Sustainable and Accessible Integrated Heritage Science Datasets

Fenella. G. France; Library of Congress; Washington, District of Columbia, U.S.A

Abstract

A current challenge in the scientific, humanities and cultural heritage realm is the storage of and access to the increasing volume of digital datasets, paramount to advancing knowledge and assuring the future of research. The European Union has established a number of research infrastructures focused on addressing access and sustainability of digital data, while initiatives in the Unites States appear less integrated. The establishment of standardized digital protocols for storing and accessing scientific cultural heritage data is critical to ensuring interoperability between heritage institutions, and the preservation of international culture in libraries archives, and museums. The Preservation Research and Testing Division of the Library of Congress has moved forward with an initiative to standardize and make accessible, data from a range of scientific instrumentation, including related metadata files and assuring open access file formats.

Introduction

Advances in technology and digital access have improved utilization and interpretation of scientific analyses for cultural heritage and humanities studies. Integrating scientific and curatorial knowledge; moving from the focus on Science, Technology, Engineering and Math (STEM) to Science, Technology, Engineering Art and Math (STEAM) is a critical multidisciplinary approach, and necessary to add and link content knowledge to the original object through digital surrogates. The increasing volume of data collection and the need for effective interpretation of this data is a major challenge for data analytics, and critical to the future success of a sustainable structured approach to the collection of accessible heritage science data.

The need for true multidisciplinary research has never been more prevalent than in the field of cultural heritage. Here, the integration of scientists, curators, conservators, information technology and data management specialists, and other research specialists, is necessary to ensure accurate interpretation of collected data. One example that outlines this integrated approach is the development and advance of customized spectral imaging. Hyperspectral imaging has evolved as an important tool not just for accessing previously inaccessible and obscured content information, but a truly non-invasive method for digital documentation in its ability to map and track spectral and spatial changes in condition, identify and characterize substrates and colorants, and link the mapped chemical data with other non-invasive analytical techniques. Spectral datasets are a good example of the large volume of data that can be quickly collected, the size making access complicated, and the range of processing software techniques that increase the number of iterative images created [1]. Further, spectral and other instrument datasets are usually not linked, being reported in separate publications, and this dispersion provide challenges to heritage professionals in their ability to link and interpret disparate data files.

The basic underlying problem with accessing and integrating heritage science is that of non-standardized data formats, complicated by the resistance from manufacturers to provide open access data files and formats and linked instrument metadata. Most researchers can only share data with other users utilizing the exact same instrument software, making collaborative efforts unwieldy, and encouraging the constant "reinventing of the wheel" with people replicating previous efforts to gain access to useful data. Heritage objects are rarely pristine, have varying histories of environment and treatments, and defy the concept of standardized materials. While reference datasets are critical for interpretation and characterization of heritage materials, the rich and untapped personal datasets of heritage scientists contain significant levels of preservation data, inaccessible due to the lack of a platform or method for ease of sharing the data.

Development of a Sustainable Integrated Heritage Science Dataset Model

The concept behind this initiative was to create a "database" to integrate scientific preservation data from international libraries, archives, museums and other heritage institutions. This initiative would necessitate an open exchange of standardized scientific data, the utilization of open technologies and data standards to ensure broad access and application sustainability. It would also support the Library of Congress's goals of international access and sustainability through; non-proprietary file standards for image and text data, the establishment of standards and protocols for rigor in scientific practice and data collection, and the use of a flexible data model. This Library initiative had been titled the "Center for Library Analytical Scientific Samples – Digital" (CLASS-D), but given the desire to include a range of colleagues it would be more appropriate for this acronym to change to the "Center for Linked Analytical Scientific Samples – Digital" (CLASS-D).

Discussions over the past seven years with colleagues nationally and internationally, made it apparent that there was a willingness to share data, but not the time or resources to explore what a shared web-accessible open source database of heritage science would look like. Research into existing databases, both within the heritage and scientific fields explored what approaches were being used, and the limitations and advantages of these datasets [2]. It was apparent that no central system had been developed for integrating data from unrelated instruments, with each dataset focusing on access to, and standardization of, that specific data; whether infrared spectroscopy (IRUG), Raman spectroscopy, Mass-Spectroscopy, microscopy, or X-ray diffraction patterns. Heritage scientists constantly employ a range of types of instrumentation to re-create and uncover centuries of history. Given the unknown history of most heritage materials, one single scientific technique cannot provide all the required data and information, hence the need for a coordinated approach to linking access to datasets for data related to the same heritage object or materials.

