5-4-2016

IBIS interview report.docx

Suzie Allard
University of Tennessee - Knoxville, sallard@utk.edu

Miriam Davis
University of Tennessee, Knoxville, miriams@utk.edu

The survey results provide an understanding of the prevalence of scientists behaviors and attitudes regarding their data. The results of the in-depth interviews allow us to gain insight into the details of how scientists "do" science and how they handle their data. This report outlines the interview process, our analysis and key findings.

PROCEDURE

Interview participants were recruited from survey respondents. At the end of the survey, respondents were asked if they would be interested in being interviewed. Those that were interested were asked to provide their contact information, which was collected on a form that was separate from their survey responses so that their survey responses could not be attached to an individual. The interview procedure was conducted pursuant to the human subject research form approved by the University of Tennessee's Institutional Review Board.

IBIS researchers contacted those survey respondents who volunteered to be interviewed and scheduled an interview. Interviews were conducted by phone or face-to-face, depending on proximity and on participant preference. Interviews were recorded. The recordings were used to create profiles that captured the key information shared by the respondent and to capture some quotes. The interviews were not transcribed word for word because analysis was conducted on the thematic level to capture major concepts. Recordings were destroyed after the profiles were created.

Most of the interview was conducted using guiding questions. Some very specific background questions were asked to establish the primary role of the individual regarding biodiversity, the kinds of biodiversity activities in which the individual was engaged, the participant's primary work sector and the primary subject discipline. Guiding questions were then asked regarding the context of the biodiversity work and the respondent's data. Respondents were asked how they collected and stored their data, the format of the data, the data tools used, and the metadata standards. Questions also addressed the respondent's data sharing attitudes and behaviors. After the interview was complete, respondents were asked if they had any data they would like to make available via NBII.

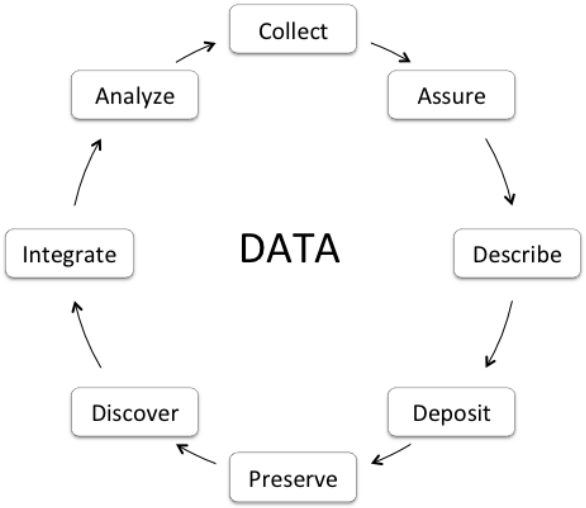
Twenty-eight interviews were completed. The following table outlines the demographics of interview participants.

Role	Total	Male	Female	Academic	Government	Non-Profit
Total	28	23	5	14	7	7
Administrator	8	7	1	1	5	2
Field Scientist	8	7	1	4	1	3
Lab Scientist	8	5	3	7	0	1
Curator	4	4	0	2	1	1

ANALYSIS

Analysis on the emerging themes was conducted using the DataONE data lifecycle (see figure below) as a guide for grouping discussions of behaviors and attitudes by each work role: administrator, field scientist, lab scientist, and curator. The data lifecycle identifies eight unique stages that data may progress through, although data does not necessarily progress through all eight stages – it may miss a stage at some point. Also one person may not be responsible for handling the data as it moves through these stages. Therefore, an individual may only be involved with a couple stages of their data's lie.

The data lifecycle begins at the point a scientist collects her data. The scientist may assure the quality of the data by reviewing the records and checking for proper notation. Data is then described which is when metadata is created. Data may then be deposited in a trusted repository where is can be preserved. After deposition data can be discovered by other scientists. Data modelers may discover the data and integrate multiple data sets to create a new understanding. Data can also be analyzed and used to create visualizations.



ADMINISTRATOR

Eight interviews were conducted with individuals who would be classified as biodiversity data administrators. The majority of these individuals (5) work in a government work sector. Two participants are in the non-profit work sector and one was in academe.

