Q&A with Glen McAninch

Kentucky Department for Libraries and Archives

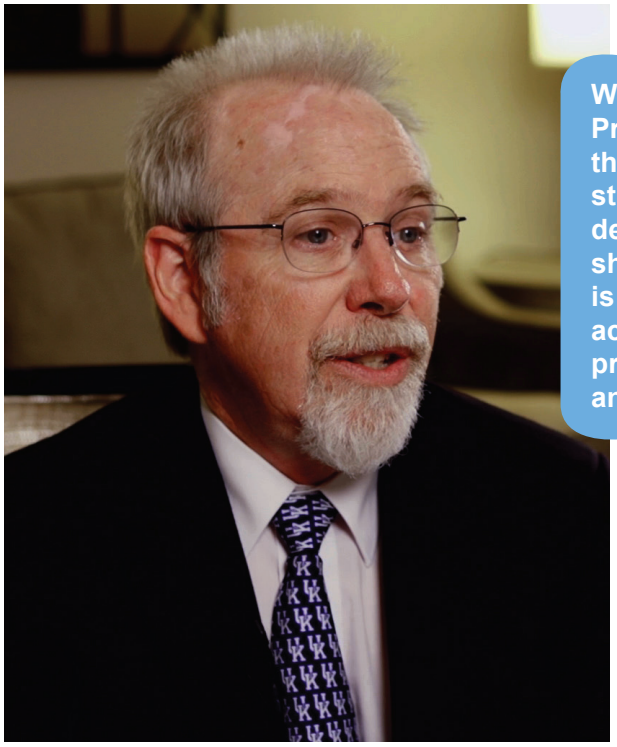
Preservica
Active digital preservation

The Importance of Provenance, Context and Metadata in Preserving Digital Archival Records

Of the many discussion topics among library archivists, the issue of provenance, as it relates to digital archival records, is especially important. Whether the archival objects are those termed "born digital," such as publications, photos, minutes, videos and geospatial records, or are records that have been digitized from early non-digital or analog documents, capturing and maintaining critical provenance data can be challenging.

In his, A Glossary of Archival and Records Terminology, 2005, Richard Pearce-Moses (past President of the Society of American Archivists and a distinguished Society Fellow) defines provenance as "information regarding the origins, custody and ownership of an item, or collection." He describes items for inclusion in this category as "arrangement, context, creator, custodial history, entity of origin, office of origin, and original order." In an additional note, Pearce-Moses also refers to the principle of primacy of provenance in archival description that "… holds that the significance of archival materials is heavily dependent on the context of their creation, and that the arrangement and description of these materials should be directly related to their original purpose and function."

In a recent interview with Preservica customer, Glen McAninch, Branch Manager for the State of Kentucky's Department for Libraries and Archives, we discussed this topic of provenance and contextual history of archival records and how this information is captured and recorded for preservation of digital materials in the Kentucky State archives.



What we try to do in Preservica is maintain that organizational structure, provide descriptive data that shows what the record is and then also add accession information to provide the who, what, and when for the record.

Q: There seems to be a lot of discussion in your industry on this topic. How significant is provenance data for digital archives and what elements comprise the most important data to capture? Is the context of their creation an essential part of that as well?

While the origins of the term "provenance" come from the world of paper, the context and custodial history is still significant in the digital world. It is important to establish who created the document; who has held the record; whether the document is an official authenticated copy; and what is the full context of the record. In a government setting, what the agency is, what its function is, and what the function of the particular record is, defines the larger context and therefore the provenance of the record.

This kind of information should be made part of the metadata set that is incorporated into each digital record upon accession as part of the record's identifier. It should also include how the data was obtained, who and where it came from, and what format it was in when it was received. There are other notes that can be made at the time of accession to further establish provenance, such as why and how it was used, and whether it came from a web site.

Q. What else might be included in this accession data that would be important to future researchers?

The accession information is generally administrative in value as a means for the archive to document the transfer from the creating agency to the archive. However, the accession information for archival records takes on added significance to provide context for future researchers. The origins of records are often buried in the electronic records. This includes what software was used to create the record and if it has undergone transformation before coming to the archive. When we take the records in, we add another layer of administrative data – accession data – that puts the records in a hierarchy of agency and series with the record creation date and the circumstances of getting the records, such as by downloading it or as a direct transfer from the government agency. The accession metadata helps establish the authenticity of the record, because it's generally thought that an authentic record has to account for its custodial history including both the creation and the transfer of the record.

