Submitted on: 19.07.2016

The Geospatial Data Curation, Management, and Discovery in Academic Libraries

[Sub-theme: Partnerships, collaboration, expertise – what new partnerships must evolve between research fields and libraries, within institutions and in the research process]

Nicole Kong
Assistant Professor, GIS Specialist
Purdue University Libraries, United Satets
kongn@purdue.edu

Abstract:

As many funding agencies began to require their funded projects to share the data publicly, researchers start to sort for resources in the libraries to curate and publish their datasets. Different with library's traditional collections, datasets are unique products derived from research projects, which require more extensive metadata in order to make it meaningful and reusable. Also, as datasets were generated along the full research process, which datasets need to be preserved and at which stage the libraries should get involved still remains unknown. In addition, inside the libraries how librarians could manage the datasets efficiently is another question that we should pursue in building up our data services.

In this research, we used geospatial data as an example to explore the questions raised above. The reason that we chose geospatial data is because it is a special but widely used data type – it allows us to reach different disciplines yet maintains the scope of this research manageable and practical. Geospatial data has been widely used in many science and technology disciplines, from its traditional application areas in Civil Engineering, Earth Science, and Agricultural studies, to the more recent application areas in computer science, industry engineering, and social science. We collected our information from multi-disciplinary collaborations, focused interviews, and training session feedbacks. From these information, we generalized the workflow of typical research projects in science and technology, and lined up the research workflow with data lifecycles. We have identified the library's role in each research stage, and experimented different software tools, cyberinfrastructure set-ups, and practices that the library could provide in order to facilitate the geospatial data curation, management, and discovery. The initial evaluation of our practice from researchers' feedbacks and data usage statistics have suggested that we have successfully addressed data problems in the collaborated projects.

Introduction

The digital research data generated in science, technology, and engineering are increasing in a rapid speed (Milner, 2009). However, most of these data remain hidden or obsolete (Nordling, 2010). Many major funding agencies have recognized the value of data sharing and the risk of data loss and mismanagement, and therefore, began to require funded projects to have data management plan to ensure proper data archiving and subsequent data sharing. To accommodate the data sharing demands, many academic libraries began to explore their roles and potential services in research data curation, management and preservation (Gold, 2010). Libraries have long played critical roles in knowledge management and sharing. But unlike traditional library collections, research datasets are often generated in massive volume without detailed documentations throughout the full research lifecycle. In order to provide appropriate data services to the research community, detailed studies are still needed to answer key questions of how to effectively identify valuable datasets, create metadata, manage and share useful and usable datasets. It has been suggested that librarians need to collaborate with researchers in the early data production stage for good data management practices, as well as take the custodian's role in data management, reporting, and "downstream" data preservation (Brandt, 2007; Gold, 2007).

In an effort to answer the questions raised above, I discuss data management and curation practices through project participation, faculty interviews, and graduate students trainings, to gather information about their project workflows, data practices, and expected services from library. Based on these information, I generalized library's role in different stages of project development, different practices library could offer in assisting data management and sharing, and research opportunities library can take when providing data related services.

Research data generated in different disciplines vary tremendously. Even within the same discipline, it is common to see various types of datasets are used and generated. This big variation in data types increases the difficulty to explore library's role and tailor the best data services for each specific data type. In this paper, I focus specifically on geospatial data as an example to explore the library's contribution in research data sharing. The primary reason for choosing geospatial data is because this type of data are widely used in multiple disciplines, which allows me to make generalizations across different disciplines, while makes the scope of

this research manageable and practical. Geospatial data has been widely used in many science and technology disciplines, from its traditional application areas in civil engineering, earth science, and agricultural studies, to the more recent application areas in computer science, technology, industry engineering, and social science. Yet, they share some common characteristics such as metadata requirements, prevailing data models, visualization methods, and organizational rules, which allows me to tailor our data services for specific data type.

