking that each library on the LibGuides platform must create their own database assets. For example: of the 1,834 resource-links for ICPSR found in this study, there were 131 different link names used, pointing to 220 different URLs, some of which were broken. The proliferation of broken links on guides especially signals the potential upside of a collaborative cross-institutional system for creating and sharing database assets, though the hodgepodge of institutional and paid proxy platforms complicates this prospect. It's a surprise, however, given the prevalence of Springshare's LibGuides A-Z Database tool, that Springshare does not offer a central service for managing database content across guides. There is potential for an organized effort on the part of a library collective, consortium, or professional organization to partner with Springshare to pilot some level of cooperative maintenance of database assets. The time-gains across libraries could be substantial, as would improvements in quality control regarding ongoing maintenance.

Finally, limiting the initial scope of library guides to be collected to R1 institutions provides an incomplete view of broader trends across academic libraries. One compelling direction for future research in this area would be to look more holistically across guides from different types of academic libraries and consider differences in the kinds of resources that are shared by institution size, type, location, and so forth. For example, do smaller institutions rely more on free and hybrid data resources than academic libraries at R1 institutions? Are smaller institutions more likely to take advantage of the community aspect of LibGuides and repurpose guides from institutions with more resources to devote to compiling data resources? The authors unsystematically noted geographical differences in data resources shared at different institutions (e.g., state data repositories are almost always only available on guides provided by institutions within the state), but a more purposeful look at key differences in resource sharing across LibGuides could make for a fruitful examination.

Conclusion

There were 10,448 different published LibGuides pages related to data or statistics topics at 123 R1 institutions in 2021. Librarians and other data experts clearly find LibGuides to be an essential tool for organizing and sharing these kinds of resources. While most common free resources shared on LibGuides were duplicative of resources compiled for earlier published bibliographic data guides, librarians shared links to licensed data resources more frequently on LibGuides than they shared free ones. The continued relevance of data and statistics sources noted in earlier studies suggests that resources from many major government agencies and intergovernmental organizations are, as a whole, relatively stable and could be managed more intentionally on LibGuides. A lack of consistency in the titles and links to data resources both within and across institutions, along with a preponderance of dead links to outdated sites, likewise suggests that academic libraries would have much to gain from some degree of centralized management of their data resources. Local inclusion of the most common free data and statistical resources identified in this study on LibGuides' A-Z Database lists, for example, would help ensure that those resources are more visible to users and are better maintained over time.

Along similar lines, a current over emphasis of licensed data resources on LibGuides undersells the value of free resources from government sources and points users instead to platforms geared towards a more traditional bibliographic market (e.g., ProQuest Statistical Insight). U.S. academic libraries could also better promote data formats such as IPUMS microdata, which was not found on

guides from almost half of the institutions studied. It's likely that there are still data knowledge gaps in libraries at many institutions such that IPUMS and other valuable resources fall through the cracks while formats that are easier to understand are emphasized instead, leaving the needs of researchers with higher-level data skills unmet. One relatively simple stopgap measure, which could be implemented locally, would be for "accidental data librarians" to draw more freely from guides at institutions that have more robust data knowledge and support. The somewhat shocking preponderance of dead links to American FactFinder on these guides—when news of the retirement of the site was widely shared among government documents and data librarian communities—similarly points to the benefit of under-resourced data librarians drawing from guides maintained at other institutions, where librarians may be better equipped to monitor the field of data and statistics resources.

Overall, libraries at R1 institutions showed remarkable consistency in recommending a similar core list of data resources across their guides. Data librarians at many institutions, however, would also benefit from reviewing the list compiled in Table 1 (or this more in-depth online table) to ensure that key free resources are available to their users. While local needs differ, it's hard to imagine that users at the seven institutions (5.69%) that do not link to data.gov on any of their guides would not benefit from that resource, or that users at the 33 institutions (26.83%) without links to the CDC Data & Statistics platform would not benefit from those tools. While it's not feasible for an individual librarian to keep up with an ever-expanding catalog of data and statistical resources, the compilation of common data and statistics resources collected here provides a snapshot of data resources that a network of library data peers deems fit to share and promote on their own guides.

