More Than a Pretty Face: The Louisiana Digital Library as a Data Hub

Scott Ziegler and Cara Key, Louisiana State University

Abstract

The Louisiana Digital Library (LDL) is an online platform for libraries, museums, archives, and historical organizations across the state. The books, manuscripts, oral histories, maps, and photographs held in the LDL showcase the cultural resources of Louisiana. The interface enables users to discover materials through search and browse, and to view and interact with these materials. The metadata about these items is also a great asset. When explored in their entirety, the data held in the LDL is as valuable as the digital facsimiles. The LDL may be conceptualized as a data hub, a place to gather and share the metadata of the participating institutions. Open data is a growing trend in archives and special collections, enabling new types of interactions with collection material. Exposing the data held in digital libraries, in ways that extend beyond traditional digital library discovery and access, opens pathways for researchers to investigate complex questions. This paper contextualizes the field of open data in historical institutions, and explores uses for downloaded metadata from the LDL.

Keywords: digital libraries, digital collections, digital repositories, digital archives, collections as data, open data, data hubs, data analysis, data visualization, metadata, digital scholarship, historical data

The Louisiana Digital Library (LDL) is an online library that brings together photographs, maps, letters, diaries, and other material from a growing number of Louisiana-based cultural institutions. Taken together, these items represent an important aspect of Louisiana history and culture, and the LDL is a central platform through which access to these items is provided. Beyond being a platform for viewing digital facsimiles, however, the LDL also functions as a data hub -- a location for the collection and dissemination of data about the collections. By emphasizing the use and reuse of data, as well as the digital facsimiles, the LDL follows a growing trend in the digital library and research data management fields.

Literature Review: Cultural Institutions, Digital Repositories, and Data

A growing interest in data held within cultural institutions has precipitated increased attention from librarians, in the form of grants, workshops, and literature. In both 2016 ("Collections as Data," 2016) and 2017 ("Collections as Data Impact," 2017), the Library of Congress hosted conferences on the use of library collections as data. Also in 2016, the Institute of Museums and Library Services (IMLS) funded a two-year project, Always Already Computational - Collections as Data, "to foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research" ("Always Already Computational," n.d.). In addition to workshops and meetings, the Always Already Computational (AAC) project compiles information on like-minded projects, and has released "The Santa Barbara Statement on Collections as Data," a document of guiding principles for treating collections as data ("The Santa Barbara Statement," n.d.). In 2018 the project was funded by the Mellon Foundation to explore sustainable models for using

collections as data ("Part to Whole," n.d.).

"The Santa Barbara Statement on Collections as Data" (n.d.) compares the decades-old practice of building digital collections in cultural institutions to the needs of researchers working with data. These researchers "have drawn upon computational means to ask questions and look for patterns." Such work "goes under a wide variety of names including but not limited to text mining, data visualization, mapping, image analysis, audio analysis, and network analysis" ("The Santa Barbara Statement," n.d.). The digital collections in our libraries, archives, and museums, however, are generally ill-suited to allow researchers to perform data-intensive work on our collections. For example, many of our systems are built to encourage searching for specific words, browsing through specific collections, and filtering for specific subjects. These are all meaningful ways to navigate digital collections. However, none of these allow a faculty member, say, to mine all text from a student newspaper, or a data journalist to build visualizations using historic photographs of floods.[1]

The work of the AAC project and similar initiatives is starting to bear fruit within both the practice and the literature of librarianship. Data-oriented topics related to cultural institutions are appearing in syllabi, library strategic goals (Lied Library, 2018), and position papers (Weber, 2017). Within this broad field, there is also growing interest in treating digital collections as data, and using this data outside of the specific web environment of a digital repository. Many digital library practitioners are beginning to understand that there are many questions that can be asked of the data held within digital repositories. These questions are often not anticipated by the repository's interface. As

[1] These examples are based, in part, on the growing Collections as Data Personas. https://collectionsasdata.github.io/personas/

such, these questions become difficult to ask, and even harder to answer.

A recent conference panel at Open Repositories 2018 explored a number of approaches to offering bulk cultural heritage data through repositories (Bailey, Lynch, & Jordan, 2018). The panel participants represented efforts along a spectrum of options -- from offering application program interfaces (APIs)[2] for accessing and downloading content from digital repositories, to building systems specifically for the purpose of working with data in different environments. All panel participants "offered insight into how repositories are being adapted and extended to provision access to collections as data […]," write Sarah Potvin and Hannah Frost in their report from the conference (Potvin & Frost, 2018).

Recent initiatives from the British Library exemplify this move to provide access to collections as data. Starting in 2017 the British Library began offering packaged datasets for download. "Our vision for the British Library," states a promotional post related to the increased attention to data, "is that research data are as integrated into our collections, research and services as text is today." ("Announcing the new British Library research data strategy," 2017). "As part of its work to open its data to wider use," the British Library (n.d.) offers these datasets "for research and creative purposes." The datasets are downloadable, and include information such as "data format (images, full text, metadata, etc), licences, temporal and geographic scope, originating purpose (e.g. specific digitisation projects or exhibitions) and collection, and related subjects or

[2] An application programming interface (API) is a set of standards that allow for the sharing of data between different systems. In digital librarianship, APIs often serve as a means of ensuring that metadata created in one system can be shared with other systems in a way that doesn't require extensive human interaction.

themes" (British Library, n.d.). The British Library has long had a digital repository for finding and viewing collections; the datasets, however, offer its data for new uses.

Another example is the University of Pennsylvania's OPenn. Designed from the beginning to accommodate use of its data outside a specific interface, OPenn is a repository that "contains complete sets of high-resolution archival images of cultural heritage material," ("OPenn," n.d.) and encourages users to bulk download content. Unlike other digital repositories, OPenn lacks a user interface for searching and browsing, and is "intended for aggregators, digital humanists, and scholars who have been directed here to procure high-resolution images of manuscript pages" ("OPenn: ReadMe," n.d.).

OPenn and the British Library dataset initiative are examples of how many digital library developers and administrators are responding to a growing understanding that researchers need large sets of data. Researchers need to ask many questions of our data; the interfaces we build cannot anticipate these questions. Therefore, offering researchers direct access to data enables a powerful means of promoting the use of our collections.

