Metadata Capature and Geospatial Records

Elizabeth Perkes , Utah State Archives and Records Service, Salt Lake City, Utah/USA; Lisa Speaker, North Carolina State Archives, Raleigh, North Carolina/USA

Abstract

When the electronic records that you are trying to preserve are unique, complex, and storage-hungry, they will quickly put an institution's feet to the fire to come up with solutions. This has been the case for Utah, North Carolina, and Kentucky as we have tried to grapple with the needs and requirements of geospatial records in the grant-sponsored GeoMAPP project (http://www.geomapp.net). Much of what we have learned while studying geospatial records can be broadly applied to other types of electronic records. For instance, digitized images of the earth will have similar preservation requirements as documents that have been scanned, but with the added metadata needed to make sense of geospatial imagery. Geospatial data in the form of shapefiles or geodatabases also come with their own descriptive metadata, which must be captured along with the technical metadata, and reused for purposes of access and preservation. This session will focus on the nature of this metadata and the commonalities found with other types of electronic records, while we share the specific strategies and tools that we are developing. One such tool is an application created by the Utah State Archives, called the APPX-based Archives Enterprise Manager (AXAEM). This platform and database-independent open-source software is used to manage the entire workflow of the archives, and recent development has added the ability to ingest metadata of various types into the system and link it to the bibliographic data of series. A demonstration of this tool will be given.

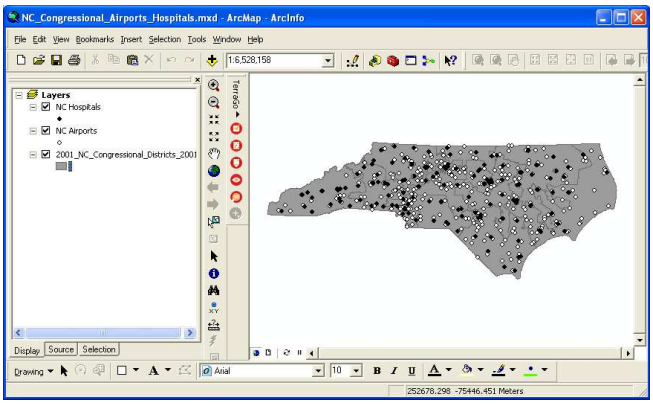

***Figure** 1: Esri ArcMap view of 3 datasets: North Carolina (N.C.) Hospitals (white dots), N.C. Airports(black dots), and 2001 N.C. Congressional Districts*

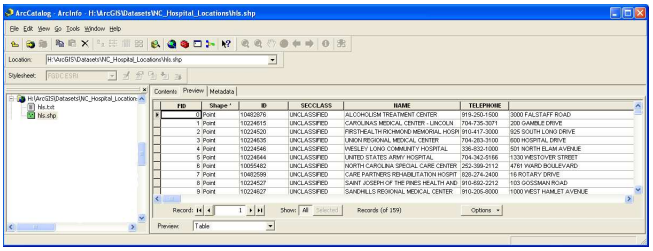

***Figure** 2: Esri ArcCatalog view of data in the N.C. Hospitals dataset*

Digital Geospatial Datasets and Their Metadata

When preserving geospatial datasets, archivists encounter the usual challenges associated with preserving born digital objects, such as dependence on special software applications, transferring and preserving "authentic" or "trustworthy" digital artifacts, and creating an appropriate archival metadata record that facilitates and ensures the access and manageability of digital assets into the future.

Geospatial datasets are produced from geographical information systems (GIS) which combine graphical representations depicting geographical features with tabular data that store information related to those features. At one level, GIS can be considered as a sort of electronic map that is supplemented with an underlying database [1]. A GIS dataset for hospitals can hold the geographical point locations for each of the hospitals in a state, plus store additional information associated with each hospital such as its name, address, telephone number, emergency services, and number of beds (see Figures 1 and 2).

Geospatial datasets are similar to other digital assets in that they are generally created by specialized application software, and specialized application software is also required to read or update existing geospatial datasets. In many cases, the format of the geospatial dataset is vendor specific, and can only be read and/or written by tools provided by that software vendor. There are some formats, such as Esri's Shapefile format [2], which have been published, and have non-vendor-specific rendering tools available. However, geospatial data formats are more complex than most other common digital formats. Unlike digitized document files, image files, and audio files where the digital asset and its associated metadata are contained in a single file, geospatial datasets are often composed of numerous files, and often have a separate rich metadata file.