Collect: Data collection is very heterogeneous in terms of collection instruments, scale and scope. Data sheets are important data collection instruments and many of these are still hand-written although there is a preference to move towards machine based data sheets if it were economically feasible. Other instruments used are data loggers and vouchers. Some data is collected using instruments such satellite imagery. The scale of the data also varies greatly, ranging from focusing on the species level and creating relatively small files to collecting satellite imagery that is being stored at the terabyte scale. These administrators are working with data that represent up to 60 years of observations. Other collections are just over a decade old. These differences in scope mean that administrators are dealing with very different legacy issues.

Assure: This is a stage that seems to get little attention. Only two of our respondents noted any behaviors associated with this stage. One noted that quality requirements revolved around how the data gathering locations were chosen, so quality control was focused on the quality of the initial observations. The other participant noted that quality assurance occurs at the point data is actually being put into the database since it is collected in many different formats and the data manager conducts quality checks.

Describe: Reponses that address this stage illustrate the range of understanding of metadata, and only six people made comments that addressed this stage. Some respondents were very conversant with metadata and noted specific schema (e.g. FGDC) that were used to describe records or datasets. However the majority of respondents did not work with metadata or know if those who created the data worked with a formal metadata standard. One respondent noted he was not interested in learning more about metadata since "My plate is run over."

Deposit: Most of these respondents felt data was deposited when it was resident on their own machines. Moving the data to a larger repository such as NBII raised concerns about assuring that proper credit was given to the dataset creator, and about the negative outcomes that could result from data sharing. As one person noted, "I have mixed feelings about making data available in a place like NBII. If I've gone to all the hard work, why should someone else personally profit from it?"

Preserve: Everyone addressed the issue of preservation but the strategies varied. One noted that they kept a hard copy of all final reports and also kept paper copies of the data. Others noted that their servers have backup and security features to address archival issues. Several respondents are preserving their data using proprietary software such as Oracle or Access, Excel or SAS. There does not seem to be any concern about the longevity of these software platforms.

Discover: Just about all the respondents do not make an special effort to make data visible to others to encourage data discovery. One respondent noted that if no one knows they are collecting the data, then no one will miss it. There is also concern about making discovery easy for those who may use the data for commercial

ventures thus making money off it. A respondent said, "I really hate doing all the grunt work and having someone else come in and take advantage of it and get grant money." There are some important reasons cited for not sharing data such as protecting rare species and providing location information about the data.

Respondents who did make their data visible to others, even if selected others, often kept records of who had the data and how they would use the data. Some of these records are written, others are based on oral agreements with "cooperators."

There was concern with making raw data available to the public, and instead it was considered important to complete synthesis first.

Integrate: Only four participants discussed data integration, and all of these people commented on the problems related to integration. For example one participant said, "XX wanted our data in a format so detailed that it was going to cost a lot of money to package it the way they wanted it." Another noted, "Sometimes in the research world, you have so much technical information, being able to synthesize information to where you could use those things to educate people is the hard part."

Analyze: Participants noted how their data were used, in most cases to inform decision making for resource management. One noted that the data were used to educate landowners and agencies and to fund research to understand the value of a particular ecosystem.

FIELD SCIENTIST

Eight interviews were conducted with individuals who would be classified as field scientists who participate in some activities related to biodiversity. About an equal number of these participants work in an academic environment (4) or a non-profit work sector (3). Only one works in the government sector.

Collect: These scientists are very concerned with all aspects of data collection. How the scientists collect the data is driven by the kind of observations they need to record. All these scientists recorded data at the site of the observation or specimen collection. These data were recorded in a variety of ways including keeping hand-written notes, using standardized forms, writing field reports, and/or on handheld computers. Handwritten notes are digitized back at the lab.

The data were collected to record the presence of a species and to record location information using GPS coordinates. Most of the scientists noted that location data is essential and must be recorded with each observation or specimen, sometime using hand-held computers that record the GPS coordinates. One scientist who collects specimens noted that data has always been described by noting a site, data and collector, and that now a second label has ecological information and GPS coordinates. Another noted that the observations were recorded in a natural heritage format that includes species list and a narrative description of the site.

Most of these scientists are collecting data at a number of sites (some from as many as 29 sites) or their data is being added to a dataset that represents a large number of sites and often a large region. One scientist noted that s/he collected two kinds of data – quantitative data from within the designated 15 meter circular plot, and qualitative data from between those plot.

Historical data was also mentioned as being important to data collection since this helps the scientist have a baseline for knowing what existed in an environment in the past. Those scientists who mentioned historical data noted that they collected information that allowed comparisons but also added some data points that could not be collected in the past such as GPS coordinates. One scientist noted that s/he were starting a new database which would include data collected primarily in the last 20 years although there were some observations from as far back as 1911. The data points include lat/long, date, political variables, collection method and disposition of the specimen.