Louisiana Digital Library

The Louisiana Digital Library (LDL) falls closer to the British Library on this spectrum. The LDL is built to provide a platform for the discovery and access of digital facsimiles. The first digital collections in what is now known as the LDL were published online in 1993. In 2003, the "LOUISiana Digital Library" was launched on CONTENTdm under the administration of LOUIS, the Louisiana Library Consortium. In 2014, the Louisiana Digital Consortium formed to provide guidance to the continued growth of the LDL. Starting in 2015, a team of developers and librarians at Louisiana

State University migrated the majority of LDL collections from CONTENTdm to a locally customized instance of Islandora. Islandora is an open source digital repository software platform that combines the Fedora Commons digital asset management architecture with the Drupal content-management framework. The redesigned site was launched on its new platform in August of 2017.

As of the time of this writing, there are 20 institutions with collections published in the production site, and another several slated to publish in the near future. The membership is quickly diversifying in type of institution. While academic libraries and state institutions make up the majority of LDL members, there is a growing number of public libraries and independent historic institutions. Each member organization has control over the description of their items, with the ability to edit, add to, and delete items in their collections. The result of the varied membership is a robust collection of items from across the state representing the holdings of a wide variety of institutions. This growing pool of digital material is supported by and presented through a dynamic and innovative interface.

LDL as Data Hub

Currently the LDL is primarily used as a means of accessing digital facsimiles of photographs, letters, diaries, maps, oral histories and other historic material through the website's interface. As a data hub, however, the LDL functions as a means to explore aggregated metadata. This data can allow researchers to examine materials in different ways -- by creating new visualizations, for example, and by asking questions about how people and places are represented in a state digital library. This section will introduce the idea of a data hub, and then will demonstrate a number of examples of the use of the LDL

as a data hub.

Data Hubs

Data hubs are central locations where data from various sources are stored and shared. In various industries that deal with big data, data hubs have a more technical definition, distinguishing between different means of uploading data and storage options.[3] Other uses of the term focus less on the technical structure of the hub, and more on the function. In this sense, a data hub is a place that individuals and institutions can store data, and a place from which individuals and institutions can retrieve data.[4] For the purposes of this paper, we will use data hubs in this less technical sense. When we explore the LDL as a data hub, we are investigating the potential of the LDL as a collective repository for data from multiple institutions, where that data could be accessed by users for a range of scholarly pursuits.

Currently, the dominant means of using the data in the LDL is through the interface. Through search and browse mechanisms, users access digital facsimiles of historic items. The LDL interface is configured to promote search and discovery, and has been recognized for both its design and functionality.[5] However, the interface is not built to accommodate all research needs, and there are times that working with the data outside

[3] See, for example, "Data Lakes, Data Hubs, Federation: Which One Is Best?" MarkLogic, [Blog] July 8, 2016. https://www.marklogic.com/blog/data-lakes-data-hubs-federation-one-best/.

[4] Explicating this use of data hub, Adam Kariv and Rufus Pollock, founders of datahub.io, write that they hope to build a "home for people passionate about data like us. A place to discover and share high quality datasets, to connect with others and to share knowledge." Kariv, Adam, and Pollock, Rufus. "About DataHub." DataHub. Accessed July 17, 2018. https://datahub.io/docs/about.

[5] The Louisiana Digital Consortium and LSU Libraries received the LOUIS Certificate of Award for "Best Customized System," for work on Louisiana Digital Library in Islandora. Presented October 2017 at LOUIS Users Conference, Baton Rouge, LA. For kind words from the Islandora Foundation, "Islandora 200 | Islandora Website." Accessed July 17, 2018. https://islandora.ca/content/islandora-200.

of the repository using other tools and systems would be preferable. In these scenarios, thinking of the LDL as a data hub enables different sets of questions and different types of answers.

Example: Collection Statistics

In summer 2018, the LDL released a custom-built means of counting and extracting data on the types of items held in the LDL. Congregated on one page are counts for every type of material in the LDL, separated into different sections. One section contains the counts for all items across the digital library, including newspapers, books, images, and videos. Another section contains specifics for each participating institution, with counts for the items that institution has contributed (See Figure 1).

