Jessica Clemons
Advisor: Linda Musser
Narrative Draft for individual Studies
Submitted 12/11/13

Developing and Populating a GIS Repository of Local Data

Background:

Academic institutions are centers for research, teaching, and learning. Much research involves gathering a variety of different data. From the sciences to humanities, geospatial data have been increasing in use and availability in the past 20 years. Just in the past eight years, geospatial data as an industry has been increasing twice as fast as geographic information systems (GIS) software and services (Foundyller, 2011). The management of geospatial datasets is increasing in importance and often required by major federal funding agencies such as the National Institutes of Health and National Science Foundation (NIH 2003, NSF 2013). Additionally, many federal agencies have data management requirements which promote increasing access to the data gathered as a result of federally funded research. Fulfilling these requirements pose challenges beyond simple file management and organization.

Management of data can be problematic for a variety of reasons. For example, metadata may be missing or incomplete. Metadata is important because it describes how the data were gathered, analyzed, interpreted, created, and other important descriptive details. Geospatial data are particularly problematic because multiple files are needed to accurately view and populate the map layers. Specialized software is often needed to view geospatial data and different versions of files can be very complicated to manage. Files sizes are large and take up space on hard drives and servers.

If data are so complicated to manage, why have these agencies mandated that should it be made available? The question of the value of open and available data is certainly an important topic. One only needs to think of the federal geospatial data that are shared and open, such as topographic data and aerial photography useful for environmental conservation, emergency planning and many other purposes to appreciate the innovations that are a result of open geospatial data (Boxall, 2007). Google Earth is an example of a GIS that uses open data from a variety of sources and is used by many different people. Additionally, many scientists believe that sharing data is a critical component for innovation, interdisciplinary work, and addressing “grand challenges” (Faniel & Zimmerman, 2011).

Libraries have been partners to help with these datasets and requirements from local institutional levels to national approaches (Kollen, Dietz, Suh, &

Lee, 2013). Libraries can also be critical partners to help manage and describe these data (Dietrich, Adamus, Miner, & Steinhart, 2012). Data management and technology training are emerging roles for librarians. Traditionally librarians had been focused on post-publication materials and items. Geospatial datasets can be accessed the moment they are produced, depending on the software platform. Librarians can participate in GIS education and can provide access to relevant geospatial information (Bishop, Grubesic, & Prasertong, 2013).

Developing a Repository:

Providing access to data takes planning and effort in order to be successful. The life cycle of data is an essential piece of managing these massive datasets. (Hartter, Ryan, MacKenzie, Parker, & Strasser, 2013). The portions that are particularly relevant from the library's perspective are the description, preservation, and discovery of data (Figure 1). Of this geospatial data life cycle, three equate to familiar tasks from a library perspective: description of data (creation of metadata), reliable access to and preservation of the items, and discovery.

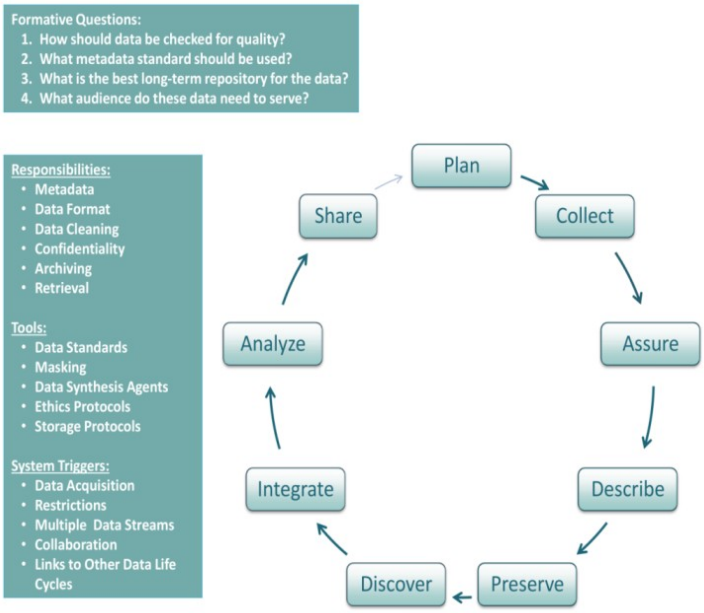

Figure 1: Image credit (Hartter et al., 2013)