The Federal Geographic Data Committee (FGDC) is a national committee that "promotes the coordinated development, use, sharing and dissemination of geospatial data on a national basis."[3] The FGDC is tasked by Presidential Executive Orders to "develop procedures and assist in the implementation of a distributed discovery mechanism for national digital geospatial data."[4] The FGDC has developed the Content Standard for Digital Geospatial Metadata (CSDGM), a rich metadata standard to describe geospatial data [5]. The CSDGM contains several

subsections that include descriptive, technical, provenance, and administrative metadata elements, and also specifies which metadata elements are required. In addition, CSDGM defines fields to record the lineage and processing history of the dataset, also useful for informing provenance-related archival records.

Archivists have long advocated for metadata creation to accompany the creation of the digital record. GIS software packages promote this best practice, as they offer interfaces for GIS developers to create the metadata to describe their datasets. The GIS creator can fill in traditional metadata fields such as creator, date created, and abstract (see Figure 3a). The GIS software might even assist the GIS developer by automatically populating technical metadata fields such as the GIS software application name and version, and host operating system, which are important metadata elements for archivists and the digital object's future sustainability. The software may also extract geospatial characteristics directly from the GIS dataset and populate the corresponding metadata fields, further increasing the reliability of the metadata and reducing human labor and the opportunity for human error. To promote the accessibility of the metadata, tools are available to export the metadata in a standard XML format (see Figure 3b), which can serve as a useful input for automating archival metadata production.