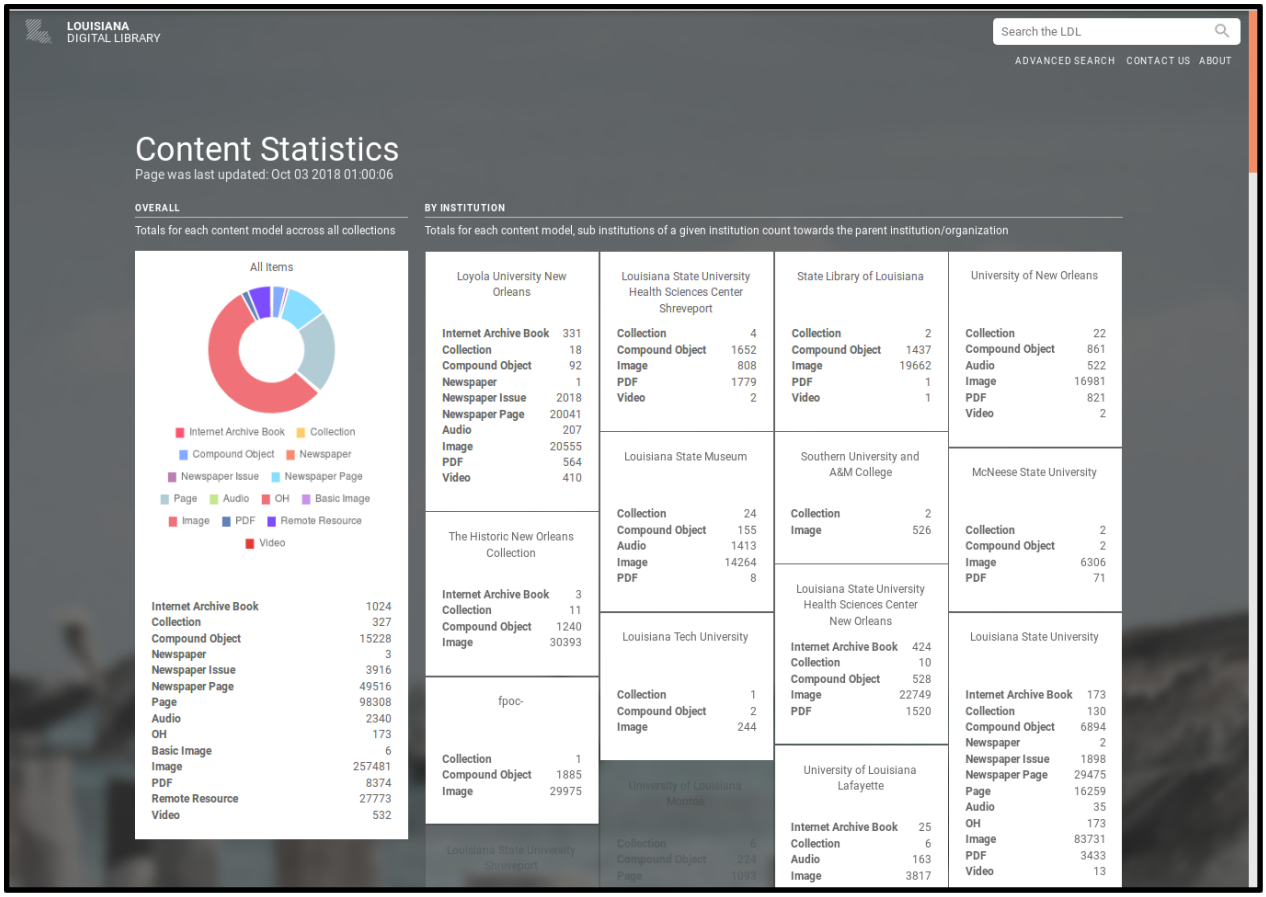


Figure 1: Screenshot of data page. On the left is the count of all institutions, on the right are panels with details for specific institutions. Content statistics are available at [*http://louisianadigitallibrary.org/data*](http://louisianadigitallibrary.org/data)

In addition to these snapshots, at the LDL- and institutional-level, the collection statistics page also offers a means for filtering, sorting, and exporting the data. For example, a user can filter to see the types of items contributed by any one institution, and then export a CSV file[6] containing this data (See Figure 2).

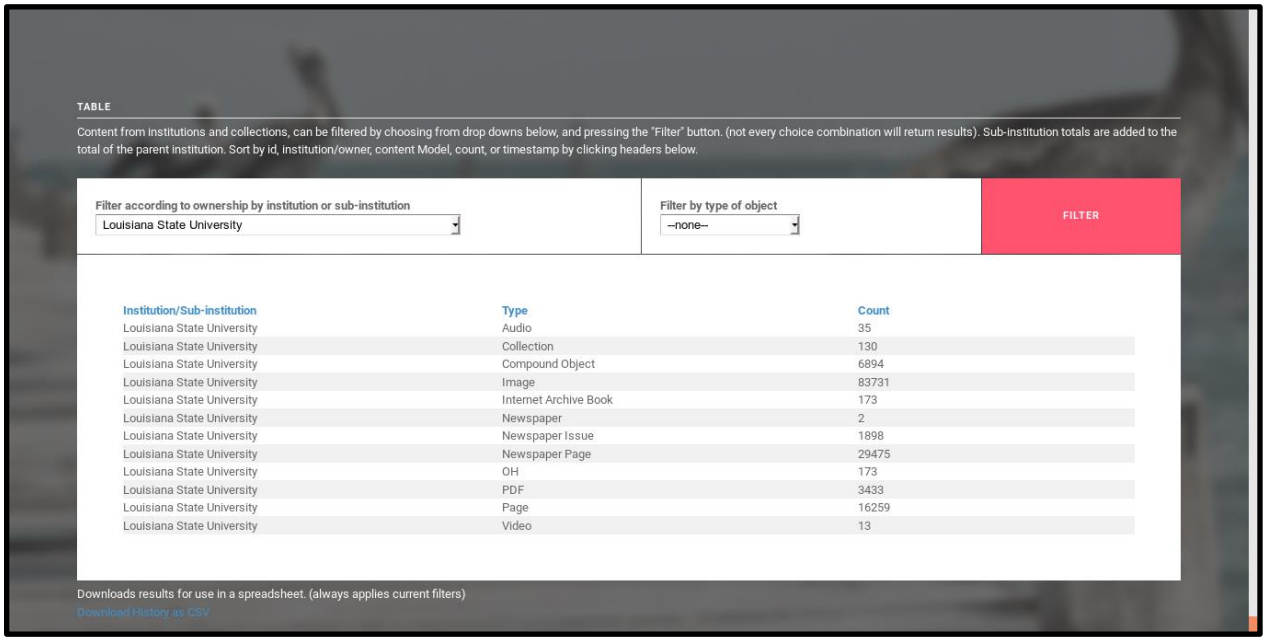


Figure 2: Screenshot of data page. This table view allows filtering by institution and type of object. All data is downloadable as CSV file. Content statistics are available at http://louisianadigitallibrary.org/data

The collections statistics page represents a significant departure from the prominent means of interacting with content in the LDL. From this snapshot of current numbers, users are able to get an across-the-board glimpse into what types of content are in the LDL and which institutions have what types of material in their collections. This type of meta-view of the repository presents the LDL content itself as an object of

[6] A comma-separated values (CSV) file consists of tabular data stored as text, which can be presented and edited as a spreadsheet or database using popular software applications. Strictly speaking, data fields in a CSV file are delimited by commas. However, there is some degree of flexibility in using the term - and the .csv file extension - for data delimited by another character such as a tab, especially when the data values may contain commas.

inquiry. Beyond treating the items in the collections -- digital facsimiles of letters, journals, etc. -- as the only objects to be investigated, by focusing on the types and numbers of items, this aspect is also raised to the level of observable object.

Example: Search Returns as CSV

Another means of extracting data from the LDL is through the option to download search returns. This feature allows all search returns for any search a user makes to be downloaded as a CSV file. Unlike the collection statistics page, the search returns download option comes standard with the Islandora software used for the LDL. Also unlike the collections statistics page, this approach focuses on descriptive metadata. As of the time of this writing, downloadable search returns data include title, subjects, date, and abstract (Figure 3).