Author statements

Hennesy: Conceptualization, methodology, data collection, cleaning, and analysis; Writing: all sections of original draft except for literature review; Co-editing: full paper.

Kubas and McBurney: Review of methodology, analysis, and data outputs; data cleaning assistance; Writing: literature review; Co-editing: full paper.

References

American Council of Education (2022) *Carnegie Classification of Institutions of Higher Education: Basic Classification Description.* Available at: https://carnegieclassifications.acenet.edu/classification_descriptions/basic.php (Accessed: 14 June 2022)

Au, R. (2020) 'Data Cleaning IS Analysis, Not Grunt Work', *Counting Stuff*. Available at: https://counting.substack.com/p/data-cleaning-is-analysis-not-grunt (Accessed: 19 January 2022).

Bauder, J. (2014) *The Reference Guide to Data Sources*. Chicago: American Library Association.

Bordelon, B. (ed.) (2009/2010) 'The Subject Content and How Researchers Use the Data', *IASSIST Quarterly*, 33(4)/34(1). dio: https://doi.org/10.29173/iq883

Boslaugh, S. (2007) *Secondary Data Sources for Public Health: A Practical Guide*. Cambridge: Cambridge University Press.

Dougherty, K. (2013a) 'The Direction of Geography LibGuides', *Journal of Map and Geography Libraries*, 9(3), pp. 259–75. DOI: https://doi.org/10.1080/15420353.2013.779355

Eclevia, M.R., Fredeluces, J.C.L.T, Maestro, R.S., and Eclevia Jr., C.L. (2019) 'What Makes a Data Librarian?: An Analysis of Job Descriptions and Specifications for Data Librarian', *Qualitative & Quantitative Methods in Libraries*, 8(3), pp. 273–290. Available at: http://qqml-journal.net/index.php/qqml/article/view/541. (Accessed: 14 June 2022)

Foster, A.K., Rinehart, A.K., and Springs, G.R. (2019) 'Piloting the Purchase of Research Data Sets as Collections: Navigating the Unknowns', *portal: Libraries & the Academy*, 19(2), pp. 315–328. DOI: https://doi.org/10.1353/pla.2019.0018

Furay, J. (2018) 'Performance Review: Online Research Guides for Theater Students', *Reference Services Review*, 46(1), pp. 91–109. DOI: https://doi.org/10.1108/RSR-09-2017-0037

Garrison, B. and Exner, N. (2018) 'Data Seeking Behavior of Economics Undergraduate Students: An Exploratory Study', *Reference & User Services Quarterly*, 58(2), pp. 103–113. DOI: http://dx.doi.org/10.5860/rusq.58.2.6930

Geraci, D., Humphrey, C., and Jacobs, J. (2012) *Data Basics: An Introductory Text*. Available at: https://3stages.org/class/2012/pdf/data_basics_2012.pdf. (Accessed: 14 June 2022)

Hennesy, C. and Adams, A.L. (2021) 'Measuring Actual Practices: A Computational Analysis of LibGuides in Academic Libraries', *Journal of Web Librarianship* 15(4), pp. 219–242. DOI: https://doi.org/10.1080/19322909.2021.1964014

Hoffman, S. (2015) 'Data Reference and Instruction in Journalism and the Social Sciences', *DttP (Documents to the People): A Quarterly Journal of Government Information Practice & Perspective*, 43(2), pp. 14–17. Available at: https://journals.ala.org/index.php/dttp/issue/viewIssue/603/360. (Accessed 14 June 2022)

Horton, J.J. (2017) 'An Analysis of Academic Library 3D Printing LibGuides', *Internet Reference Services Quarterly*, 22(2/3), pp. 123–131. DOI: http://dx.doi.org/10.1080/10875301.2017.1375059

Huck, J. (2020) 'Identifying, Accessing and Evaluating Data', *Information Outlook: The Magazine of the Special Libraries Association*, 24(1), pp. 4-6. Available at: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1000&context=sla_io_2020 (Accessed: 25 March 2022).