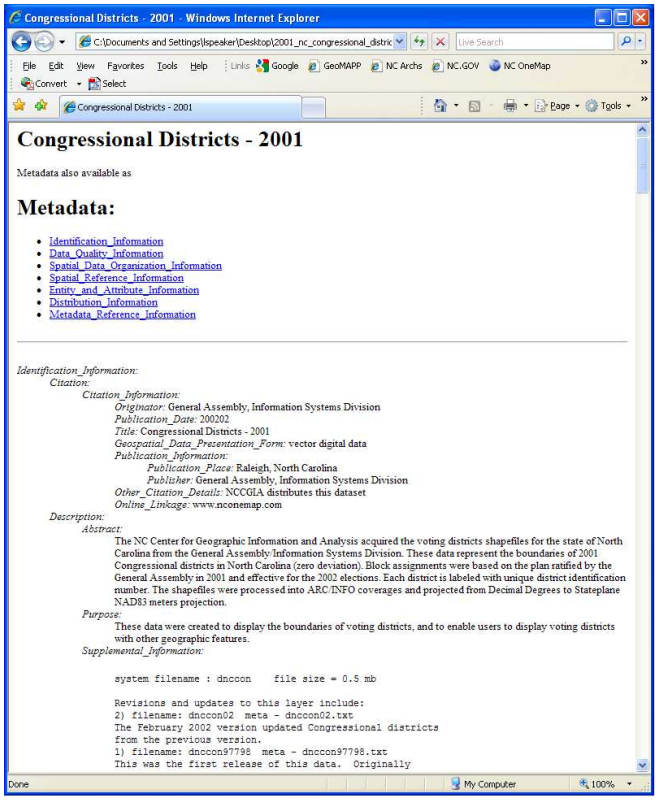

***Figure** 3a: Excerpt: GIS metadata for N.C. 2001 Congressional Districts dataset*

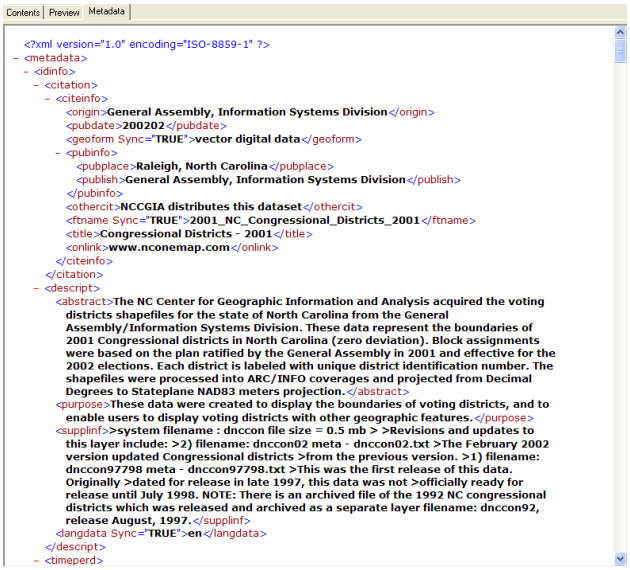

***Figure** 3b: XML Excerpt: GIS metadata for N.C. 2001 Congressional Districts dataset*

With the extensiveness of the geospatial metadata record, state geospatial coordinating councils may establish geospatial metadata standards, and the state archives may need to determine if they will establish policies and procedures regarding the (non)acceptance of geospatial datasets that are not adequately described.

As with other digital objects, the archivist is tasked with building an archival metadata record that facilitates the long term management and access of the digital geospatial object. By 'archival metadata record' we mean the entire collection of metadata associated with an archived digital object, which can include descriptive, technical, or administrative aspects of the archived digital object. This archival metadata can be repurposed to address different areas of responsibility in an archives, such as populating a catalog record or finding aid to promote access, or referencing the technical details to manage potential migrations to more contemporary data formats. At this point, unfortunately, there remains a lack of standards that define a comprehensive dictionary of metadata elements to represent archived digital objects, so each archives is building its own archival metadata dictionary to support the digital assets it manages.

Unlike the "simple" single-file digital formats, the archives will generally receive along with the geospatial data file(s), a rich geospatial metadata file, which can be useful in creating the archival metadata record. Given the extensiveness of the geospatial metadata fields, the archivist may only need to extract a subset of the GIS metadata fields that are key to the access and management of the dataset. For example, to support access needs, the archivist can extract descriptive fields such as:

- the title (<title>) to populate the archival record title field,
- the originator (<origin>) to populate the archival record creator field,
- the abstract (<abstract>) and purpose (<purpose>) to populate the archival description field,

- the time period of the content (<timeperd>) to populate the archival date created field

Much of the technical metadata required for the archival metadata record including application software vendor, name and version, and underlying operating system, may be extracted from the Native Data Set Environment (<native>) metadata field. The archivist can then supplement the GIS metadata with archival-related metadata such as fixity values, rights statements, archives accession and ingestion dates, archival processing actions, etc., which aid in the preservation of the datasets.

The above metadata fields are likely common across all types of digital objects. However, different types of digital objects will also have format-specific metadata that may be included in the archival metadata record. Consider image characteristics such as image resolution or bit depth, sampling frequency or noise reduction for audio files, frame rate for video files, or encoding method for text files. GIS datasets, similarly, will have some format-specific metadata that might be extracted from the geospatial metadata file and included in the archival metadata record, such as the:

- bounding coordinates (<eastbc>, <westbc>, <northbc>, <southbc>) that could be used as the basis for a geographically-oriented search interface,
- spatial data organization information <spdoinfo> (e.g. point/vector type of object or raster object) <direct>, or
- geospatial reference information <spref> such as the coordinate system <horizsys>.

The geospatial metadata file may also provide extensive information regarding the data attributes, such as definitions <attrdef> and data sources <attrdefs>. The attribute metadata will be of interest to future geospatial researchers, but is not necessarily relevant to the archival description, so may not necessarily be included in the archival metadata record. Even if they are not included in the archival metadata record, you can still offer the end user easy access to these additional fields by offering an HTML-version of the geospatial metadata file through your access interface.

The archive's rights policies for a dataset are likely to be different than the original rights documented in the geospatial dataset, therefore, geospatial metadata such as the use rights (<useconst>) and access constraints (<accconst>), may be best left in the geospatial metadata file.

Preparing GIS datasets and their metadata for archiving has provided techniques that can be applied to the management of archival metadata for any type of digital object, including:

- identify the metadata elements common to all digital data formats, and then
- identify format-specific metadata, such as the FGDC geospatial metadata,
- evaluate which common and format-specific metadata to extract for its archival record,
- create a crosswalk for each digital format to document the metadata mapping between the data format's metadata and the archival metadata record to facilitate metadata extraction for the archival record,
- define the metadata extraction process, whether it is manual or technology-assisted such as with the APPX-based Archives Enterprise Manager (AXAEM) described below.