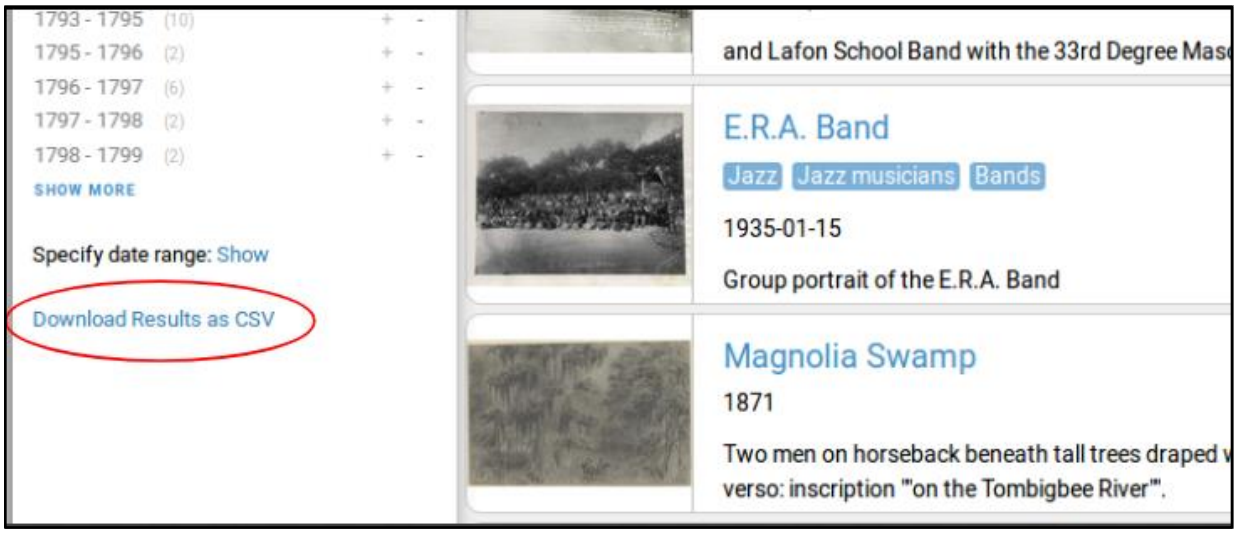
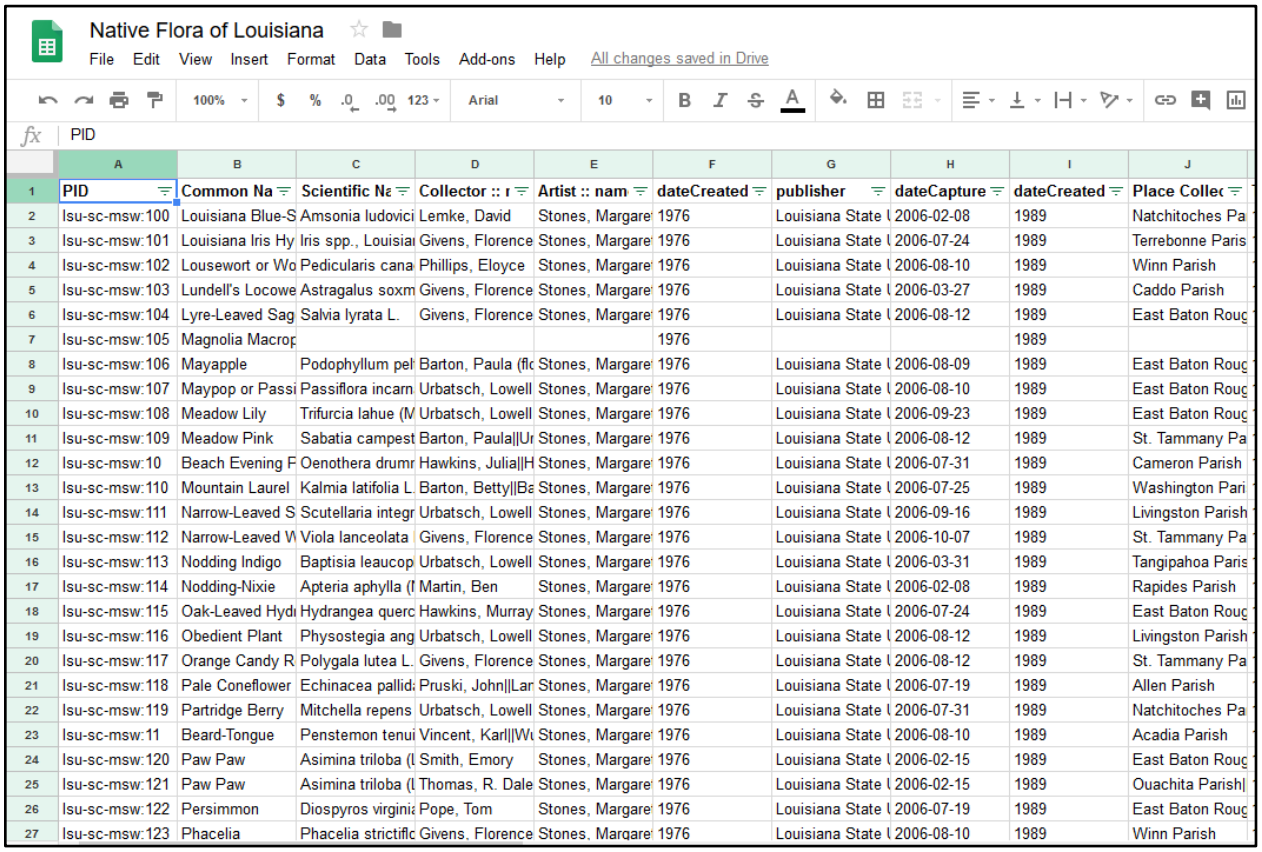


Figure 3: Screenshot of search return page with download option. Downloads include all data included in the search return, including title, subjects, dates, and abstract.