Jackson, R. and Stacy-Bates, K.K. (2016) 'The Enduring Landscape of Online Subject Research Guides', *Reference & User Services Quarterly*, 55(3), pp. 219-225. DOI: https://doi.org/10.5860/rusq.55n3.219

Johnson, C.V. and Johnson, S.Y. (2017) 'An Analysis of Physician Assistant LibGuides: A Tool for Collection Development', *Medical Reference Services Quarterly*, 36(4), pp. 323–333. DOI: https://doi.org/10.1080/02763869.2017.1369241

Johnson, E.O. (2019) *Working as a Data Librarian: A Practical Guide*. Denver: Libraries Unlimited.

Joo, S. and Schmidt, G.M. (2021) 'Research Data Services from the Perspective of Academic Librarians', *Digital Library Perspectives*, 37(3), pp. 242–256. DOI: https://doi.org/10.1108/DLP-10-2020-0106

Kalinowski, A. and Hines, T. (2020) 'Eight Things to Know about Business Research Data', *Journal of Business & Finance Librarianship*, 25(3/4), pp. 105–122. DOI: https://doi.org/10.1080/08963568.2020.1847548

Kellam, L.M. and Peter, K. (2011) *Numeric Data Services and Sources for the General Reference Librarian*. Cambridge: Chandos.

Kellam, L.M. and Thompson, K. (eds) (2016) *Databrarianship: The Academic Data Librarian in Theory and Practice*. Chicago: Association of College and Research Libraries.

McCormick, A. (2020) 'Collection Development for Librarians in a Hurry: A Survey of the Physics Resources of the Libraries of the Association of American Universities', *Issues in Science and Technology Librarianship* 96. DOI: https://doi.org/10.29173/istl68

McNulty, T. (2013) *Art Market Research: A Guide to Methods and Sources,* 2nd edn. Jefferson, North Carolina: McFarland.

Nelson, M.R.S. (2020) 'Adding Data Literacy Skills to Your Toolkit', *Information Outlook: The Magazine of the Special Libraries Association*, 24(1), pp. 10-11. Available at: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1000&context=sla_io_2020 (Accessed: 25 March 2022).

Neuhaus, C., Cox, A., Gruber, A.M., Kelly, J., Koh, H., Bowling, C. and Bunz, G. (2021) 'Ubiquitous LibGuides: Variations in Presence, Production, Application, and Convention', *Journal of Web Librarianship*, 15(3), pp. 107–127. DOI: https://doi.org/10.1080/19322909.2021.1946457

Ohaji, I.K., Chawner, B. and Yoong, P. (2019) 'The Role of a Data Librarian in Academic and Research Libraries', *Information Research*, 24(4). Available at: http://informationr.net/ir/24-4/paper844.html. (Accessed: 23 March 2022)

Ornat, N., Auten, B., Manceaux, R. and Tingelstad, C. (2021) 'Ain't no party like a LibGuides Party: 'cause a LibGuides Party is mandatory', *College & Research Libraries News*, 82(1), pp. 14–17. DOI: https://doi.org/10.5860/crln.82.1.14

Osorio, N. (2014) 'Content analysis of Engineering LibGuides', in *2014 ASEE Annual Conference & Exposition Proceedings*. Indianapolis: ASEE, pp. 24.318.1-24.318.23. DOI: https://doi.org/10.18260/1-2--20209

Reback, J. et al. (2022) *Pandas (1.4.0rc0)*. Computer software. https://zenodo.org/record/5824773. (Accessed: 20 January 2022).