Since then, a number of comprehensive reports have evolved, one being the Smithsonian Institution report on sharing digital biological data [3]. Out of this approach at the Library of Congress, the Center for Linked Analytical Scientific Samples – Digital (CLASS-D) developed, beginning the approach by standardizing the capture of data from reference samples characterized by a range of scientific instrumentation, before moving on to utilize the approach for the more complicated task of linking larger datasets to cultural heritage objects. The plan was to develop an open source software architecture and platform, initially modeled on a customized resource description framework, to allow addition of modules, with international access to data with data interoperability and standardized file formats. Through characterizing a wide range of reference heritage sample materials; including sample replicates that were either new, naturally aged, or had undergone accelerated aging, a collection of physical materials (reference papers, books – Barrow collection, pigments, leather, stone, fibers, modern media, etc.) would then have linked digital files from a range of different analytical methods (hyperspectral images, FTIR, Raman, XRF, SEM etc.). A challenge for heritage scientists is that conservation documentation usually focuses on treatments rather than access to actual scientific research datasets, a critical reason for addressing the need.

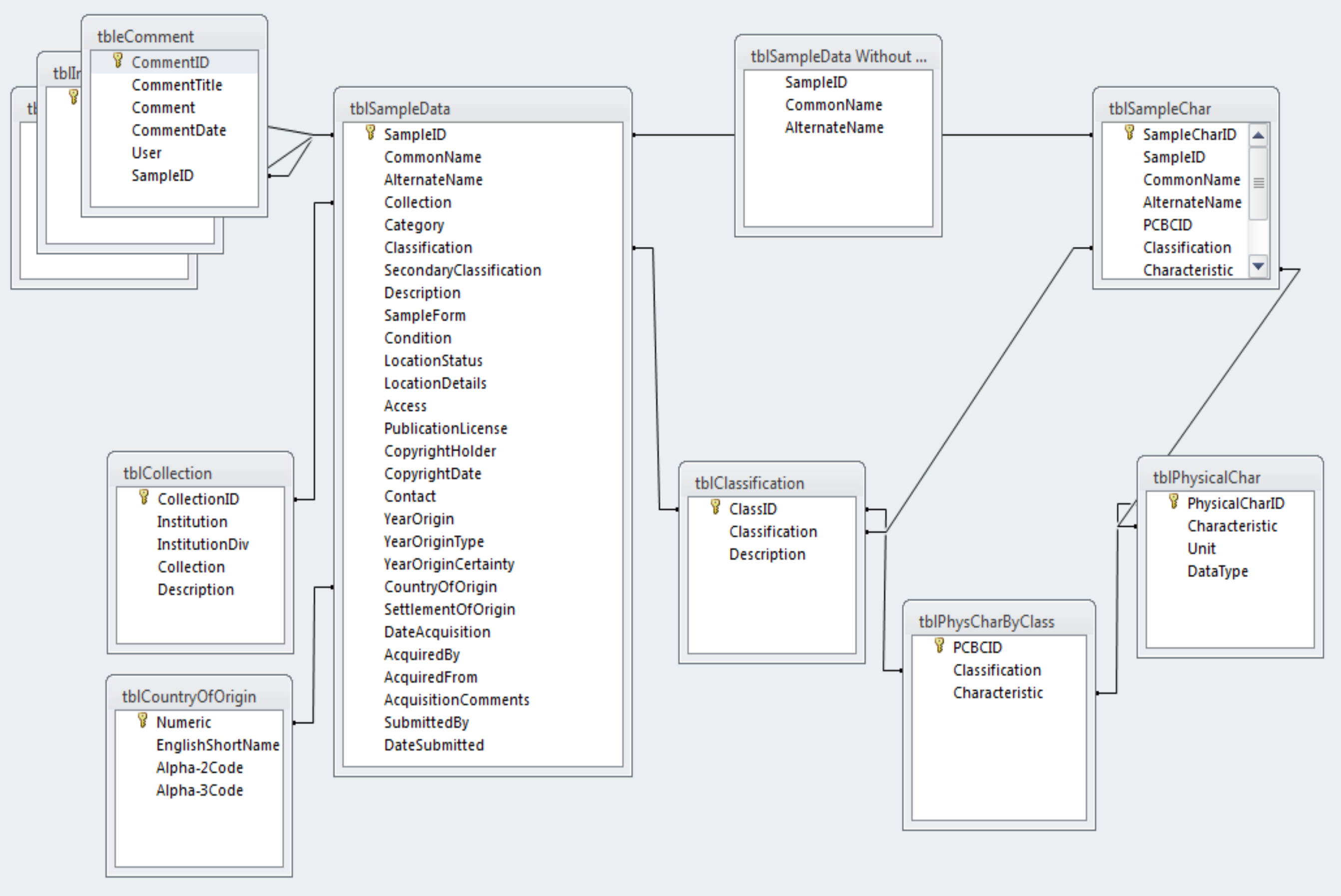


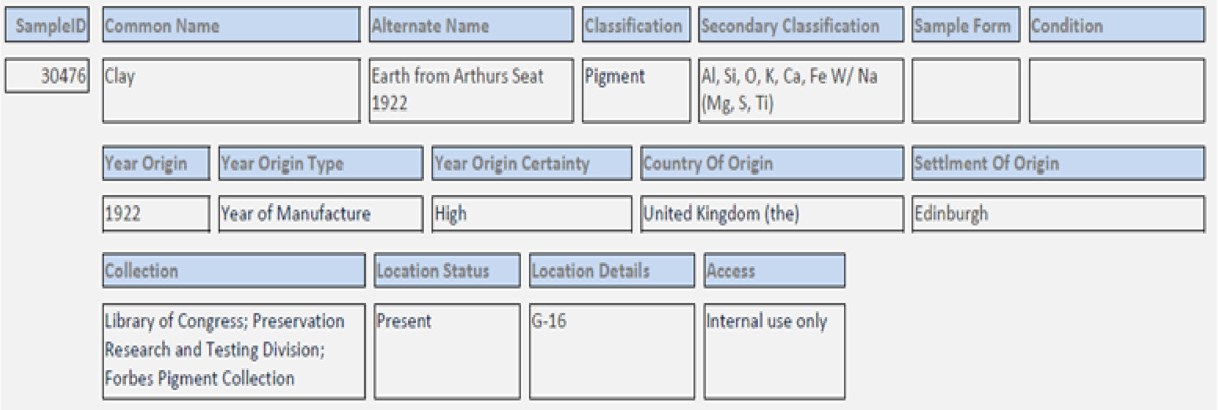
***Figure 1**. CLASS-D Database Architecture*

Results

Initial challenges revolved around determining what levels of metadata were critical, and what could be left open to allow incorporation of extant datasets, while ensuring a robust model database that would meet the needs of diverse institutions. An initial prototype proved cumbersome, and a simplified model was developed with over 1900 records loaded into the dataset to assure its applicability to various material types.

One of the key factors was linking specific data with material type (book, paper, pigment, fibers etc.) to allow only those required fields to show, but allow each material sample to have the specific properties needed to characterize and identify. Consistent nomenclature for samples and analyses in the database was imperative, and the naming structure needed to ensure that names were consistent with names of other database components. Having consistent names for components