GIS Archival Metadata Case Study

In Utah, the State Archives for many years has been using a system it developed to manage its records. Record creators (governmental entities and related persons) as well as their functions are identified first, and then data about records are entered as new retention schedules are needed. For those records appraised as having historical value, other data is added which builds upon the initial retention schedule description, including details for finding aids, indexes, microfilm, etc. This system has been fully integrated with a third-party box inventory system used by the records center for space management. Physical records are then known and quantified, whether the disposition is "destroy" or "transfer to Archives," and tied to bibliographic descriptions for access.

With the advent of the GeoMAPP project, a concerted effort has been made to allow this system to ingest electronic records of all types and capture their metadata. This system has been named the APPX-based Archives Enterprise Manager (AXAEM), and is available as an open-source application. The features listed below are in various stages of development. Some have been completed and are now being run in a production environment, and others are still being programmed and tested for future release. The new Electronic Records menu as it currently exists is seen in Figure 4.

Figure 4: Electronic Records menu in AXAEM

In identifying metadata for its electronic records, the Archives referenced the standard developed by the Mountain West Digital Library (MWDL), a consortium of institutions of which the Archives is a part. This document [6] outlined the various Dublin Core metadata elements and their preferred usage within the MWDL. Processes were added in AXAEM that reflected this basic metadata organization.

Dublin Core, however, did not offer the specificity needed for geospatial data, nor did it acknowledge the technical details that some metadata extractor tools are able to capture for a variety of formats. Also, since geospatial records tend to be multi-file and multi-format, additional functionality within AXAEM was needed.

The database structure was edited so that there would be one table identifying individual files, called Electronic Records, plus sub-tables for metadata elements that are repeatable or unique to specific formats, and another table identifying Object Groups. An AXAEM "Electronic Record" refers to a single row/entry in the Electronic Record table. An Object Group may consist of one or more Electronic Records, as well as other Object Groups. The Object Groups may then effectively nest together to reflect the structure of the actual record. Finding aids may point to either an individual Electronic Record or an Object Group for description purposes.

A geospatial shapefile usually consists of up to six files, each a different format, which will result in six Electronic Records, bound together within one Object Group. Without all pieces of the shapefile in place, the data cannot be opened or accessed, so recording the relationship between these files is critical. In AXAEM, the metadata are captured within each Electronic Record and reflect details pertinent to that item. Technical metadata will differ from file to file, but descriptive metadata will be the same between all elements of a shapefile or geodatabase.

AXAEM can create the metadata record for an Electronic Record from several sources: 1) an XML parser was added to the underlying AXAEM software, allowing AXAEM to ingest metadata supplied in XML files, 2) a file ingest feature integrated with metadata extraction tools, allowing AXAEM to extract metadata directly from a variety of file types, and 3) import metadata from .csv files. With the XML parser, AXAEM can now map any XML schema or standard to the metadata fields of an AXAEM Electronic Record, and then populate the metadata fields by importing the XML file. The data ingest feature uses a process of copying files to the server, then running a specified metadata extractor tool (e.g. JHOVE [7], New Zealand Metadata Extraction Tool [8]), which then produces the XML metadata values to write to the Electronic Record. To the end user, this is a one-click operation after adding initial identifying data.

As more metadata fields are determined to be desirable for specific file types, fields will be added to AXAEM. In this sense, AXAEM can accommodate any metadata standard, and adding fields is very easy. This system is intended to be highly flexible. For instance, in the event that a single file contains more than one format, as may be the case with some complex TIFF files, the data structure supports the identification of each embedded format within a single Electronic Record entry.

For geospatial datasets, AXAEM intends to retrieve metadata using both the file ingest feature to extract the file-specific metadata data, and the FGDC metadata supplied in the geospatial XML file. As there may be similar aspects in the technical metadata extracted by JHOVE and the geospatial metadata, it is intended that all of the metadata will be rationalized and merged into a single metadata record for each electronic record. This multi-step metadata extraction and assignment process will appear to the user as one single process.

The data ingest screen (see Figure 5) asks for data such as the location of the original files being ingested, storage location of where data should be sent, name of XML schema map definition being used, record series ID, records transfer/accession ID, batch ingest ID, and a pointer to a digitization workorder if applicable (such workorders contain data related to hardware and software used in digitization projects). The ingest process includes capturing a checksum of the file(s) being ingested, and storing the checksum value in the record in the database, just as it does the extracted metadata. The files are placed in a storage location accessible to the application. When the ingest is complete, a report is automatically generated identifying the key(s) of the newly added records, as well as any error messages encountered in the ingest process. This report may be ingested on its own merits and entered as a related record to any Electronic Record.