Reitz, K. (2015). *Requests (2.7.0)*. Computer software. Available at: https://pypi.org/project/requests/2.7.0/ (Accessed: 14 June 2022)

Rice, R. and Southall, J. (2016) *The data librarian's handbook*. Chicago: American Library Association.

Richardson, L. (2015). *Beautiful Soup* (4.8.1). Computer software. Available at: https://beautiful-soup-4.readthedocs.io/en/latest/ (Accessed: 14 June 2022)

Selenium (4.2.0). Computer software. Available at: https://pypi.org/project/selenium/. (Accessed: 14 June 2022)

Semeler, A.R., Pinto, A.L. and Rozados, H.B.F. (2019) 'Data science in data librarianship: Core competencies of a data librarian', *Journal of Librarianship and Information Science*, 51(3), pp. 771–780. DOI: https://doi.org/10.1177/0961000617742465

Smith, E. (2008) *Using Secondary Data In Educational And Social Research*. New York: Open University Press.

Springshare (2022). *LibGuides community*. Available at: https://community.libguides.com/ (Accessed 14 June 2022)

Stankus, T. and Parker, M.A. (2012) 'The Anatomy of Nursing LibGuides', *Science & Technology Libraries*, 31(2), pp. 242–255. DOI: https://doi.org/10.1080/0194262X.2012.678222

United States Census Bureau (2021). *Transition from AFF*. Available at: https://www.census.gov/data/what-is-data-census-gov/guidance-for-data-users/transition-from-aff.html (Accessed: 15 January 2022)

Van Dyk, G. (2015) 'Finding Religion: An Analysis of Theology LibGuides', *Theological Librarianship*, 8(2), pp. 37–45. DOI: https://doi.org/10.31046/tl.v8i2.384

Wheatley, A., Chandler, M. and McKinnon, D. (2020) 'Collaborating with Faculty on Data Awareness: A Case Study', *Journal of Business & Finance Librarianship*, 25(3/4), pp. 281–290. DOI: https://doi.org/10.1080/08963568.2020.1847553

Endnotes

[1] Cody Hennesy (chennesy@umn.edu) is the Journalism & Digital Media Librarian at the University of Minnesota, Twin Cities.
[2] Alicia Kubas (akubas@gpo.gov) is a librarian at the US Government Publishing Office.
[3] Jenny McBurney (jmcburne@umn.edu) is a Social Sciences Librarian at the University of Minnesota, Twin Cities.
[4] A list of the institutions included are available, along with the full replication data for this study, in the Data Repository for the University of Minnesota: https://conservancy.umn.edu/handle/11299/228216.
[5] These forms of the terms were chosen, in part, because the LibGuides search interface returned guides with matches such as *statistics* and *statistical* for the keyword *statistic*, and terms such as *datasets* for the keyword *data*.
[6] Python code used for data analysis and the generation of tables is available at https://github.com/chennesy/lg_data/
[7] https://chennesy.github.io/lg_data/
[8] While most datasets on ICPSR are freely available, some data from specific datasets (e.g., the American National Election Study, and the U.S. Transgender Survey) are only available to users at institutions with paid memberships. For this reason, the authors tagged platforms such as ICPSR and Statista as 'hybrid'and tended to group them together with 'paid' resources instead of grouping them with freely available sites such as census.gov.
[9] The Example URL column in Table 1 refers to the most common URL associated with a particular resource in the dataset. Any number of URLs could be associated with a particular resource. Most

URLs for Roper Center iPoll, for example, are proxied at an institutional level so that the most common URL is specific to the institution that linked to iPoll the most often.

[10] Normalized (uncapitalized) resource names are retained in the tables to enable clearer references to resources in the original dataset.

[11] The mean number of links per hybrid resource skews high due to the high number of links to the top three resources (ICPSR, Statista, and OECD iLibrary) and the relatively few hybrid resources overall. The median number of links for hybrid resources was 67.50, compared to 66.00 for paid, and 53.00 for free.