has several advantages; it is easier to determine the nature of a component when it has a name that conveys the meaning and use of the component, accessing and finding a component is easier when it is named predictably, and it is easier to create a name for a component when clear guidelines exist. It was also important to ensure ease of search-ability by having data terms reflect the real world. Creating a standard for data sharing means the facilitation for unambiguous understanding of database elements and schemas, and to do so, the names and structures need to represent and model the informational aspects of objects and concepts that users are most familiar with. This initiative attempted to ensure that elements did not model collections of data or institutional biases.

The use of authority tables has been incorporated and will greatly facilitate future movement, and this is an area of heritage science that needs extensive collaborative effort.



SampleID	Common Name	Alternate Name	Classification	Secondary Classification	Sample Form	Condition
30476	Clay	Earth from Arthurs Seat 1922	Pigment	Al, Si, O, K, Ca, Fe W/ Na (Mg, S, Ti)		

Year Origin	Year Origin Type	Year Origin Certainty	Country Of Origin	Settlment Of Origin
1922	Year of Manufacture	High	United Kingdom (the)	Edinburgh

Collection	Location Status	Location Details	Access
Library of Congress; Preservation Research and Testing Division; Forbes Pigment Collection	Present	G-16	Internal use only

***Figure 2**. Example of Data Fields for a Reference Material*

Expanding the data model to incorporate instrumentation also raised the challenge of temporal components; how to link data that had been analyzed multiple times and natural aging environmental data, adding in accelerated aging instrumentation and conditions, and including assessment of treatments for digital documentation. Capture of instrument and analytical technique metadata (rather than relying on the lab notebook) was also complicated by the lack of non-proprietary file formats on some instruments, and the need to probe deeply into the underlying software structures to extract linked metadata. In many cases to simplify extraction a simple text file was created and linked to the analysis file. Other issues included determining what formats were most appropriate for viewing in the database format, for example with spectral imaging datasets, jpegs were linked as the viewing platform with a link to the full dataset stored on another server.