Assure: While the interviews yielded rich description of data collection activities, only two scientists made comments regarding the assure stage, and these comments were generally very brief. One noted that certain criteria need to be met for an observation to be added to study. This scientist noted that quality control included having a data sheet reviewed by others around the world. Another simply noted that all data is quality checked before it was added to the biotics database in order to be sure it is correct and complete, but this scientist did not elaborate on how this was accomplished.

Describe: The assignment of metadata is varied and often is not done in a way that will allow the full power of metadata to be exploited. Four of the scientists explicitly stated that the metadata being assigned has been created just for these data and as one noted s/he "just created a list of information I thought I would need." This scientist did note that this personalized metadata was consistently applied.

One scientists noted that s/he believed that a metadata standard was being used but did not know which one and that all was required was a periodic "dump" of the information and someone else dealt with the metadata.

Another scientist noted that s/he ha no idea of a metadata standard was being employed and that "As long as I don't know about it, I'm happy. (I)don't want to get involved."

One scientist noted that his group was now partnering with NBII that drove the need for them to create metadata from the field notes, although the group needed help to do this successfully. This was a difficult task and the scientist expressed a desire for NBII to provide a template of required metadata elements for each data set.

Another scientist noted that there was difficulty with transcribing the data from specimens into the database.

Only one scientists specifically named a standard metadata schema, Darwin Core, This same individual also noted that the entomology collection used the ITIS standard.

Deposit: Six scientists made reference to their data deposition behaviors. These scientists did not differentiate between the different services that data deposition may offer – for example data discovery and sharing versus preservation. Comments suggest that all these scientists equate data deposition with sharing, and there are very mixed feelings about sharing data.

One scientist expressed concern about sharing data, "I'm reluctant to share with someone who is part of a big, powerful research lab especially if one of my students is working on something similar." Another concern that was ascertaining that the repository holding the data would be sure that the researchers who gathered the data would get appropriate credit for their work.

Four of these scientists were already sharing or willing to share heir data – with conditions. These included: (1)Data would be shared on request which allows the researcher to know who will be using the data. (2) Data would be shared on the condition that the person or project requesting the data would share their data. (3) Data would not be shared until the scientist had had published from it, effectively setting an embargo on the data. (4) Data would only be shared after sensitive information such as information that could reveal the location of endangered species was removed.

Two scientists specifically referenced NBII. One commented that s/he wasn't sure whether NBII wants the data only when it is a finished data set or if there could be progressive updates. Two others noted that their data was probably already being shared with NBII through the programs in which they were engaged.

One scientist noted that s/he had no control over the data and that sharing decision would need to be made at a higher level.

Preserve: All eight scientists made comments that touched on the preservation stage, however there was a wide range of expertise level. All noted their technological solutions such as what programs they used to store their data, where those files were housed, how often they were backed up and where these backups were stored. Programs used included proprietary software such as File Maker Pro, Excel, and Access. There was no mention of concern about the long-term viability of access to these programs. One specifically mentioned keeping PDF formatted reports, but did not mention if it was PDF-A or if these reports had any original data or only synthesized data. Several mentioned having copies of the data set stored on several different laptop and desktop computers and perhaps on an organizational

server. External hard drives were also a solution. Only one scientist specifically mentioned a regular backup schedule.

Paper was also seen as a viable preservation strategy. For example, one scientist noted that data sheets were duplicated and stored in two locations. Another reported creating a booklet for each county and distributing those to public libraries and to the State. A third scientist that original coies of the reports were kept in a fireproof safe in the office.

Discover:

Integrate:

Analyze:

LAB SCIENTIST
Eight interviews were conducted with individuals who would be classified as biodiversity data administrators. The majority of these individuals (5) work in a government work sector. Two participants are in the non-profit work sector and one was in academe.

Collect: .

Assure:

Describe:

Deposit:

Preserve:

Discover:

Integrate:

Analyze:

FIELD SCIENTIST
Eight interviews were conducted with individuals who would be classified as biodiversity data administrators. The majority of these individuals (5) work in a government work sector. Two participants are in the non-profit work sector and one was in academe.

Collect: .

Assure:

Describe:

Deposit:

Preserve:

Discover:

Integrate:

Analyze: