Access and Preservation: Addressing Challenges of Linking Cultural Heritage Datasets

Fenella G. France; Library of Congress; Washington DC, USA.

Michael B. Toth; R.B. Toth Associates; Oakton, Virginia, USA

Abstract

Development of a spectral reference database of scientific samples with data from advanced imaging techniques in the Preservation Research and Testing Division (PRTD) at the Library of Congress expands the utility of non-destructive scientific analyses for cultural heritage research. This yields enhanced preservation data and provides greater access to lost information hidden in the non-visible and obscured in the visible regions. It includes data from collection objects such as drafts of the Declaration of Independence and Gettysburg Address, maps, drawings and prints. To ensure access and usability of this data, a "scriptospatial" system for representing data from the original document supports effective integration of diverse scientific and scholarly data, layered with spatial, temporal, cultural and historical data. To support archiving of digital products with consistent quality, PRTD has established standard workflows for data management. The integration of these new capabilities into the PRTD spectral imaging system enhances this technology to advance non-destructive analysis of collection items. This initiative is developing and storing layers of data that augment scholarly and curatorial research, aid library and student investigations of history and science, while preserving and protecting the original documents.

Introduction

The launch of what became the first Landsat satellite a little over 40 years ago prompted a surge in both spectral imaging and large amounts of image data, with over 300,000 spectral images produced for civil applications.[1] With the advent of advanced spectral imaging in cultural heritage studies, the past decade has seen a similar surge in spectral image data. These need to be integrated with common standards to provide useful information to support scholarly research, preservation studies, and scientific analysis. The Preservation Research and Testing Division (PRTD) of the US Library of Congress has adapted data handling and management, data preservation, and methods and software for cultural heritage studies from those used to handle earth resource and geospatial data, but with the much more limited budget and infrastructure of a cultural heritage institution. This initiative has included customizing the imaging system and its hardware, software, metadata and processing with a standardized approach to enable rapid access to linked data by scholars and scientists.

Spectral Imaging Studies

Digital spectral imaging of cultural heritage objects in the US Library of Congress has capitalized on over a decade of research and development into not only spectral imaging and processing, but also the development of standard spectral image products.[2]

Advances over the past decade by an experienced team have led to an advanced capability to study cultural heritage objects with a robust spectral imaging system that provides large-format, high quality images and standardized data output using advanced commercial off-the-shelf components.[3] Building on pioneering research and development efforts in systems design, spectral illumination and imaging, and data management during spectral imaging studies of the Archimedes Palimpsest, David Livingstone Diaries and St. Catherine's Monastery Palimpsests, combined with continued research and development at the Library of Congress, an integrated spectral imaging system is now providing new insights into cultural heritage objects from the Library collection.[4] This capitalizes on best practices in government and industry in applying not only new technologies, but also work practices and personnel. The system includes integrated image collection, processing and data storage technologies and enhanced work processes, which trained and experienced personnel are transitioning from leading edge technology to a working tool.



Figure 1. Spectral Imaging System at the Library of Congress

The Library's spectral imaging system has developed into a multi-functional capability for extracting the maximum amount of data from a single spectral imaging session. This reduces the impact on often fragile collection items, while capturing provenance data, baseline condition and changes due to environment and treatments, and recovery of lost and obscured text and images.[5] Experience has shown there is a clear need to balance the access to preservation based data, while retaining and

preventing loss of contextual scholarly information inherent in a cultural heritage object. Too often objects are treated essentially as separate entities for preservation and scholarly purposes, leading to a segregation of any data collected between disciplines, and a divisive perspective based upon the original purpose of the study. The ability to non-invasively capture, store and access a cube of research data that bridges the gap between science and humanities shows the successful integration of technology into multi-disciplinary research.

Careful assessment of the layers of information present in an object can reveal significant details that are important to both preservation experts and scholars. While tracking changes to ensure optimal preservation, pigments, colorants and substrates can be identified without taking a sample. The captured information can also provide context as to the time period when the pigment was available, and if this is commensurate with the age of the object, the geographical location with which it was historically associated, potential evidence of trade routes that transported pigments distant from the known origin of the object, and even later additions or changes made to the object. Given digital capabilities, identification of these pigments can allow a digital 2D or 3D rendition of the object, restoring faded or deteriorated colors to their original palette, without the need to restore rather than conserve the object. Further, identification of the substrate, watermarks in paper, trace elements in particulates, or materials that link these back to specific locations, adds additional layers of information to the object. Physical layering of media, colorants, inks and other substances on a document or object can be better understood with spectral imaging, allowing scholars to analyze construction methods, printing techniques and original obscured and deteriorated text. This allows the original intent of the author or artist to be captured. For example, the change from "subjects" to "citizens" on Jefferson's handwritten draft of the Declaration of Independence, fingerprints on Lincoln's draft of the Gettysburg address, redactions in papers and journals from the Founding Fathers, all provide more information from spectral imaging data than can be accessed from normal visual interaction with the object.

Digital Cultural Objects

The integrated layers of data now create what is termed the "digital cultural object" (DCO) - an object that cannot replace the original, but adds non-visible information contextually in both temporal and spatial layers. Historic documents do not readily lend themselves to context analysis, since documentation about the creation of the document may not be readily available. The additional contextual information is not apparent in conventional digitization techniques for these objects, so the integration of the spectral data assists in mining the layers of data stored within the objects. In this way the DCO provides a range of data and information. This allows a shift from the use of interpretive virtual heritage applications, which focus on the artistic rather than the investigative and inferential, toward the development of interdisciplinary scientific data as part of cultural heritage scholarly studies. In this way, cultural heritage studies, science and technology are intertwined, advancing the capacity to mine and analyze historic data from multiple viewpoints. Coordinating the relationship between the original and the DCO entities enables greater access to scholarly information, since identification of

materials enables access to provenance, geographical and temporal information to extend knowledge about the historic document.

Waldseemüller Map

The application of this technology to the Waldseemüller 1507 World Map provided a number of insights for cartographic researchers. Johannes Schöner (1477-1547), a German astronomer and cartographer and pupil of Waldseemüller, drew red lines over the map from the Middle East north to the Black Sea, an area that must have been of interest.[6] Over time, these inscribed red gridlines on Sheets 6 and 7 had faded significantly and become virtually imperceptible. These gridlines represented important added features in the cartographic history of this map, so it was important for curatorial collaborators to find a way to retrieve this lost information. With spectral imaging, these lines could be distinguished through their unique spectral response. The reconstruction allowed researchers to assess overlapping of the lines - including which lines were laid down first, where lines began and ended. This began to broaden the interpretation and understanding of Schöner's thought processes from the early 1500s. While the Waldseemüller 1507 World Map is famous for the first reference to 'America', it remains a source of intrigue since the cartographer is known, but not the precise time or location of printing [7].

The 1507 World Map was discovered in 1901 by Joseph Fisher, a Jesuit historian conducting research in the library collection of Wolfegg Castle, Wurttenberg, Germany. He was convinced that the World Map sheets were printers' proof sheets, so the ability to gain greater understanding about how the woodcut was produced was of great interest to scholars, cartographers and historians. Specialized processing created a virtual 3-dimensional rendering of the map, giving a visual perspective of the original woodcut from which the sheets were printed. Since the original woodcut no longer exists, the processed images allow an analysis of possible techniques used in the early 1500s to inform scholars about creation techniques and gain a better understanding of the materials used in the process. These layers of data comprise the DCO, adding a rich resource of non-visible information that need to be stored and accessible as a critical data set to support expanded contextual knowledge of the document.



Figure 2. Principle Component Analysis Enhancement of Faded Map Grid Lines Image here of Library object spectral image

Herblock Cartoons

Spectral image data is important for assessing the impact of fugitive materials, tracking changes in condition, and scholarly understanding of the author/artist process. The Herblock foundation donated the entire collection of editorial cartoonist Herbert L. Block (Herblock), including 14,400 of his daily published cartoons and 50,000 rough sketches. The bequest to the Library required that some of the collection must be on exhibit at all times, and Herblock began using light sensitive materials in the 1970s. The media he utilized included contemporary materials such as porous point pens, whiteout, blue and red artist pencils, other felt-tip pens, and Avery labels, on a paper board substrate containing optical brighteners. In 2009, PRTD began a long-term longitudinal study of the conservation of contemporary media and substrates. Exposure to lighting while drawings were on exhibition was an issue, but conservators and curators had also observed fading of a selection of some of the drawings while in dark storage.



Figure 3. ."Isnt This Better" 2001 Identification of Felt-tip Pen Inks

To assess changes occurring in Herblock collection materials, and accurately identify the source of the media involved in these changes, sample sheets of all the media used were created and aged, both naturally and with a range of accelerated aging techniques. This encompassed both light and dark aging to determine the mode of degradation, with analyses before and after successive aging periods to create a database of reference samples that could then be compared with the original object to identify the specific felt-tip pens that had changed in color after various aging methods. Successful matching of reference sample data with Herblock drawing changes and additional data captured in continued studies will allow preservation of the originals and add to knowledge of the creation techniques.

The integration of these new capabilities into the spectral imaging system at the Library of Congress enhances the Library's ability to conduct non-destructive analyses of collection items. Using these tools for the establishment of integrated access and presentation of scientific and scholarly data advances the development of a spectral reference library of scientific samples to support standardized identification of materials with scientific rigor to confirm the identification of materials. This initiative is developing layers of data that augment scholarly and curatorial research, aid library and researcher investigations of history and science, while preserving and protecting the original documents.



Figure 4. Clockwise from Top left; Felt-tip Pen Reference sample sheet, spectral curve from original faded ink, aged samples in visible and infra-red allowing identification of ink

Reference Database

The development of a PRTD spectral reference database of scientific samples with data from advanced spectral imaging techniques at the Library of Congress poses digital archiving, access and collection management challenges. The data expands the utility of non-destructive scientific analyses for cultural heritage research, yielding enhanced preservation data and greater access to lost information hidden in non-visible and obscured areas on the object. This requires transitioning the Library's scientific data to a reference archive for broad access, including data from collection objects such as drafts of the Declaration of Independence and Gettysburg Address, maps, drawings and prints. In addition to the spectral reference database, other materials have been collected to expand the dataset to encompass the range of materials styles, as well as various scientific instrumentation utilized in PRTD. The sample database has been titled the Center for Library Analytical Scientific Samples (CLASS), which comprises the physical reference samples. The CLASS Digital data files generated from these samples and collection objects are referred to as CLASS-D. Samples are tested non-invasively on a range of instrumentation so complementary data pertaining to the same sample can be generated, stored and accessed.

To test the CLASS-D database, PRTD developed a prototype of the application using example entries from the PRTD reference sample collection with representative example scientific analyses. These included both sample and analysis records, where a sample was designed as a record of a research sample, such as a Forbes pigment sample, or a collection object, such as the L'Enfant Plan of the city of Washington. An analysis record with data was designed to correspond to a single instance of a test or analysis, such as a single paper tensile test or a single XRF analysis of an object. The five samples tested for the prototype included a Forbes pigment, a Barrow Book, recycled paper samples used in an accelerated aging test, an audio compact disc from an aging test, and the L'Enfant Plan of the city of Washington. The primary goal of CLASS-D was to encourage interaction across and ease of use of data. Based on the prototype, some changes were made to the data model and application. To date a large portion of the extensive physical collection data has been transferred to the CLASS-D database.

It was apparent that the analysis record needed to be simplified to accommodate a single analysis or a group of related analyses of a single type (such as a logical group of Fourier Transform Infrared spectra for an object). Each analysis record needed to have a single structured archive of data files associated with it that complied with strict standards for structure and content. A tool was also needed to ensure that the database could accept bulk uploads of XML data for large data sets (such as the Barrow Books collection). Significant discussions were held with users and potential users both internal and external to the Library. These users included scientists, conservators, curators, researchers and information specialists. Based on their feedback, it is clear that this initiative towards standardization of scientific research data collection will help ensure CLASS-D database compatibility.

Continued developments in spectral image acquisition and access have allowed the Library to refine and standardize not only data formats and metadata, but also work processes to develop the spectral imaging system as an efficient laboratory tool.