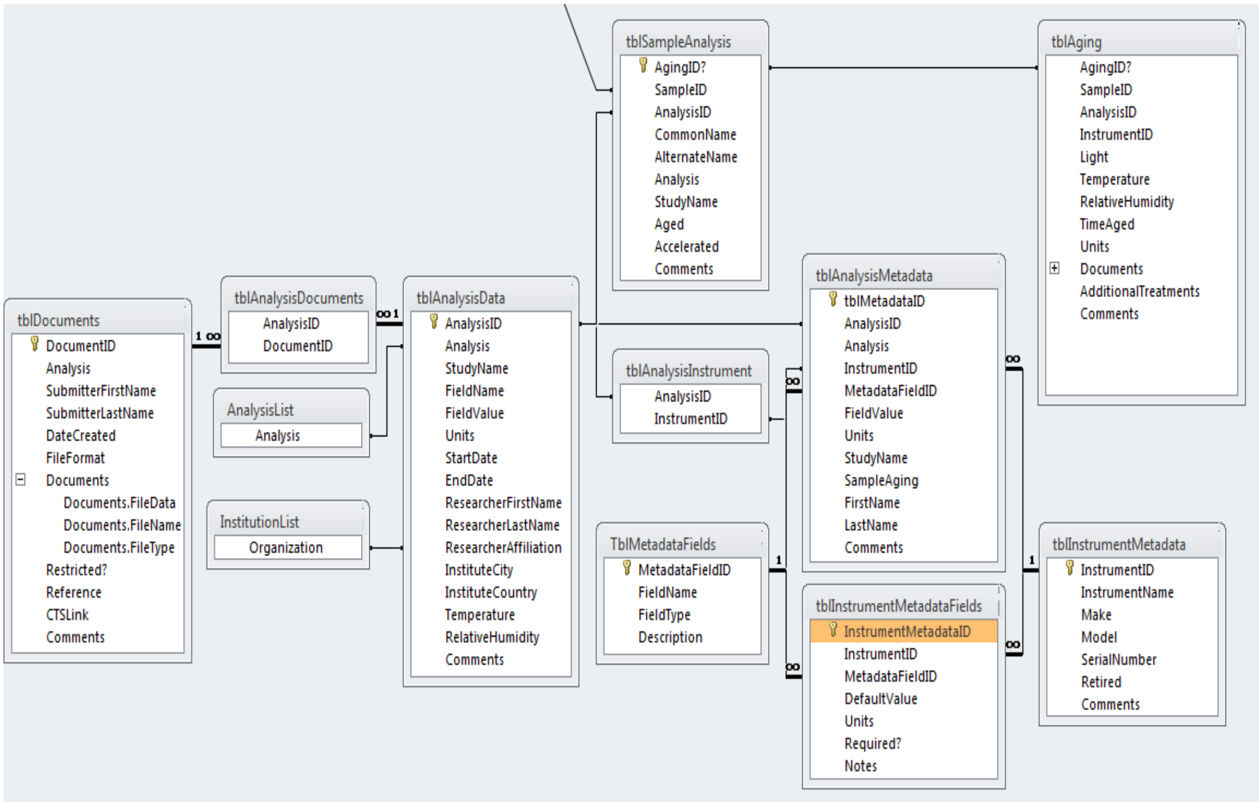


***Figure 3**.Database Schematic for the Inclusion of Scientific Analytical Results*

To encourage other institutions to contribute, it will be critical to have scripts to generate XML from CSV files, or Access database records, allowing a simple bulk upload of data. A single central repository, that links reference samples, analyses, metadata, and associated data files is a major component of the above schematic, to ensure relevant data is linked accurately. The success and sustainability of the database also requires careful structuring for the capacity to search and access any data needed for preservation research. It is hoped that as this initiative expands and becomes more fully accessible through the new platform, that it will encourage high standards for the collection of scientific data, metadata, and datasets, including the linking of protocols and data specifications to each database record. Another vital component for sustaining high quality heritage datasets is the integration of the workflow within the data model, with a structured validation and approval process for verifying accuracy of submitted data by qualified experts.