Project Workflow and Partnership Development

To understand the research community's needs and explore appropriate data services, information was collected from the following three major sources: faculty interviews, project participations, and students' interactions in data related trainings. Faculty interview is the most effective way to collect information about researchers' data needs across various disciplines. A total of 19 faculty and professional staff from 10 different departments in science, engineering and technology disciplines were interviewed using the question sets adapted from Data Curation Profiles (Carlson & Brandt, 2010). The interview questions cover the areas of project/data lifecycle, data acquisition needs, sharing and access expectations, common data tools, and potential library services. In order to obtain more hands-on experiences of data management, I also identified and participated in eight research projects in these disciplines, in which, I heavily involved in data collection, management, and sharing. Collaborating in these projects allows the development of better understanding about the specific data needs in different stages of project development. Finally, I also interacted with graduate students in formal or informal trainings to learn about their data practices and experimented different strategies to solve data problems.

From the interview transcripts analysis and project participation, four major stages of research data development were generalized from researchers' perspective, which includes initial raw data collection, data cleaning and processing, analysis, and final research products publication (Figure 1). Depending on specific project needs, some projects include extra stages, while others omit one or two stages in the generalized diagram. In the following paragraphs, I describe the characteristics of data and potential roles of library in each stage.

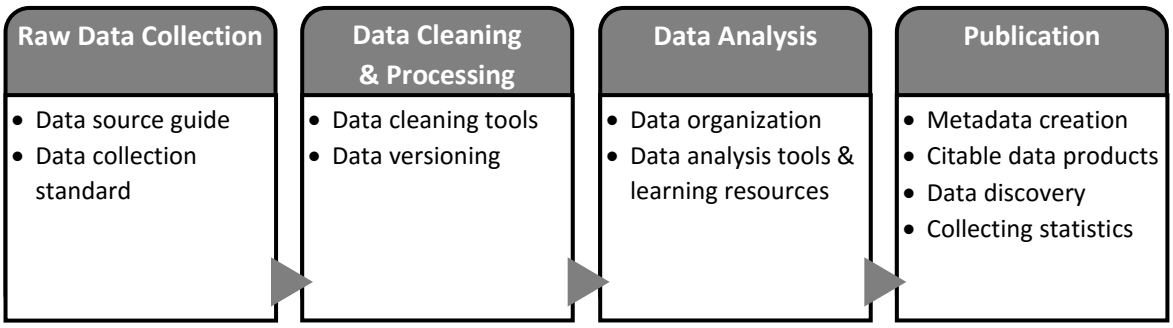


Figure 1. Project workflow identified by researchers and library's role in each workflow

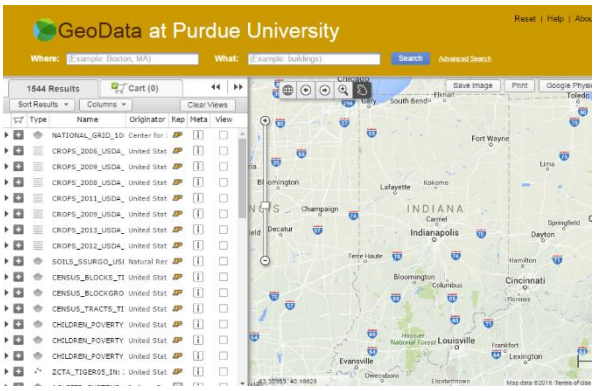
1. **Raw Data Collection:**

In general, there are three major ways of collecting initial datasets: data downloading from public domain, data shared from colleagues, and self-collected datasets. Overall, public domain data are the most often used data for various projects. This is because in the geospatial information world, many datasets are available in public domain via various data portals (Kerski & Clark, 2012). These datasets were collected and shared with different assumptions and potential errors, and were often updated at various schedules without user notifications. In research labs, lab members usually download the datasets and pass them along to other lab members without much metadata information including the downloading date, data quality, etc. Researchers have expressed two concerns in public domain data collection process. First, how to efficiently locate and search useful information among various data portals. Second, how to effectively organize the raw data along with metadata information so that students can develop better understanding about data quality, research assumptions, as well as updating dataset with latest versions.

For data shared from colleagues and self-collected datasets, researchers have similar concerns about data organization and standard. In many team-based data collection efforts, there usually lacks a general design of the data organization structure, the definition of different coding or conditions, and documentation about the background information. Although the collected datasets and naming conventions make sense to the data collection team, they start to confuse others when data are reused.

