Challenges to Sharing Data Among
Environmental Scientists and Data Managers in the Southeastern United States

Poster Abstract for:
34th Annual College of Information and Communication Research Symposium
"Leaping Forward in Communication and Information Research"
University of Tennessee
February 29, 2012

Mary Beth Ross West and Miriam L. E. Steiner Davis, Ph.D.

Introduction

As data storing and data sharing technologies change within scientific communities, it is essential to understand how environmental science data sets are developed and stored in order to develop more accessible systems for those interested in using and accessing environmental data. A data set refers to a specific type or group of related data collected by a researcher.

In this study, on behalf of the United States Geological Survey's National Biological Information Infrastructure (NBII) Southeast Information Node (SEIN), a program designed to identify and provide access to environmental data in the Southeast, researchers interviewed 29 southeastern scientists and scientific data managers. Interviews included questions about data set information (e.g., data set contents, size, variables, collection and management methods), data formats and storage (e.g., how the data was stored, format of storage, metadata standard), current data sharing (e.g., data availability to others, issues or barriers to data sharing, conditions of access, willingness to share data with the NBII, and requirements for sharing). Geographically, the US Southeast (defined in this study as Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina and Tennessee) is of particular concern due its high levels of biodiversity and endemism. By understanding the parameters of environmental and resource information in the Southeast, the USGS can provide and distribute data and information in ways that increase effectiveness and accessibility.

Background

In 2009, SEIN partnered with the University of Tennessee's Center for Information and Communication Studies (CICS) on the *Increasing Biological Information Sources: Technical Assistance and Support for Delivery and Technology Transfer (IBIS)* project. The project aimed to identify relevant data sets in areas of research that served NBII's stakeholders: environmental decision makers, researchers and scientists including citizen scientists, and teachers and students. The IBIS project also developed biodiversity information tools and services to address the accessibility of USGS provided biodiversity information.

Research Methods

In 2010, the IBIS Biodiversity Users Information Needs survey was created and distributed online to scientists and environmental decision makers in the Southeast. The final survey question invited respondents to participate in further research about their biodiversity data. Of the over 400 survey responses, 29 scientists and scientific data managers agreed to be interviewed by the IBIS researchers and discuss their data. The IBIS researchers interviewed the 29 scientists and data managers, each interview was recorded and notes were taken. After each interview an interview profile and one or more data set profiles were created per interview.

IBIS researchers created data set profiles via a template that captured each data set's information relative to data management practices. The information was coded and sorted based on its relevance to their data collection and management practices, which for this study included availability of the data and the data sets, methods and formats of data storage, metadata creation and standards implementation, and requirements or conditions for data sharing. Once all of the interviews had been completed, the IBIS researchers worked to identify the core issues that affected biodiversity information practices in the Southeast. The data sets were also analyzed in terms of their defined work sector, belonging to either an academic, nonprofit or government agency.

Research Findings

Forty data sets were identified: 23 academic data sets, 10 nonprofit data sets, and 7 government data sets. The discovery of 40 data sets from 29 interviews indicates a one to potentially many relationships between any given researcher and the data sets they work with. In addition, the findings reveal that individual researchers often engage in varying practices with respect to the data sets they work with. In other words, researchers' can work with more than one data set at once and their data practices can vary among the data sets they work with resulting in data sets with differing parameters being identified and discussed within one interview.

Four core issues resonated in each of the data sets; 1) the availability of the data set, 2) how the data set was stored and formatted, 3) how the data set was organized (i.e., whether a metadata standard was adopted and used), and 4) the restrictions or conditions imposed on the data set before the data set could be shared.

Accessibility of the Data Sets

Data sets from all three sectors were distributed or made available to others through various means discussed below. Some data sets were distributed via only one method (i.e., the data set is made available online for others to review), while other data sets were distributed in multiple ways (i.e., they are made available to the general public by request, but they are available to specific agencies via online channels). Data set distribution methods varied greatly by sector due to the types of trade or embargo agreements in place. The majority of the academic sector data sets (N = 13) were available by request, while the majority of the nonprofit sector data sets (N = 7) were made available online. Government sector data sets (N = 3) were also primarily available online.