Data Access and Interface

The concept of scriptospatial mapping of data enables direct sharing and visualization of data to support heritage analyses, with the integration and linked capture of standardized instrumentation parameters and object metadata. This scriptospatial concept greatly enhances the ability to share data to effectively support cross-disciplinary research collaborations and analysis. Examining and explaining the spectral, optical and chemical properties of heritage materials with scriptospatial tools, while linking this with curatorial knowledge, permit scholars to relate these scientific analyses to the social context of how they were created and used. These relationships support valuable and innovative creative approaches to data integration, while strengthening effective art and scientific collaborations.

Essentially, the term scriptospatial refers to the development of an object-oriented approach to data access requiring the integration of data from other sources (in a variety of formats) projected onto a digital image of the heritage object. This approach requires effective spatial metadata to allow linkages to specific locations within the image, or images if the image of the object is linked through a spectral image data cube. The spatial metadata is necessary not only to register locations on the same section of a manuscript leaf in various spectral bands, but also to link other images and transcriptions with the spectral images. A camera collecting images over a heritage object is similar to a satellite collecting geospatial data over the Earth. Using technologies developed for "geospatial" systems to link each point on the globe with images from earth resource satellites and data collected from other scientific analyses, spectral imaging can link the "scriptospatial data" from each point on an object or manuscript with images and data from various scientific instruments. This method provides a standardized method to support links between images and data from the same location on the object.

With multiple data entries for samples, precisely defining the specific point where the sample or scientific data analysis or collection takes place is critical in comparing data from different research types or objects. For samples (non-invasive and invasive) taken from a larger, non-uniform, heterogeneous object such as a manuscript, textile or painting, the spatial location of the sample point on the object must be defined to be able to integrate the data from various research tools. Spatial metadata elements will allow linkages to specific locations on an object, potentially within images of the objects. Scriptospatial data can serve as an interface for scientific dialogue in "one shared layer," linking data from various sources for in-depth studies and analyses of a specific research topic or object.

In addressing the preservation science challenges for sharing scientific data from diverse instruments, institutions and research goals, it is important to look at the challenges faced in another discipline three decades ago. In highlighting the importance of "Open Geospatial Information Systems (GIS)," ESRI's 2003 Spatial Data Standards and GIS Interoperability White Paper highlighted the progress made by the GIS community, which faced many of the same challenges:

"*In early years, the constraints of computational speed and cost limited our ability and caused us to focus on practical solutions such as direct file conversion. Data sharing between organizations with different GIS vendor systems was limited to data converters, transfer standards, and later open file formats. Sharing spatial data with other core business applications was rarely achieved. Today, most GIS products directly read and sometimes dynamically transform data with minimal time delay. The point here is that the GIS community has been pursuing open interoperability for many years, and the solutions to achieving this goal have changed with the development of new technologies.*

"Another factor to be considered is the still evolving view of the role that GIS plays in an organization. In the early days of GIS, the focus, with rare exceptions, was on individual, isolated projects. Today the focus is on the integration of spatial data and analysis in the mission-critical business processes and work flows of the enterprise and on increasing the return on investment (ROI) in GIS technology and databases by improving interoperability, decision making, and service delivery.

"Finally, it is worthwhile to remember why we implement geographic information system technology in the first place. Even if we have specialized responsibility for gathering and managing geographic data, we need to remember that a GIS is not an end in itself. A GIS must produce useful information products that can be shared among multiple users, while at the same time provide a consistent infrastructure to ensure data integrity. It is important not to get caught up in the technology and forget this basic principle. Interoperability enables the integration of data between organizations and across applications and industries, resulting in the generation and sharing of more useful information."

The progress and successes of the GIS community over what is now three decades in establishing an open architecture in which diverse data types can be integrated are dependent on standard interchange formats and open file formats. One can look at progress made to date with metadata standards in the preservation science community as a similar foundation for development of an open scientific data sharing architecture. With the introduction of spatial metadata to common standards, the preservation and heritage science community can expand and adapt from what is now 30 years of collaboration and standardization by the GIS community to rapidly develop an open "scriptospatial" architecture. The cultural heritage and preservation community can capitalize on the investment made by earth science, national security and defense organizations and contractors into common systems and standards for sharing spatial data.