Based on the concerns raised by researchers, two library services have been identified that could be built around the data collection stage. The first is to provide an efficient data source guide for public domain geospatial data. In practice, we have created a library guide to introduce different data sources by geographic region and by research theme (Figure 2a). In this guide, we tried to organize as many public domain geospatial data portals as possible with brief descriptions and links, so that users can browse through and find the relevant data sources. On average, there were 1,000 visits per month on this data source guide page in 2015. The visits on data source by geographic region are particularly higher than the sources listed by theme, indicating that users primary data search interest is by location. As data portals are increasing rapidly in the era of "big data", browsing through various data portals might not be the most efficient way to find data. As an effort to address this challenge, we are building a geospatial data portal to link the metadata information harvested from different public domains so that users can have a one-stop shop to search for geospatial information (Figure 2b).

The second service that library could build in this stage of project development is data organization and documentation consultation. As many researchers have recognized the importance of using standard in data collection, they need more librarian's expertise in setting up the data documentation standard, file structure, naming conventions, and detailed data code book before data collection so that data can be better understood and shared later on. It requires a close partnership between librarians and the research team to gain understandings about the project and the team culture in terms of data practice, in order to provide appropriate recommendations and trainings to the team.



(a) GIS data guide (b) Geospatial data portal

Figure 2. Geospatial data source guide provided by our library

2. **Data Cleaning and Processing:**

From the collected information, data processing is an important stage the researchers have in order to convert the raw data into a usable format for analysis. This process varies a lot among different disciplines. It includes cleaning noisy information in the raw dataset, data format conversion, merge/clip geospatial data for selected study area, etc. In our interviews, we have designed two questions to ask about the importance of data processing. In a scale of 1 to 5, the average responses to both questions are 4.6, which rank the highest among our overall questions.

While most of the data cleaning and processing tasks need to be accomplished by research team, they expressed needs to have a complete list of useful tools for data processing from the library. These tools should be organized according to the users skill set and effects they have on the datasets. In the past a few years, we have received many questions in regard to the recommended tools for geospatial data cleaning and processing. We are keeping the records of these questions, so that we can create a data processing guide to answer those frequently asked questions.

3. **Data Analysis:**

Analysis is the basic stage of each project, which yields the majority of research data products. Researchers have expressed particular data management needs in the following three areas. The first expectation is to have lab data practice recommendations for their team to document each analysis procedures. Analysis usually involves a lot of experiments, trials and fails. Without a good way to document these detailed steps, it will be challenging to find the successful data results and generate metadata information. In our collaborated projects, we designed folder structure and naming conventions to simplify this documentation needs, which can be easily adopted in the lab practice.

The second expected library service is about sharing data within team during this stage. Some projects require collaborations between departments or institutions. Researchers have tried different ways such as Google Drive, Dropbox, or mailing external drives to share their data in these cases. Different with many other datasets, geospatial data usually require a large storage space, which creates challenges in this stage. While we recommend to use general data sharing platforms, such as Purdue University research repository (PURR), to share the data in this stage,

I also explored specific solutions for easy geospatial data sharing. Depending on different project needs, we are in the experiment stage of providing enterprise geodatabase, GIS server, and web-based data editing/analysis tools to facilitate data sharing between teams (Kong, 2015). It has been approved that these tailored data sharing solutions have greatly simplified and improved the traditional data sharing practices by reducing the data uploading/downloading operations, and providing data quality control within teams.

The third expected library service is specific to geospatial data. Since geospatial data requires specific GIS software for analysis, and not all the research team members are necessarily trained to use the software, faculty expect their students can get flexible training opportunities from the libraries to learn how to use these spatial information. According to this need, I am providing online learning modules, workshops, in-class visits, and consultations to students as needed. For each training opportunity, I tailored the training topics specific to serve for the students' backgrounds and research questions. The increased library visits, requests for workshops, and workshop evaluations have suggested that these training opportunities are very important for graduate students to truly integrate geospatial information into their research.