Figure 5. Creation of Aged Pigment Samples to Aid Preservation of Historic Maps and Manuscripts

Spatial Standards

The Library uses the "Archimedes Palimpsest Metadata Standard," (APMS) which has been used on a range of spectral imaging projects at the Library and elsewhere, despite the eponymous name based on the original cultural heritage project using an initial spectral illumination and collection system.[6] Based on Dublin Core and the Federal *Content Standard for Digital Geospatial Data*, it includes "Section 2: Spatial Data Reference Information or Coverage: the spatial extent or scope of the data set".[7] This allows users to define the exact location of every artifact on a cultural heritage object, just as geospatial standards are used to identify points on the globe with Geospatial

Information Systems (GIS). This is necessary not only to register locations in image data from each of the spectral bands, but also to link other data with the spectral image data.

Scriptospatial Data

Just as latitude and longitude are used to define points on the globe that are linked to images from earth resource satellites with *geospatial* systems, a Cartesian coordinate system is used to define points on each object that are linked to points on the images *scriptospatial* systems (Fig 1). This standardized method provides links between the object, various images and other data.



Figure 6. Scriptospatial Imaging of Manuscripts based on GIS Model

Scriptospatial mapping of documents with an accurate coordinate system allows the layering of scientific and scholarly analyses to the DCO. This allows inferences to be drawn to generate new knowledge through analysis of the data linked to spatial points. This approach to viewing the DCO applies a GIS methodology toward uncovering and interconnecting information layers of cultural heritage artifacts, just as in the case of archeological strata. Utilizing an object-oriented approach in conjunction with the spatial data layers allows the mapping of spatial and temporal data with increasing complexity. Examining and explaining the physical, spectral and chemical properties of these historic materials permit scientists and scholars to link these scientific analyses to other data about the creation of the object.

The Spatial Data section of the APMS (Table 1) defines the spatial metadata elements that must be included to provide the links between locations on the cultural heritage object and on the image of the object, based on those cited in the *Content Standard for Digital Geospatial Data*. Each element then defines the specific coordinate system and measurements of the specific location on the image and original object. This is currently used just for two-dimensional objects, but elements for the z axis are included to capture metadata about cockling of the surface of manuscripts, as well as application to three-dimensional objects.

Mature spatial technologies at the micro level can be used to "scriptospatially" link a variety of scientific data from spatial points on the object, and at the macro level to link the object to the geographic area where it was developed, fabricated, stored and displayed. The goal is to integrate the spatial data into a GIS interface adapted to cultural heritage needs. At one level, the GIS display could link the scientific data collected from specific spatial areas on the object with a spatial reference to each data collection point. On another level, the GIS interface could link the object to the geographic location where it was used, displayed and/or manufactured in traditional mapping fashion.

Table 1: Spatial Data Section of Metadata Standard

2.1	Coordinate Units: unit type used in quantitative spatial metadata
	DOMAIN: "nivol" "continutoro" "millimotoro"
	DOMAIN. pixel, centimeters, minimeters,
	"micrometer"

- 2.2 X Resolution: number of resolution elements per coordinate unit in the x-direction CORE: YES TYPE: SINGLE DOMAIN: REAL NUMBER
- 2.3 Y Resolution: number of resolution elements per coordinate unit in the y-direction CORE: YES TYPE: SINGLE DOMAIN: REAL NUMBER
- Upper Left XY Coordinates: upper left coordinate of the limit of imaging expressed in resolution units.
 CORE: YES TYPE: COMPOUND DOMAIN: x= n and y= n
- 2.5 Lower Right XY Coordinates: lower right coordinate of the limit of imaging expressed in resolution units. CORE: YES TYPE: SINGLE DOMAIN: x= n and y= n
- 2.6 Bounding Coordinates: the limits of coverage of the complete data set for the entire sample expressed by Cartesian coordinate values assigned by the x:y table software in the order upper left, lower right. CORE: YES TYPE: COMPOUND DOMAIN: REAL NUMBER
- 2.7 Grid Coordinate System: Orientation of and definition of Cartesian coordinate so that spatial positions can be readily transformed to and from plane coordinates. CORE: YES TYPE: COMPOUND DOMAIN: TEXT
- 2.8 Vertical Coordinate System Definition: the reference frame or system from which vertical distances (altitudes or depths) are measured. CORE: NO TYPE: COMPOUND DOMAIN: TEXT
- 2.9 X-Positional Accuracy: an estimate of accuracy of the horizontal positions of the spatial objects. CORE: YES TYPE: SINGLE DOMAIN: REAL NUMBER
- 2.10 Y-Positional Accuracy: an estimate of accuracy of the horizontal positions of the spatial objects. CORE: YES TYPE: SINGLE DOMAIN: REAL NUMBER
- 2.11 Vertical Positional Accuracy: an estimate of accuracy of the vertical positions in the data set. CORE: NO TYPE: SINGLE DOMAIN: REAL NUMBER