In many current research databases, the metadata elements for spatial location are not provided to capture detailed data on where an instrument collects data, or a sample is taken. This is not an issue for uniform, homogeneous samples of paints, pigments, media or other samples, but is critical for samples taken from a heterogeneous object like a painting, manuscript or textile. By defining a Cartesian coordinate system on an object or image of an object, as well as the degree of precision required, specific sample points on an object can be defined. This allows integration with other images of the same object and scientific samples from the same point. The Content Standard for Digital Geospatial Data (FGDC-STD-001-1998) serves as a basis for defining metadata elements for cultural heritage and scientific research. Use of this standard will also allow use of geospatial software and systems to manage and integrate "scriptospatial data" from object samples. This will provide a standardized method to support links between various samples from the same sample point or object by a range of users.

Discussions with international colleagues have advanced potential future developments for this initiative through the push for global research infrastructures. Currently, the European Union (EU) has begun developing infrastructures in both the science and humanities fields, but not yet integrated the two. Utilizing existing linkages with the EU such as the Research Data Alliance (RDA) working groups seems a logical progression, given that RDA group outcomes are focused on tangible acceleration of progress for global data sharing and increasing data-driven innovation (https://rd-alliance.org/).

Conclusions

Collaborations with international colleagues has demonstrated the willingness to integrate and establish research infrastructures that are truly multidisciplinary, and to do so, there needs to be an organization that can support access and sustainability of the platform, have standardized open access data that is easily accessible both through a robust working platform to integrate and research the datasets, and the capacity to use this heritage science reference data as a source for addressing challenging heritage research problems. As was noted above, for effective linking of heritage and humanities data, the interface is critical.

Developing and implementing the CLASS-D initiative will allow participating institutions to leverage existing investments into research equipment, infrastructure and information systems by building the semantic bridges to connect the data and metadata from their research. CLASS-D metadata standards will enable different institutions and data systems to share and exchange information, irrespective of the research equipment and methodologies used by each institution. The definition and maintenance of CLASS-D metadata and data collection standards and protocols ensures that cultural heritage institutions and supporting industry partners can reap significant cost benefits through adoption and reuse, rather than building proprietary, single-use research techniques and data collection methodologies from the baseline. The further distance we can move from reliance on closed access proprietary systems of files and software for scientific data collection, the grater our ability to collaborate, integrate and access heritage data and datasets.

With the focus on standardized data elements this database initiative allows data sharing and preservation by making research data available to users for efficient and intuitive search and discovery. Users will then be able to discover and reuse existing heritage science data that meet their research requirements, and extract and reuse specific content to support their own research. Development of CLASS-D as an effective data sharing capability will require implementation of four key practices outlined above; an effective governance model, standard metadata elements, structured machine and human readable language for sharing, and the ability to map standard metadata elements across institutions and partnerships. With partner institutions supporting these practices, CLASS-D will serve not only as an effective tool for storage, dissemination, and searchable access to standardized heritage science data and datasets from global research, but also a framework for the preservation and access to additional data that can enhance research beyond the source institution or scientific equipment

References

1] C. Ferrari, G. Foca, and A. Ulrici., "Handling large datasets of hyperspectral images: reducing data size without loss of useful information", Analytical Chimica Acta, 802, (2013) 29-39.

[2] F.G. France, D. Emery, D., and M.B. Toth, "The Convergence of Information Technology, Data and Management in a Library Imaging Program", Library Quarterly special edition: Digital Convergence: Libraries, Archives, and Museums in the Information Age, Vol. 80, No. 1: 33-59 (2010).

[3] Smithsonian Institution, "Sharing Smithsonian Digital Scientific Research Data from Biology," Office of Policy and Analysis, Washington D.C, March (2011).

[4] AWP Research; "Charting the Digital Landscape of Conservation," Survey Results – Foundation of the American Institute for Conservation of Historic and Artistic Works, August (2014).

[5] ESRI White Paper, "Spatial Data Standards and GIS Interoperability", (January 2003) http://www.esri.com/library/whitepapers/pdfs/spatial-data-standards.pdf (accessed March 30, 2015).

[6] Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf (Accessed March 30, 2015).

Author Biography

Dr. France, Chief of the Preservation Research and Testing Division at the Library of Congress, researches spectral imaging techniques and addressing integration and access between scientific and scholarly data. An international specialist on environmental deterioration to cultural objects, her focus is connecting mechanical, chemical and optical properties from the impact of environment and treatments. Serving on standards and professional committees for cultural heritage she maintains collaborations with colleagues from academic, cultural, forensic and federal institutions.

.