Example: Metadata Export at the Collection-Level

Another tool built by the LDL development team will soon allow users to download the entirety of a collection's metadata as a single CSV file. The CSV file can be opened by a standard spreadsheet application and utilized for sharing, editing, and analysis, without requiring technical knowledge of the metadata schema or XML encoding in which the metadata is natively stored. Unlike the search returns download above, which exports a small number of system-specified fields, this option includes all of the descriptive metadata present in the item records (Figure 4).



Native Flora of Louisiana

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

fx PID

	A	B	C	D	E	F	G	H	I	J
1	PID	Common Na	Scientific Na	Collector :: r	Artist :: nam	dateCreated	publisher	dateCapture	dateCreated	Place Collec
2	lsu-sc-msw:100	Louisiana Blue-S	Amsonia ludovici	Lemke, David	Stones, Margare	1976	Louisiana State U	2006-02-08	1989	Natchitoches Pa
3	lsu-sc-msw:101	Louisiana Iris Hy	Iris spp., Louisia	Givens, Florence	Stones, Margare	1976	Louisiana State U	2006-07-24	1989	Terrebonne Paris
4	lsu-sc-msw:102	Lousewort or Wo	Pedicularis cana	Phillips, Eloyce	Stones, Margare	1976	Louisiana State U	2006-08-10	1989	Winn Parish
5	lsu-sc-msw:103	Lundell's Locowe	Astragalus soxm	Givens, Florence	Stones, Margare	1976	Louisiana State U	2006-03-27	1989	Caddo Parish
6	lsu-sc-msw:104	Lyre-Leaved Sag	Salvia lyrata L.	Givens, Florence	Stones, Margare	1976	Louisiana State U	2006-08-12	1989	East Baton Roug
7	lsu-sc-msw:105	Magnolia Macrop				1976			1989	
8	lsu-sc-msw:106	Mayapple	Podophyllum pel	Barton, Paula (fc	Stones, Margare	1976	Louisiana State U	2006-08-09	1989	East Baton Roug
9	lsu-sc-msw:107	Maypop or Passi	Passiflora incarn	Urbatsch, Lowell	Stones, Margare	1976	Louisiana State U	2006-08-10	1989	East Baton Roug
10	lsu-sc-msw:108	Meadow Lily	Trifurcia lahue (M	Urbatsch, Lowell	Stones, Margare	1976	Louisiana State U	2006-09-23	1989	East Baton Roug
11	lsu-sc-msw:109	Meadow Pink	Sabatia campest	Barton, Paula||Ur	Stones, Margare	1976	Louisiana State U	2006-08-12	1989	St. Tammany Pa
12	lsu-sc-msw:10	Beach Evening F	Oenothera drumr	Hawkins, Julia||H	Stones, Margare	1976	Louisiana State U	2006-07-31	1989	Cameron Parish
13	lsu-sc-msw:110	Mountain Laurel	Kalmia latifolia L	Barton, Betty||Ba	Stones, Margare	1976	Louisiana State U	2006-07-25	1989	Washington Pari
14	lsu-sc-msw:111	Narrow-Leaved S	Scutellaria integr	Urbatsch, Lowell	Stones, Margare	1976	Louisiana State U	2006-09-16	1989	Livingston Parish
15	lsu-sc-msw:112	Narrow-Leaved W	Viola lanceolata	Givens, Florence	Stones, Margare	1976	Louisiana State U	2006-10-07	1989	St. Tammany Pa
16	lsu-sc-msw:113	Nodding Indigo	Baptisia leaucop	Urbatsch, Lowell	Stones, Margare	1976	Louisiana State U	2006-03-31	1989	Tangipahoa Paris
17	lsu-sc-msw:114	Nodding-Nixie	Apteria aphylla (I	Martin, Ben	Stones, Margare	1976	Louisiana State U	2006-02-08	1989	Rapides Parish
18	lsu-sc-msw:115	Oak-Leaved Hydi	Hydrangea querc	Hawkins, Murray	Stones, Margare	1976	Louisiana State U	2006-07-24	1989	East Baton Roug
19	lsu-sc-msw:116	Obedient Plant	Physostegia ang	Urbatsch, Lowell	Stones, Margare	1976	Louisiana State U	2006-08-12	1989	Livingston Parish
20	lsu-sc-msw:117	Orange Candy R	Polygala lutea L.	Givens, Florence	Stones, Margare	1976	Louisiana State U	2006-08-12	1989	St. Tammany Pa
21	lsu-sc-msw:118	Pale Coneflower	Echinacea pallid	Pruski, John||Lar	Stones, Margare	1976	Louisiana State U	2006-07-19	1989	Allen Parish
22	lsu-sc-msw:119	Partridge Berry	Mitchella repens	Urbatsch, Lowell	Stones, Margare	1976	Louisiana State U	2006-07-31	1989	Natchitoches Pa
23	lsu-sc-msw:11	Beard-Tongue	Penstemon tenui	Vincent, Karl||W	Stones, Margare	1976	Louisiana State U	2006-08-10	1989	Acadia Parish
24	lsu-sc-msw:120	Paw Paw	Asimina triloba (I	Smith, Emory	Stones, Margare	1976	Louisiana State U	2006-02-15	1989	East Baton Roug
25	lsu-sc-msw:121	Paw Paw	Asimina triloba (I	Thomas, R. Dale	Stones, Margare	1976	Louisiana State U	2006-02-15	1989	Ouachita Parish|
26	lsu-sc-msw:122	Persimmon	Diospyros virginia	Pope, Tom	Stones, Margare	1976	Louisiana State U	2006-07-19	1989	East Baton Roug
27	lsu-sc-msw:123	Phacelia	Phacelia strictiflc	Givens, Florence	Stones, Margare	1976	Louisiana State U	2006-08-10	1989	Winn Parish

Figure 4: Spreadsheet containing all descriptive metadata for the collection Native Flora of Louisiana- Watercolor drawings by Margaret Stones.

The object of analysis in this case is the aggregation of the items in a given collection, or a subset that can be culled by the user. The CSV download allows a researcher to investigate collection items along one or more dimensions of interest, and to

do so using a widely prevalent format for which many data analysis tools exist.