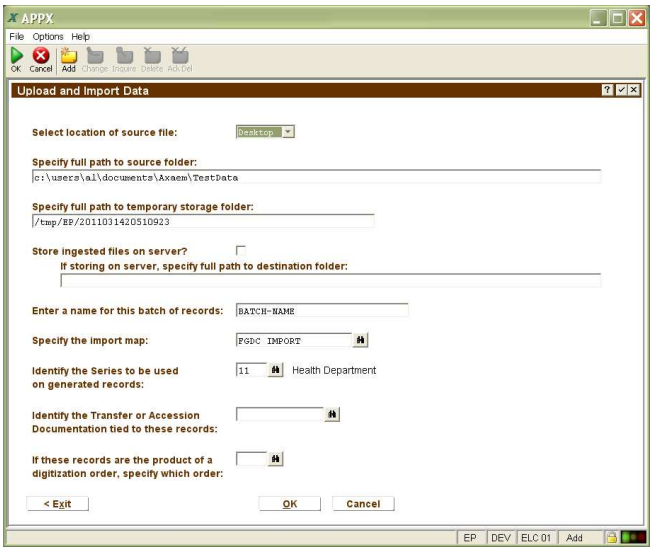

Figure 5: Data ingest screen

The screens in AXAEM that are used to view and edit data about records are organized around *Describing Archives: A Content Standard* (DACS) [9] principles. Figures 6 and 7 reflect newly-ingested records of municipal boundaries into the Electronic Records table. The Electronic Records were created using the FGDC metadata file.