The spatial links between data allow analyses that provide meaningful scientific outcomes of the content: Retrieval of obscured or faded text; Characterization of inks and pigments that can be traced to specific geographical locations; analysis of the intensity of handwriting to understand the author's original intent; and gleaning the provenance and source of paper through the capture and analysis of the watermark. A continued focus on collaboration between people, data and processes is a major factor in promoting access and integration of scientific research through a scriptospatial system of linking data for analysis and study.

PRTD System Integration

To support archiving of digital products with consistent quality, PRTD has established standard workflows for data management. These address the imaging challenges encountered with cultural objects; fragility, security, impact of previous treatments and unique properties of collection materials. Work processes have been automated and incorporated into the integrated image capture and processing. An additional challenge is capturing metadata from processing and manipulation of the image datasets to support and supplement the image acquisition data. This is a critical component to effectively link the "digital cultural object" with the original in terms of characterization of materials and new features revealed from the spectral imaging. This includes spatial metadata, which serves as the key link for the integration of data from other scientific instruments and scholarly studies.

The PRTD Spectral Imaging Project required development and integration of the following capabilities to effectively integrate the imaging, metadata capture and data integration:

- *Image Capture:* Efficient work processes and spectral imaging technologies to apply standard imaging methods to digitizing US cultural treasures as part of the spectral imaging program. This includes some automation of the capture sequence so that data and metadata collection is accurate, efficient and cost-effective.
- **Data Management:** Efficient work processes and technologies for PRTD and Library personnel to manage, store, archive and access the image data and data products with integrated metadata.
- *Image Processing:* Standard technical methods and tools to assess the quality of the raw images captured, and to ensure accurate integration of collected and processed images with and metadata into standard image products.
- **Quality Control:** Methods to review the processed images and metadata, creating a feedback loop within the imaging process so that the image data and metadata meet PRTD standards.
- **Dissemination and Access:** Storage and availability of the digital data, both within the Library and externally, with appropriate PRTD information technology systems.

Data Preservation

A key element in development of a mature imaging system is the ability to transfer and archive digital data for the curatorial, scholarly, preservation and imaging communities for research and analysis. Archived datasets are created with clear access procedures, to ensure data integrity for digital storage repositories, preventing introduction of mislabeled and incorrect metadata.

Overall management of large volumes of image datasets, with independent verification and validation of data and metadata, is critical to the successful integration of spectral imaging as a tool to gain new scientific information about library items, with verified archiving of raw and original file formats. The spectral data archiving is guided again by NASA with the *Reference Model for an Open Archival Information System (OAIS)* developed by the Consultative Committee for Space Data Systems (CCSDS). [8]

OAIS serves as the impetus for the PREMIS (*Preservation Metadata: Implementation Strategies*) standard, which implements the OAIS reference model for digital data preservation.[9] The OAIS model is intended for an integrated system of receiving, storing, and disseminating data with data sets that are self-documenting and allow for access and discoverability in both near- and long-term time frames.

Preservation of digital data in ever-changing storage media, software systems and protocols is a continued problem for the cultural heritage community. As Internet pioneer Vint Cerf noted, "What happens when files are there and we don't know how to interpret them anymore?"[10] PRTD and the Library follow several standards and work practices developed during previous spectral imaging programs starting with the Archimedes Palimpsests to try to offer better potential for the long-term storage and retrieval of archived spectral data. These include:

- Embedding all metadata in the EXIF headers of the TIFF files to ensure the metadata is retained with the images they represent;
- Including all imaging, processing, spectral and spatial metadata in accordance with Archimedes Palimpsest Metadata Standard;
- Conformance with the PREMIS data model and data dictionary, based on the Reference Model for an Open Archival Information System (OAIS).