Working with Exported Data

This section explores a number of modest visualizations built using data exported from the LDL. These examples are offered as concrete examples of specific questions that can be answered from bulk data, but which would be difficult to address using only standard access through the interface. The visualizations were each constructed using the LDL collection *Native Flora of Louisiana- Watercolor drawings by Margaret Stones*[7]. The metadata for this collection represents a rich, multi-dimensional dataset. In addition to standard bibliographic descriptive information such as title, author, publisher, and format, metadata fields in the collection also include the dates and locations where the plant specimens were collected as well as the biological nomenclature for the specimens -- the Latin family, genus, and species names, along with their common or English equivalents. Many LDL collections feature unique, "nonstandard" descriptive information, as the *Native Flora of Louisiana* collection does. Broadly speaking, collections that feature more enhanced metadata records illuminating the informational content of the objects or some aspect of their history or provenance, as in this example, can be more easily leveraged as data.

With the included location data, we can analyze the flora by the location gathered. Using a Google base map, we add points in the Parishes where the specimens were collected. The sidebar enables users to select specific plants by their common name to focus their inquiry. In this example, we are also grouping the plants the date they were collected (See Figure 5).

[7] louisianadigitallibrary.org/islandora/object/lsu-sc-msw:collection

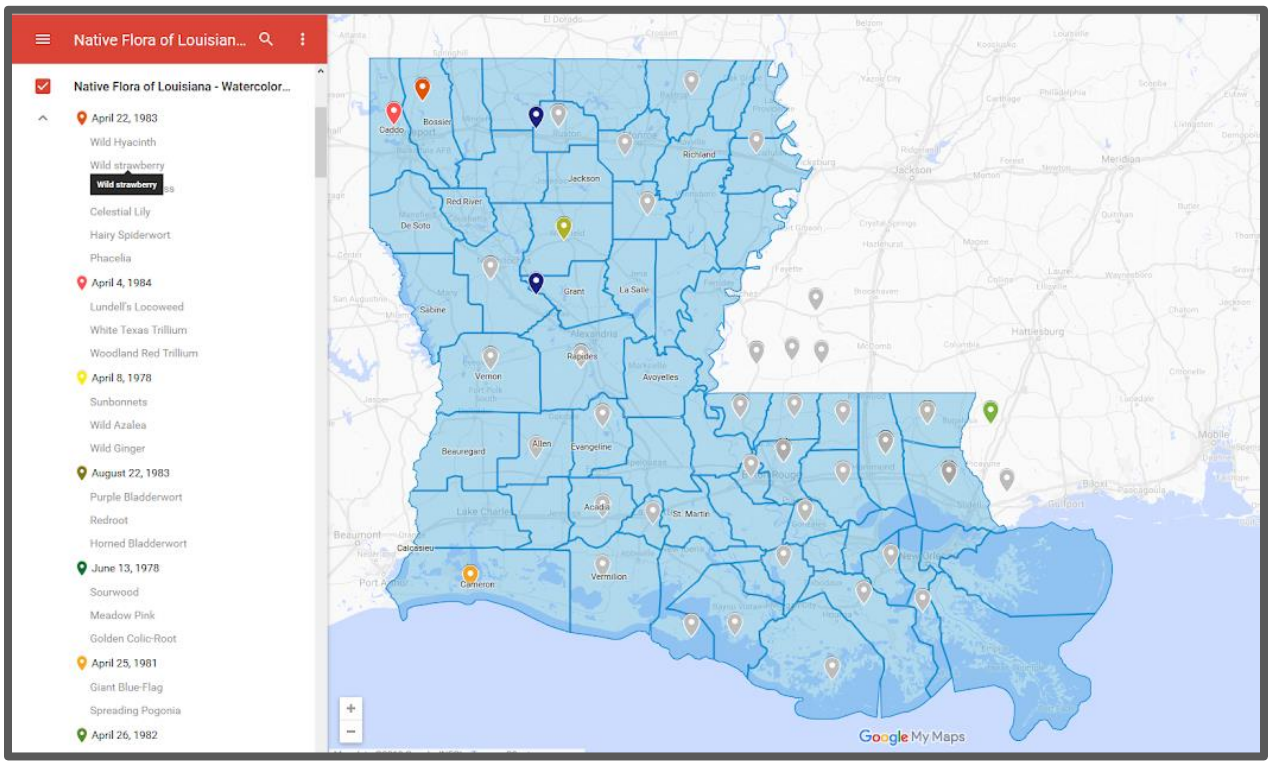


Figure 5: Map created to show location data in the collection Native Flora of Louisiana-Watercolor drawings by Margaret Stones. Map available: http://bit.ly/NativeFloraMap

We can also ask questions about seasonality (Figure 6). Different flora samples were collected at different times of the year. An interested researcher might, for example, be interested in exploring the collecting time and comparing this to the known growing season of the particular sample. Additionally, the seasonality of plant collection might become a growing area of interest for climate science and other researchers looking into effects of climate change on plant species.