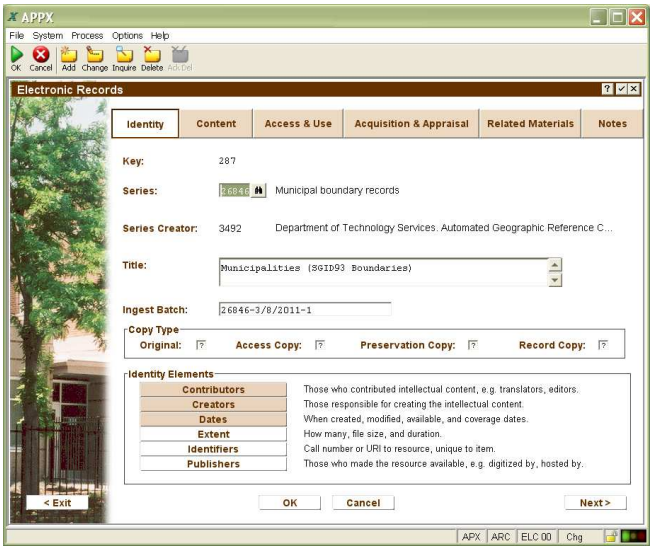

Figure 6: Municipal boundaries

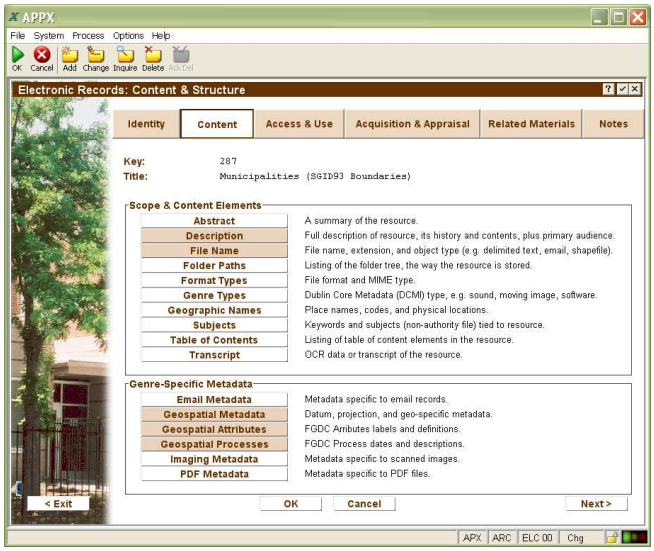

Figure 7: Metadata content for municipal boundaries

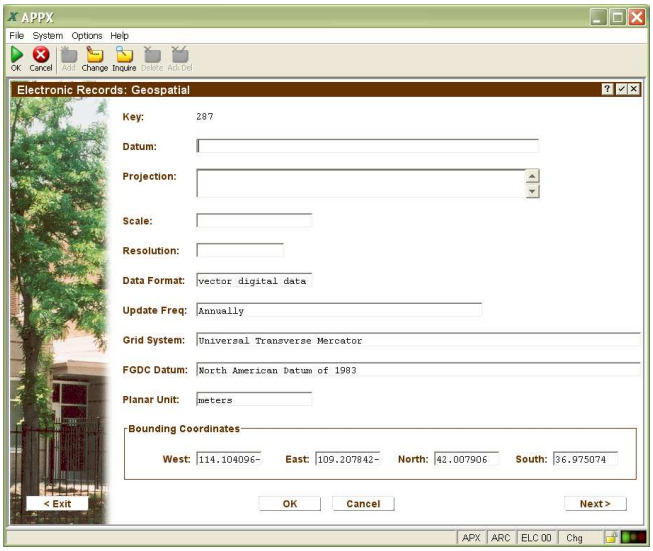

Figure 8: Geospatial metadata in municipal boundaries

The decision about which metadata to capture from the FGDC record was made from feedback received from the GeoMAPP participants, which included GIS practitioners and archivists. They indicated what metadata fields were most important to them to understand, utilize, and preserve the datasets. Such metadata fields included datum, projection, resolution/scale, and publication date. Presumably, these details would need to be visible within a finding aid or other advertisement of resources available from the Archives.

One difficulty that comes from auto-ingesting metadata is a lack of control and consistency over data entry formatting. For example, different GIS developers can describe the same grid system or FGDC datum differently. AXAEM can provide drop-down lists or lookup values on fields to keep data neat and consistent, but that cleanup would need to be done within the AXAEM application after ingest (see Figure 8).

Another way to update metadata within the Electronic Record is to use AXAEM's export to .csv feature, which automatically opens the data in Excel. The auto-fill features contained within the spreadsheet software may be used to populate fields, then the data may re-imported from the saved .csv file. This option is available on the Electronic Records menu (displayed in Figure 4). To export, simply choose the metadata categories you wish to edit (creators, formats, subjects, etc). Then the query will ask you which set of Electronic Records you want to change, such as by record series ID, ingest batch ID, or a range of Electronic Record IDs.

Facilitating the descriptive effort for each Electronic Record, much of the metadata pertinent to an Electronic Record will be inherited from its record series or collection description information. This might include usage rights, scope and content, technical access notes, related materials, and appraisal data, which will be common to most types of digital objects. Figures 9 and 10 display screens from the series record, and data are organized here around DACS principles just as the Electronic Record screens are.