4. Publication:

The publication stage is usually considered as the final stage of a project by most of the researchers, although they might have very different definitions about publication. Some consider it as the publication of their research paper or technical reports; some consider this stage as the launch of their software or website; and only very a few of them also consider this stage including data publication. It is not surprising to find this attitude toward data publication because data citation and linking is an emerging topic in the past a few years (Ball & Duke, 2011). Library could provide education and data publication information for researchers in this data stage.

In this research, only a few disciplines have existing data repositories. In addition, these repositories only accept limited data products with strict format requirements. In order to encourage the researchers to share their data, library needs to provide a reliable data publication platform with benefits to the researchers. In our libraries, PURR allows researchers to publish datasets and make them available to the world with a unique Digital Object Identifier (DOI) that

can be used to reference the data in paper publications, so that their data could be cited if shared with others.

For geospatial data, the researchers' data publication needs are usually beyond a simple data sharing with DOI. Many researchers are interested in sharing their geospatial data with non-professionals as an interactive web or mobile map. This service requires a reliable GIS server to publish the geospatial data as a standard map service. For the eight collaborated projects, I have provided researchers with this capability and also provided them map visit statistics by time and by region so that they can learn how people interacted with their research data. It has been proved that this cyberinfrastructure is a great success and attracted researchers to use library resources more often for their data publication.

One of the biggest challenges in this stage of data service is about metadata generation. Without an effective metadata, the published dataset will be useless. The importance of metadata has been recognized by every researcher I interacted, yet nobody has been putting efforts on it due to other priorities. In my collaborations, I have learned that there are two major issues preventing researchers from efficiently generating metadata. The first issue need to be addressed is metadata standard and the basic requirement. Geospatial metadata have many existing standards and each standard has a long list of elements to be documented. Library needs to set up the metadata standard serving for the university research community, and provide guidelines about the required and optional fields in metadata generation, so that research labs can easily get started and follow. Second, library needs to set up metadata content standard to guarantee the quality of metadata documentation. Many researchers are not familiar with the statements in metadata about the authorship, distribution information, license agreement, etc. It will be beneficial if library could provide recommendations in these fields and provide examples for them to follow.

5. Discovery and Preservation:

In research data lifecycle, the last but not least is long-term data preservation and discovery. This stage is easily to be ignored by most of the researchers, probably because this is not a required task in their performance evaluation metrics. Once library takes on the data service role, it is an important step which cannot be ignored. Since data service is an emerging service area in libraries, not a lot of studies have been done in this area.

For data discovery, in addition to the effective metadata organization, a powerful and easy-to-use platform is the key element to increase long-term data usage. For geospatial data at Purdue, this research data discovery functionality was integrated into our institutional geodata portal (see Figure 2b for the web interface). Originally, this geodata portal was designed for external geospatial data search as mentioned in the "raw data collection" stage of research project. We ingested the published research data into this portal so that these data can be easily found, previewed, and downloaded for data users.

As for data preservation, we are in the pilot study stage of using Lots Of Copies Keep Stuff Safe (LOCKSS) program for geospatial data preservation. To prepare data for this system, we are in the process of identifying most valuable datasets for long-term preservation, selecting data formats that possibly won't be impacted by software upgrading and retirement, and evaluating the best file batch size for preservation. Geospatial datasets are usually large. Different algorisms need to be developed in order to save the storage space yet preserve the datasets without information loss.

Library's Research Topics in Data Service

While developing partnerships with researchers in the above five stages of a project, there are many research areas need to be explored further by librarians in order to provide appropriate data services. First is about data literacy education. It has been mentioned frequently during our interviews that graduate students need more education opportunities to learn about available data sources, data quality and limitations, data processing skills, and best lab practices for data. It is important for librarians to identify the most valuable and applicable topics to connect with graduate students, so that they are better prepared for data management before the data are published.

Second challenge for librarians is about how to identify the most important dataset to publish and preserve. There are massive datasets generated along the research process. It is not necessary to publish all the data when the research is completed. In my current practice, I recommend to only preserve the datasets that cannot be easily reproduced and document the detailed data processing and analysis procedures if the data can be easily reproduced. For

example, we publish and preserve the field collected data, manually edited data, but don't publish the data if it is just a simple conversion using one particular software function. However, in many practices, it is not clear to draw a definite line between the valuable datasets versus reproducible datasets.