Description:

There are many metadata standards available which can either be embedded in the geospatial files themselves or added to a file record ("Open Geospatial Consortium Standards and Supporting Documents," 2013). The Federal Geographic Data Committee (FGDC) is the primary federal organization that contributes geospatial data guidelines. Major attributes include basic identification, data quality, spatial references, features, and

accessibility. The FGDC has several different standards depending on the type of geospatial data (Metadata Ad Hoc Working Group, 2012). However there are no specific geospatial data standards for expressing semantics such as title, spatial tags, or use constraints (Bose & Reitsma, 2006).

Preservation and Storage:

In many ways, managing data and information is what libraries have been doing for centuries. Dealing with the data deluge, particularly geospatial data, is only going to become more important as data is collected and shared. Recent improvements in cloud computing could be an effective way to manage these important collections but using that resource leaves many questions unanswered (Olson, 2010). Archiving geospatial data in its many forms is an extremely complex challenge. The National Digital Information Infrastructure and Preservation Program (NDIPP) is an important resource when beginning to archive geospatial data but evaluating and inventorying data can be complex since there are so many sources available (Morris, 2009). In today's market data storage is cheap so large datasets are not overly expensive to store. However, multiple files associated with each dataset present challenges (Erwin & Sweetkind-Singer, 2009). There are resources available to track past and present file formats which can be useful for legacy and modern datasets (Hoebelheinrich, 2012)

Discoverability:

Good metadata and preservation standards will aid the discovery of available geospatial data since it will be searchable and recoverable. It is extremely difficult to discover data if only the creator knows about it. This "word of mouth" type of discovery is serendipitous, cumbersome, and unlikely. Libraries have begun to collect geospatial data, add metadata, and develop geospatial data catalogs to help researchers (Kollen et al., 2013). States have been developing and cultivating geospatial data collections for decades. New York has a GIS Clearinghouse that includes a catalog of datasets and information on how to obtain the data ("Data Sharing Cooperative," n.d.). There are a host of online geospatial data repositories, from industry to grassroots organizations. It is important to determine where the data would be most discoverable in order to have an effective, accessible geospatial data collection.

Populating a Repository:

One of the biggest challenges of beginning a geospatial repository is uncovering where the data currently are. On campus there are datasets running on basement servers, ArcGIS Online, hard drives, flash drives, and others. Finding the people with the data will be a critical first step. Once some specific datasets are identified (probably from some local or Adirondack research stations), there will be a pilot

project established. The idea behind this test is to exhibit to other faculty what can be done with the data and build participation. The datasets will be added to a SUNY Digital Repository collection http://dspace.sunyconnect.suny.edu/ as a test case to host sample files and metadata. The file type(s) will depend on the sample data that is procured. It will likely be a ArcGIS zipped file and/or KML files which are used in Google Earth. Next a metadata schema will need to be established. The files will be uploaded to the repository with metadata. To aid in the discoverability, a local data finding aid will be created and posted to the web. Depending on what the data are, the library could also add it to its catalog. This is a simple but time-intensive process. Once completed, it will be used to generate interest and support of the project.

Assessment:

A project such as creating and developing a geospatial repository of local data is no small task when the amount of data available is considered. This is not a project that has a discreet beginning and ending. Assessing the usability and usefulness is important and will enhance project funding from outside the campus. Usage statistics can be gleaned from the SUNY Digital Repository. This may be helpful because I can see how people are searching for information and how they find the collection. The online finding aid offers usage statistics. I would offer a brief poll to determine user satisfaction (probably 3 questions) on the finding aid. I also plan to list my contact information so people may directly contact me to add data to the repository or answer questions. My target group initially will be those identified people who have and use geospatial data on campus. After we have more than two usable datasets (probably at least 10 sets to represent a variety of information) I plan to survey the faculty to determine if this fits the needs of campus or if there is a way to partner with other institutions.