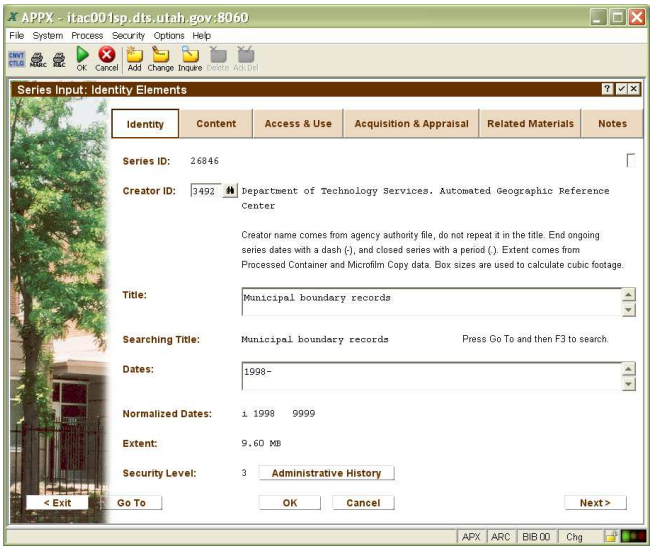

Figure 9: Record Series entry for municipal boundaries

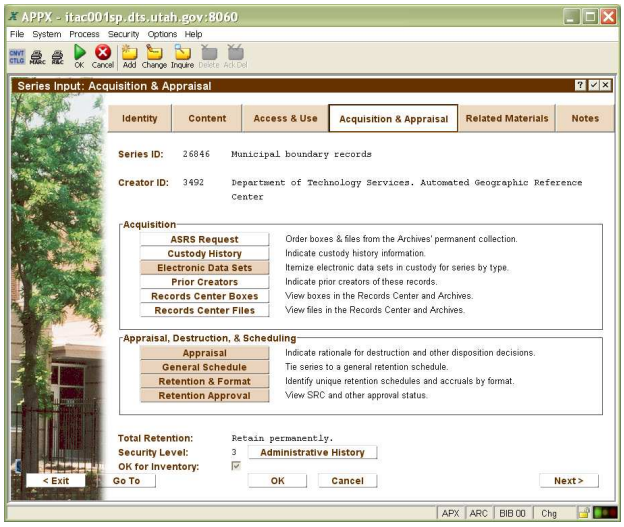

Figure 10: Acquisition and appraisal data for the series

Conclusion

In preserving digital artifacts, creating and managing the metadata remains a significant challenge, compounded by the variety and complexity of digital formats being preserved. Automating the creation of the archival metadata record remains a goal for archivists, to reduce processing time and to improve the quality of the archival record. In order to automate the creation of the archival metadata, first the metadata that will comprise the archival metadata record must be identified. We suggest a phased approach to defining the archival metadata, by 1) identifying common metadata attributes, such as those elements based on Dublin Core, 2) identifying format-specific metadata for the particular data formats you manage, and 3) finally creating a cross-mapping between the digital object's metadata and the archival record metadata. Fortunately, geospatial datasets have a rich metadata standard that includes attributes that can promote their long-term use and management, from which to populate the format-specific archival metadata. In addition, tools such as JHOVE can extract format-specific metadata from many common digital formats.

The availability of a well-defined XML-formatted geospatial metadata file, and the XML-export from tools such as JHOVE and the New Zealand metadata extractor, lend themselves well to automating the population of the archival metadata record. The development of solutions such as AXAEM, which supports both simple single-file and complex multi-file digital data formats, offers archivists an attractive solution for automating data ingestion, and creating and populating the archival metadata record. With the metadata in place, and the files safely ingested, preservation of digital objects is one step closer to being realized, and a system is in place to facilitate the long-term sustainability of a variety of digital data formats.

References

[1] A to Z GIS. Ed. Tasha Wade and Shelly Sommer. Redlands, CA: ESRI Press, 2006. Print.

[2] Environmental Systems Research Institute. "ESRI Shapefile Technical Description: An ESRI White Paper."1998. Retrieved 7 Mar. 2011 from http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[3] The Federal Geographic Data Committee. "The Federal Geographic Data Committee." Retrieved 7 Mar. 2011 from http://www.fgdc.gov/

[4] The Federal Geographic Data Committee. "Geospatial Metadata: What are Metadata?" Retrieved 7 Mar. 2011 from http://www.fgdc.gov/metadata

[5] The Federal Geographic Data Committee. "Content Standard for Digital Geospatial Metadata". Retrieved 7 Mar. 2011 from http://www.fgdc.gov/metadata/csdgm/

[6] Mountain West Digital Library. "Dublin Core Application Profile." 2010. Retrieved 14 Mar. 2011 from http://mwdl.org/public/mwdl/MWDL_DC_Profile_Version_1.1.pdf

[7] JSTORE. "JHOVE-JSTOR/Harvard Object Validation Environment."2009. Retrieved 18 Mar. 2011 from http://hul.harvard.edu/jhove/

[8] National Library of New Zealand. "Metadata Extraction Tool." Retrieved 18 Mar. 2011 from http://meta-extractor.sourceforge.net/

[9] The Society of American Archivists. "Describing Archives: A Content Standard." 2004.

Author Biography

Elizabeth Perkes is the electronic records archivist at the Utah State Archives and Records Service, where she has worked for the past twenty years managing the Archives' data systems and collaborating with stakeholders on electronic record policies and guidelines. A Certified Archivist, she received her Master of Library and Information Science degree from Brigham Young University in 1990.

Lisa Speaker is an electronic records archivist for the North Carolina State Archives focused on the GeoMAPP project. She received her Master of Science in Library Science degree from the University of North Carolina in 2010 and an MBA from the University of Michigan in 1993. She has over twenty years experience in the software industry, involved with metadata data-driven web application design and development including content management, portal, and e-commerce solutions.