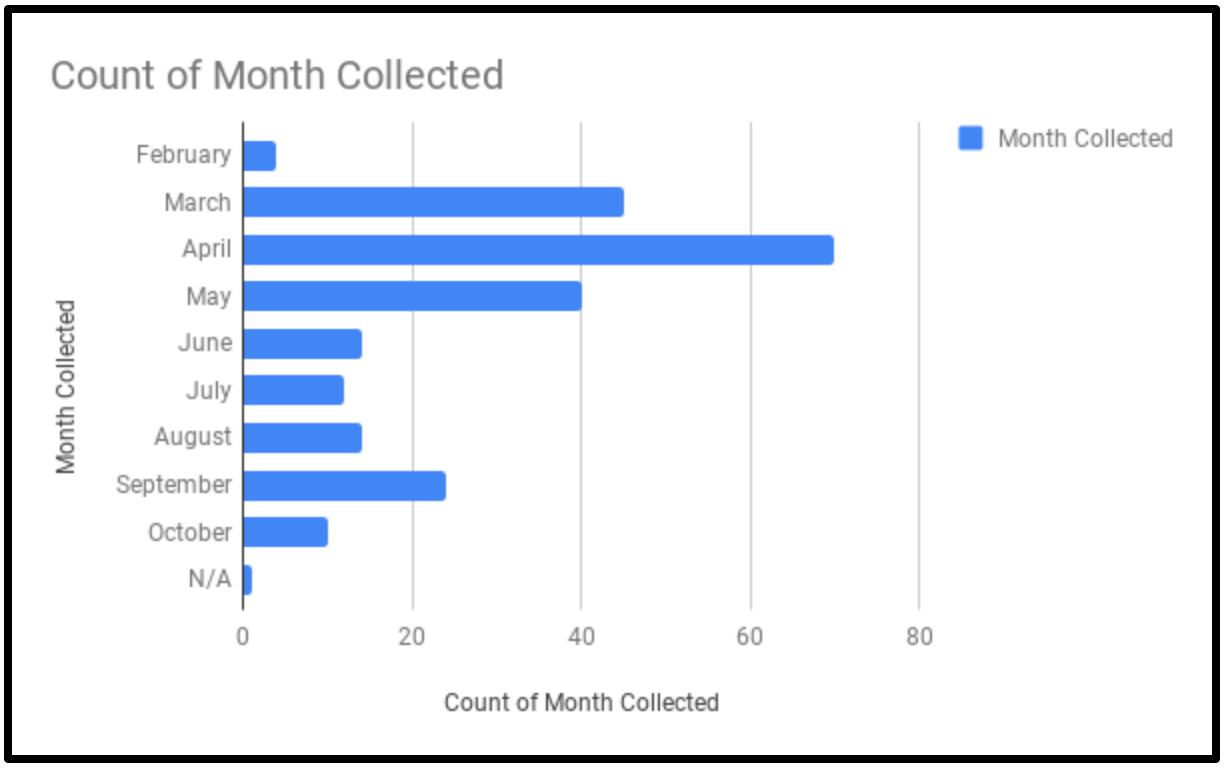


Figure 6: Chart showing number of times each month is listed as being the month a plant is collected. Chart, and associated data, available at http://bit.ly/NativeFloraByMonth

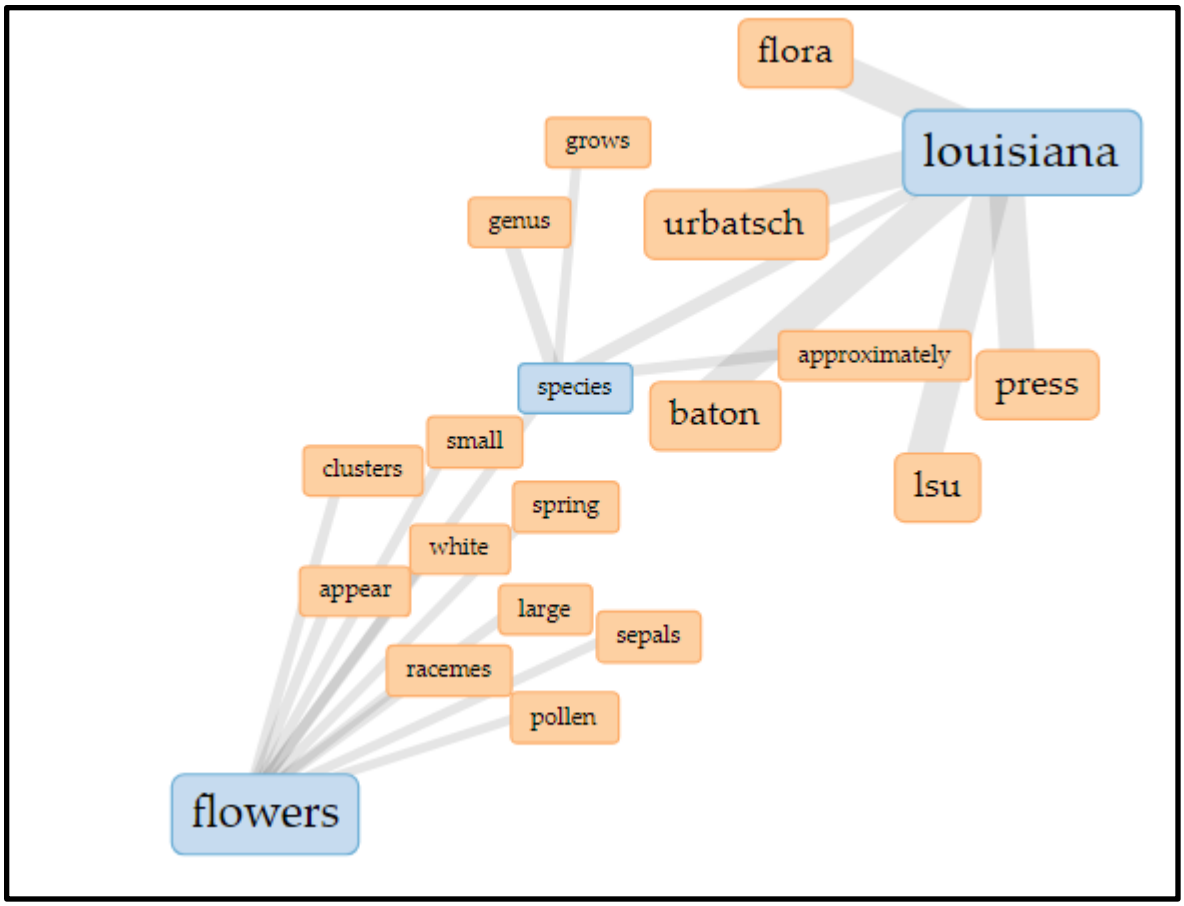


Figure 7: Collocates graph representing keywords and terms appearing in the description field of the collection. The keywords -- flowers, species, Louisiana -- appear in blue, and words that appear in close proximity appear in orange.

Text analysis of the descriptions can show connections between keywords (Figure 7). Treated as a text like any other, these descriptions become data about which we can ask specific questions.

Another network graph example (Figure 8) groups the botanical families within a particular order, and uses the item's image file, downloaded from the LDL, to represent the plant species nodes in the graph. Users of this visualization are able to click the nodes to view additional information drawn from the descriptive metadata.