Q. What do you mean by putting the records in a hierarchy of agency and series?

While most agencies create and maintain common records series, such as photographs or publications, other series are specific to an individual agency and are only found within that agency. The hierarchy of the organizational structure – cabinet, department, or branch level – also puts records in context, and that is a significant piece of provenance. The purpose or functional description of records is another context element. This should come in detail from a description for each series found in the records retention schedule, particularly for those that are agency specific.

What we try to do in Preservica is maintain that organizational structure, provide descriptive data that shows what the record is and then also add accession information to provide the who, what, and when for the record.

Q: And this is all included in the metadata that is attached to the digital records in Preservica? It sounds like an extraordinary amount of information.

It is. One of our tasks is to identify and capture accession metadata or "context" metadata related to the records coming in that is particularly important long term. Preservica facilitates this process by using a template to attach this metadata to every item within the accession.

Q: So if you're deciding what information a researcher of the future would want or need to know, it would seem to be subjective in some way at this point.

Yes, anticipating future use is subjective. We can suppose for a long-term record, particularly those that have statistical research value, someone would want to know the full context of the record including how the information was gathered. In the case of electronic records, especially database records, there could be a lot of detail about the collecting of the data that is unknown. Shared data sets, such as geospatial records, are often created at a local level under certain rules and procedures, but when the record is created, the agency responsible for the data collection often neglects to include much of the descriptive and administrative type of data needed, and in some instances doesn't include any data at all. The source of shared data sets, federal, state, and local government, as well as commercial, is a complex issue of custodial history. The date and method of the collection may not be specified in the record that is archived, though is a crucial piece of context provenance.

> **"Being able to maintain all the information possible about that object's transformations is particularly relevant to establish the context."**

There is a lot of information that could be very valuable to researchers in the future that we may have problems capturing. It would be very difficult for us to go back after we accessioned the record and see that it doesn't have the accurate date the data was actually collected. We know the date it went into a database, but whether research data on bird migration, for example, was taken in September, October or January would be very important for research.

Q: Are you putting all of this information into Preservica?

With Preservica we are focusing on accession metadata, which has to do with a part of provenance, though it's not the whole picture. Selecting which data elements are important, particularly for research or legal value, out of so many possible elements, is difficult. If you are looking at a legal or historical analysis, you want to know specifically how we got the electronic record. How authentic is it? Is it original? We had to decide which elements to include for each accession, by accession. The name of a particular agency contact, for example, would not be as important to a researcher years from now as how the archive acquired the object and the context of its creation.

Q: What are the methods you're using to try and make this record pertinent in full context as well as showing its provenance?

We've selected maybe 6 or 8 different accession data elements that we regularly include in the records. While we have a whole set of records in our DSpace repository with item level descriptive metadata (Dublin Core) that will be transferred to Preservica, we also are entering into Preservica many large accessions with only the descriptive metadata that can be extracted from folder titles and the agency/series information entered with each upload. The accession metadata thus becomes even more important for those records without the more detailed descriptive metadata.

Preservica also automatically generates information as part of the technical metadata that extends the history of the record migration path through different formats. Being able to maintain all the information possible about that object's transformations is particularly relevant to establish the context. During the migration process, maintaining as much of the original record for format sensitive media, such as video and audio, is particularly important so you don't lose the functionality, context, or the content.

This is also true for records that are harvested from the Internet which have a complex set of relationships between the linked parts. Records, such as photographs, that are published on the Internet often contain more descriptive metadata than the unpublished files that are sent to the archive by the creating agency. Other web records can best be used as a series of linked parts rather than single objects. Preservica has the capability of harvesting websites with the data all linked and distributed in various layers. The ability to display these records in full context for future researchers is crucial for web records.

On a related note, one of my colleagues, along with a colleague from Preservica, will give a presentation at the Society of American Archivists' conference that will focus on the whole question of selection of metadata elements to capture. The presentation takes into consideration descriptive and technical metadata as well as administrative metadata.