The third on-going research topic is about the best practice for data management and publication on the library's cyberinfrastructure. In different projects, we are researching the best data management strategies in order to save storage space and optimize the data sharing options (Kong, 2015). For example, historical digital maps can be organized into one mosaic datasets with time labels, which saves GIS server's resource to serve this kind of maps, comparing to save them as individual maps for each time period. This research topic requires professional skills from geospatial professionals and also depends on the specific project requirement.

Conclusion and Discussion

This research and practice suggested that academic libraries could play an important role in research data curation, sharing, and management. In our interviews, almost all the participants expected library could take one step forward in assisting them to better manage their data, curate the dataset, and share their data. The data products generated from my collaborated projects suggested that library is an important part of the project in data collection, sharing, management, and curation. Although some research fields have their own disciplinary data repositories, those repositories can only share specific datasets with strict data format requirement. Researchers need a more inclusive platform to share their data from different perspectives. Academic libraries could be the home of their data products around their full research cycle.

This study also suggested that research data curation requires close partnerships with the researchers' teams. Good data management practice needs to be enforced at every research development stage, which is often neglected by the researchers due to other priorities. The collaboration between research team and librarians can start from the very beginning of project planning stage. Librarians can contribute their specialty in setting up the standard for data collection, folder structure, naming convention, and data documentation in the collaboration.

Librarians can also learn more about the project needs from this partnership, and research for optimized solutions for data sharing down to the data publication stage.

Finally, research data are diverse. This paper only takes one special dataset – geospatial data as an example to discuss the collaboration between library and research teams. When we consider the research datasets in general, libraries are still facing many challenges. For example, we need to set up metadata standard that could be adapted to all kinds of research data; other research data needs different kinds of data publication platforms, best management practices, and data preservation judgements. There is still a long way for us to go in order to set up a tailored data service for different disciplines on campus.

References:

Ball, A., & Duke, M. (2011). *Data Citation and Linking*. Digital Curation Centre. Retrieved from http://alexball.me.uk/docs/ball.duke2011dcl/

Brandt, D. S. (2007). Librarians as partners in e-research. *College & Research Libraries*, *68*(6), 365–368.

Carlson, J., & Brandt, D. S. (2010). Data Curation Profiles: Purpose and Use of the Profiles. Retrieved June 26, 2015, from http://datacurationprofiles.org/

Gold, A. (2007). Cyberinfrastructure , Data , and Libraries , Part 2 Libraries and the Data Challenge : Roles and Actions for Libraries. *D-Lib Magazine*, *13*(9). http://doi.org/10.1045/july20september-gold-pt2

Gold, A. (2010). Data Curation and Libraries : Short-Term Developments , Long-Term Prospects. *Office of the Dean (Library)*, *27*, 1–33.

Kerski, J. J., & Clark, J. (2012). *The GIS Guide to Public Domain Data*. ESRI Press.

Milner, J. (2009). A UK Research Data Service (UKRDS): the way forward for research data management%3F - Purdue University - West Lafayette. *Serials*, *22*(1), 83–85. Retrieved from http://purdue-primo-prod.hosted.exlibrisgroup.com/primo_library/libweb/action/openurl?date=2009&aulast=MILNER&issue=1&isSerivcesPage=true&spage=83&title=Serials+%3A+the+journal+of+the+United+Kingdom+Serials+Group.&dscnt=2&auinit=I&atitle=A+UK+Research+Dat

Nordling, L. (2010). Researchers launch hunt for endangered data. *Nature*, *468*(7320), 17. http://doi.org/10.1038/468017a

Short Bio:

Nicole Kong is an assistant professor and GIS specialist at Purdue University Libraries. Nicole holds a PhD in Ecology, and has many years of GIS teaching, development, and data management experiences before she joined Purdue University Libraries. At Purdue, her primary research interests include spatial education in different disciplines, geospatial data management, and spatial information discovery.