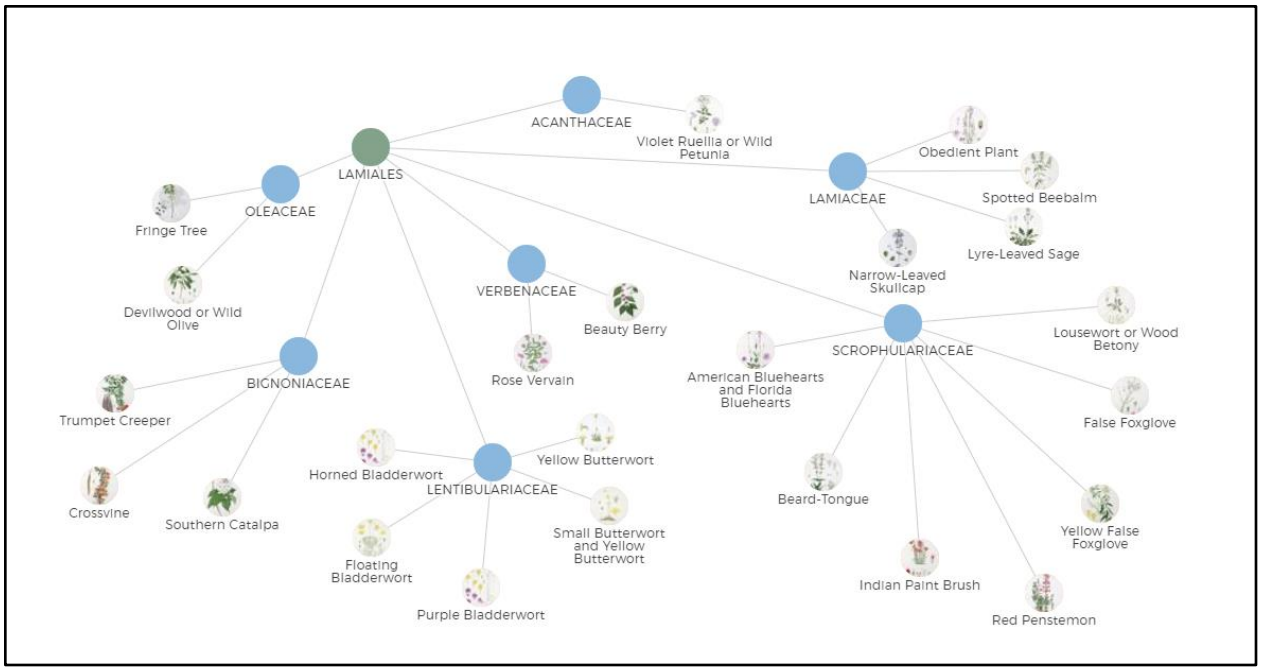


Figure 8: Network graph showing relationship between botanical families. Graph available at https://onodo.org/visualizations/46536/.

These are just some quick examples of the types of questions that can be asked of the data, and visualizations that can be created with data extracted from the LDL. The larger point is that we could never recognize all the uses to which the data could be put. And this is a good thing. Giving direct access to the data opens opportunities to researchers who want it.

Conclusion

The Louisiana Digital Library is primarily used as a means of accessing individual digital items. This “one-by-one” approach works well when a researcher needs access to a single photograph, say, or several letters from the same collection. To complement this approach, additional means are being added to allow for bulk access to the data held within the LDL. Each of the three options detailed above -- Collection Statistics, Search Returns as CSV, and Collection-Level Metadata Exports -- represent

steps toward a new way of conceptualizing and interacting with the LDL. These new directions in working with the data in the digital library speaks in meaningful ways with larger conversations in librarianship. Efforts to promote data-focused approaches to collections will continue as the LDL grows and develops.

References

A release and a call - collections as data facets. (n.d.). Retrieved from https://collectionsasdata.github.io/facets/

Always already computational. (n.d.). Retrieved from https://collectionsasdata.github.io/

Announcing the new British Library research data strategy. (2017, August 27). [Blog post]. In *Digital Scholarship Blog*. Retrieved from http://blogs.bl.uk/digital-scholarship/2017/08/announcing-the-new-british-library-research-data-strategy.html.

Bailey, H., Lynch, K., and Jordan, M. (2018, June). *Enabling computational access at scale: Are repositories serving collections as data?* Panel presented at the Open Repositories 2018, Bozeman, Montana. Retrieved from https://www.youtube.com/watch?v=2_3gc-_TBL8

British Library. (n.d.) *data.bl.uk.* Retrieved from https://data.bl.uk/.

CAD methods profiles. (2018). Retrieved from https://collectionsasdata.github.io/methodsprofiles/

Collections as data 2016. (2016). Retrieved from http://digitalpreservation.gov/meetings/dcs16.html

Collections as data impact. (2017). Retrieved from http://digitalpreservation.gov/meetings/asdata/impact.html

Data lakes, data hubs, federation: Which one is best? (2018, October 9). Retrieved from https://www.marklogic.com/blog/data-lakes-data-hubs-federation-one-best/

Islandora 200 | islandora website. (2017). Retrieved from https://islandora.ca/content/islandora-200

Kariv, A., & Pollock , R. (n.d.). *About datahub*. Retrieved from https://datahub.io/docs/about

Lied Library. (2018). *Collections as data national forum 2*. Retrieved from https://www.youtube.com/watch?time_continue=23&v=ENaPV2XmO9I

Native flora of louisiana - watercolor drawings by margaret stones. (n.d.). Retrieved from http://louisianadigitallibrary.org/islandora/object/lsu-sc-msw:collection

OPenn. (n.d.). Retrieved from http://openn.library.upenn.edu/

OPenn: Read Me. (n.d.). Retrieved from http://openn.library.upenn.edu/ReadMe.html

Part to whole. (n.d.). Retrieved from https://collectionsasdata.github.io/part2whole/.

Potvin, S., & Frost, H. (2018). *Enabling computational access at scale: Are repositories serving collections as data?* [conference report]. Retrieved from https://collectionsasdata.github.io/or/

The Santa Barbara statement on collections as data. (n.d.). Version 2. Retrieved from https://collectionsasdata.github.io/statement/

Weber, C.S. (2017). *Research and learning agenda for archives, special, and distinctive collections in research libraries.* Dublin, OH: OCLC Research. Retrieved from https://doi.org/10.25333/c3c34f