In terms of online availability, even though each sector had data sets available online, "being online" was not indicative of free or unrestricted access to the data. Many of the data sets that were available online, had extensive security measures, authentication protocols and restrictions on what was available.

Data Set Storage and Formatting

Similar to the variations seen in data set accessibility and distribution, data set storage methods varied greatly. Each of the 40 data sets have components stored in a digital format (i.e., they were saved as a file or files in a computer program or database), some of the data sets are fully digitized, while others are in the process of partial or full digitization. The greatest variation in digital data storage came in the form of how they were stored (e.g., on a personal computer versus a networked drive or a server) and in what program the scientist used to enter or input the data (e.g., Microsoft Access, Microsoft Excel, ARCGIS, SPSS, Oracle, etc.).

Twenty-six data sets were also stored using physical storage methods; the majority of these were from the academic sector, with a total of 16 data sets using physical data storage (e.g., printed data sheets, field notebooks, specimens, etc.). The nonprofit sector had a total of 6 data sets that used physical formats. The sector with the fewest data sets using physical storage was the government sector, as only 3 data sets used a physical method for data set storage. Physical formats ranged from field notebooks, data spreadsheets printed on paper, physical specimens and specimen photos, to printed baseline documents, tapes, and hard copies of field data.

Metadata Creation Standards

The use of metadata standards was not consistent among the data sets. Only 21 of the 40 data sets were associated with a metadata standard; 8 of these were from the academic sector, 8 were from the nonprofit sector, and 5 were from the government sector. Even within those 21 data sets associated with metadata, there was not consistency in the standards used. Four data sets had metadata created without the use of a standard, while others were associated with owner/manager/scientist-created standards (N=2), Darwin Core (N=4), Federal Geographic Data Committee (N=3), NatureServe (N=3), and Ecological Metadata Language (N=2). Three data sets were associated with an unspecified or unknown (to the interviewee) metadata standard used.

Conditions to Data Sharing

Data set sharing was associated with many various conditions. 28 data sets included at least one condition for sharing. These conditions could be specific agreements that had to be met before the data could be shared (e.g., an easy to use system had to be setup for the researcher to import their data set into or proper attribution had to be given to the data set's collecting agency), restrictions preventing the data sets from being shared with certain groups (e.g., the general public or only sharable with named government agencies), or restrictions preventing the data sets from being shared at all (e.g., endangered species data for this region could not be shared in order to protect the species studied). Of the 28 data sets with conditions for data sharing, 13 data sets were from the academic sector, 9 data sets were from the nonprofit sector, and 6 data sets were from the government sector. Only 12 data sets were not associated with any conditions for sharing. Ten of these were from the academic sector, 1 was from the nonprofit sector, and 1 was from the government sector.

Conditions to Sharing with National Biological Information Infrastructure (NBII)

Of the 40 data sets analyzed, only 33 data sets were indicated to be sharable with the NBII. Of those 33 data sets that could be shared (either in part or in whole) with the NBII, 20 were from the academic sector, 8 were from the nonprofit sector, and 5 were from the government sector. 25 of the sharable data sets were identified by the respondents as having requirements to sharing their data sets with NBII. Of those 25 data sets, 17 data sets were from the academic sector, 6 data sets were from the nonprofit sector, and 2 data sets were from the government sector. Of those data sets that could not be shared, 8 data sets were from the academic sector, 2 were from the nonprofit sector, and 2 were from the government sector.

Conclusions

Environmental scientists, researchers, data managers and decision makers in the US Southeast work with many different data sets, each of which can have their own unique parameters with respect to their creation, description, management, sharing and availability. Variation in the four areas cited (accessibility, format and storage, metadata standards, and data sharing) such as those described here can create challenges to gathering, presenting and providing access to data sets from different managers. However, identifying the core points of variation and the reasons for them, as well as similarities and differences within and among key data set sectors, allows those, such as the USGS, wishing to identify existing data sets and improve access to them to strategically develop identification and provision initiatives and practices which can best maximize the existing potential for data sharing and accessibility.