Conclusion

The challenges of linking data in cultural heritage databases stem from the large range of data that can be acquired and the volume and size of datasets that can be generated from spectral imaging and other scientific instrumentation reference sets. The development of a spectral reference database of scientific samples with data from advanced imaging techniques links closely with the physical sample set of naturally and accelerated aged samples that generate information critical to the preservation of Library and cultural heritage collections. Expanding the utility of nondestructive scientific analyses for cultural heritage research enables both preservation and scholarly data to be linked and interpreted. It also provides greater access to lost and obscured data, not only in reference samples, but also data from collection objects that create insights into the author intent and provenance of documents. The object oriented approach through the "scriptospatial" system based on geospatial standards allows representation of the original document, while supporting effective integration of standardized scientific and scholarly data, layered with spatial, temporal, cultural and historical data. While this approach develops layers of data that augment scholarly and curatorial research, the digital component and the cultural heritage object greatly enhance access while preserving and protecting both the data and the original documents.

Acknowledgments

The authors would like to thank R. Douglas Emery of Emery IT for his disciplined and tenacious data management of this and so many other spectral imaging programs, Dr. Bill Christens-Barry of Equipoise Imaging LCC for his support in establishing the *Archimedes Palimpsest Metadata Standard*, and the other imaging

scientists and members of the spectral imaging teams who advanced the imaging techniques and systems for more than a decade. They would also like to thank the staff and management of the Library of Congress Preservation Division for their continued support and patience.

References

- [1] National Aeronautics and Space Administration "Landsat 1" http://landsat.gsfc.nasa.gov/about/landsat1.html 12 Jan
- [2] F.G. France, W. Christens-Barry, M.B. Toth, K. Boydston, "Advanced Image Analysis for the Preservation of Cultural Heritage", Proc. IS&T/SPIE Symposium on Electronic Imaging, San Jose, California, Jan (2010)
- [3] W. A. Christens-Barry; K. Boydston; F. G. France; K. T. Knox; R. L. Easton, Jr.; M. B. Toth, "Camera system for multispectral imaging of documents" San Jose, California, Proc. SPIE, vol. 7249 pg. 724908-724908-10 (2009).
- [4] F.G. France, M.B. Toth, "Spectral Imaging for Revealing and Preserving World Cultural Heritage",19th European Signal Processing Conference (EUSIPCO'11), Barcelona, Spain, pg. 1450-1454 (2011)
- [5] France, F.G. "Advanced Spectral Imaging for Noninvasive Microanalysis of Cultural Heritage Materials: Review of Application to Documents in the U.S. Library of Congress," Applied Spectroscopy, 65/6 June (2011)
- [6] Archimedes Palimpsest Program, "Archimedes Palimpsest Metadata Standard 1.0" (APMS), Revision 5. 2006, http://archimedespalimpsest.net/Documents/External/, Baltimore, Maryland: Walters Art Museum, 7 June
- [7] Federal Geographic Data Committee. FGDC-STD-001-1998. "Content standard for digital geospatial metadata" (revised June 1998), http://www.fgdc.gov/standards/projects/FGDC-standardsprojects/metadata/base-metadata/v2_0698.pdf (1 Dec 2008)
- [8] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System", 2002, Washington, DC, http://public.ccsds.org/publications/archive/650x0b1.pdf (1 Dec 2008)
- [9] PREMIS Editorial Committee, Library of Congress, "PREMIS Data Dictionary for Preservation Metadata" Version 2.0, 2008, Washington DC, March, http://www.loc.gov/standards/premis (1 Dec 2008)
- [10] J. King,, "The Grill: Google's Internet Evangelist Vint Cerf on the Hot Seat" Computerworld, July 30 (2007), pg. 20-22

Author Biography

Fenella G. France is the Library of Congress, Preservation Research and Testing Division Chief. She received with Distinctions her Bachelor of Applied Science and Masters in Textile Science from University of Otago, New Zealand, and her MBA from Deakin University, Australia. A Fellow of the American Institute of Conservation, Dr. France has focused on developing spectral imaging and other non-destructive techniques to enhance preservation and access to collections objects, including Jefferson's Declaration of Independence Draft.

Michael B. Toth is President of R.B. Toth Associates, founded in 1981 (rbtoth.com), which enables the use of new technologies in the study, preservation and display of cultural objects. Mr. Toth has supported scientific and cultural institutions across the United States and Europe, and his work is cited in the book The Archimedes Codex and numerous press articles. He has over 25 years of program management experience with the US Government and other institutions.