For me personally, I will measure success in other ways. Working with geospatial data will help me understand the nature and types of data that are collected on campus. This will help me better support the research faculty. It will help me increase the use and visibility of library services as we develop rich local collections of data. This project will also create citable units of information. According to Reilly, citing unpublished data can be difficult since there are no static citations. A geospatial data repository will allow for easier citation (2012). Faculty will see the impacts of their research as people use and adapt the geospatial data in new and exciting ways. If faculty see the library as partners in this process, it may help us secure funding and support for other initiatives.

Next steps:

There are two major steps that need to happen for this project to continue. I need to get copies of some datasets to work with as examples. I have identified a specific

person with data but he is not often on campus so coordinating a meeting is difficult. Additionally, my test database platform (Dspace) is currently being upgraded so new records can't be added until after January 1, 2014. I also plan to pursue collaboration with CUGIR to see if that is a good fit for our data. This collaboration would be very desirable from our point of view since they have a dedicated staff to handle geospatial data specific to New York State.

References:

Bishop, B. W., Grubesic, T. H., & Prasertong, S. (2013). Digital Curation and the GeoWeb: An Emerging Role for Geographic Information Librarians. Journal of Map & Geography Libraries, 9(3), 296–312.

Bose, R., & Reitsma, F. (2006). Advancing Geospatial Data Curation. Retrieved from http://www.era.lib.ed.ac.uk/handle/1842/1074

Boxall, J. C. (2007). Advances and Trends in Geospatial Information Accessibility—Part II. Journal of Map & Geography Libraries, 3(1), 57–78.

Data Sharing Cooprative. (n.d.). NYS GIS Clearinghouse - Coordination Program. Retrieved October 29, 2013, from http://gis.ny.gov/co-op/

Dietrich, D., Adamus, T., Miner, A., & Steinhart, G. (2012). De-mystifying the data management requirements of research funders. Issues in Science and Technology Librarianship, 70(1). Retrieved from http://www.istl.org/12-summer/refereed1.html?a_aid=3598aabf

Erwin, T., & Sweetkind-Singer, J. (2009). The National Geospatial Digital Archive: A Collaborative Project to Archive Geospatial Data. Journal of Map & Geography Libraries, 6(1), 6–25.

Faniel, I. M., & Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. International Journal of Digital Curation, 6(1), 58–69.

Foundyller, C. (2011). GIS/Geospatial Sales Up 10.3% to US$4.4 Billion Growth Forecast to Top 8.3% in 2011. Cambridge, MA: Daratech.

Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially Explicit Data: Stewardship and Ethical Challenges in Science. PLoS Biology, 11(9).

Hoebelheinrich, N. J. (2012). An Aid to Analyzing the Sustainability of Commonly Used Geospatial Formats: The Library of Congress Sustainability Website. Journal of Map & Geography Libraries, 8(3), 242–263. doi:10.1080/15420353.2012.700301

Kollen, C., Dietz, C., Suh, J., & Lee, A. (2013). Geospatial Data Catalogs: Approaches by Academic Libraries. Journal of Map & Geography Libraries, 9(3), 276–295.

Metadata Ad Hoc Working Group. (2012). Federal Geographic Data Committee endorsed standards. Standards publications. Retrieved November 13, 2013, from http://www.fgdc.gov/standards/standards_publications/

Morris, S. P. (2009). The North Carolina Geospatial Data Archiving Project: Challenges and Initial Outcomes. Journal of Map & Geography Libraries, 6(1), 26–44.

National Institutes of Health. (2003, February 26). Final NIH statement on sharing research data. NIH Guide Notice. Retrieved October 28, 2013, from http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

National Science Foundation. (2013, January 1). NSF Data Management Plan Requirements. Dissemination and Sharing of Research Results. Retrieved October 28, 2013, from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Olson, J. A. (2010). Data as a service: Are we in the clouds? Journal of Map and Geography Libraries, 6(1), 76–78.

Open Geospatial Consortium Standards and Supporting Documents. (2013). Open Geospatial Consortium. Retrieved November 13, 2013, from http://www.opengeospatial.org/standards

Reilly, S. (2012). The role of libraries in supporting data exchange. Presented at the International Federation of Library Associations, Helsinki.