https://doi.org/10.2352/issn.2168-3204.2017.1.0.39

Advances in Integrated Research Infrastructures for Science and Humanities Linked Data

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

Abstract

The continued challenge for data in any discipline is sustainable access, open source file formats, and the capacity for linked data. Collaborations with European and American colleagues indicates a shared concern, but with a less focused effort for establishing and recognizing the need for a more integrated approach to truly linked data, and the need for high level metadata embedded within datasets. Many related fields and disciplines have begun to focus on the need to integrate and assess approaches from colleagues – from materials science to archeology, botany, biology and chemistry. The Research Data Alliance (RDA) has brought together a more cohesive approach to data management on the global scale. Developments for linked scientific data generated on heritage materials has continued to develop within the Library of Congress Preservation Research and Testing Division has engaged with colleagues in RDA and internationally to build upon existing standards and authorities, allowing greater credence for humanities and cultural heritage linked data. Further developments in the CLASS-D database structure enable the unique capability to link a range of types of scientific instrumental analyses back to original source materials, track samples and derivatives over time, and further the capability for web-accessible access to heritage collections.

Introduction

Access and interoperability of data are critical elements for any database initiative, with many challenges being that while there is lip service given to "open access", often a deep understanding of the full requirements to achieve this are not fully understood until the completion of a project. The establishment of standardized digital protocols for storing and accessing scientific cultural heritage data is vital for interoperability between heritage institutions, and the preservation of international culture in libraries archives, galleries and museums. Many institutions also fail to consider requirements of other institutions when they build supposedly interoperable structures, and the careful design of a robust system is necessary to its longevity.

The Preservation Research and Testing Division of the Library of Congress (PRTD) continues to develop an initiative for a shared web-accessible database of heritage materials and associated reference samples to standardize and make accessible, data from a range of scientific instrumentation. This structure has incorporated careful attention to the integration of related metadata files, open access and sustainable file formats, high level metadata for searchability for data, the ability to include and bulk upload extant datasets, and competence in building a structure that is flexible enough to take account of needs of partner institutions even when these needs may not have been apparent in the initial phases of the database architecture. This streamlined approach to exposing and linking scientific data sets that relate back to one heritage object is a powerful new use of previously separated data components, and adds value for data mining and seeing trends in seemingly unrelated heritage materials.

Developments in Integrated Research Infrastructure

Research infrastructures provide a significant position in the advancement of all fields of knowledge and the way new and relevant technologies can be adapted and applied to increasing our knowledge of heritage institutions. As noted, it is critical that these structures recognize the diverse range of disciplines, researchers and stakeholders to involve in order to help shape and link scientific and scholarly communities [1]. The current data deluge has been overwhelming in all fields, especially as we integrate analog and digital systems and related information, and can be daunting when trying to anticipate the needs of all users and integrate foresee future requirements. This has been the situation in previous discussions with international colleagues, with the biggest stumbling block being how to know when to keep revising and perfecting components of the infrastructure and when to forge ahead.

Discussions with colleagues has included engagement with extremely diverse audiences. Through the role as Chair of the United States – Italy Bilateral Agreement on Cultural Heritage Further meetings there have been extensive discussions with European colleagues from a range of heritage related infrastructure initiatives. These European digital infrastructures that preserve and provide access to heritage data, include the Digital Research Infrastructure for the Arts and Humanities (DARIAH), the Integrated Project for the European Research Infrastructure ON Cultural Heritage (IperiON CH), the Advanced Research Infrastructure for Archaeological Data Networking in Europe (ARIADNE) Project, the Collaborative European Digital Archive Infrastructure (CENDARI) Project, and PARTHENOS – "Pooling Activities, Resources and tools for Heritage E-research Networking, Optimization and Synergies. The integrated European Research Infrastructure for Heritage Science (E-RIHS) moved forward with discussions about the need for links to the Research Data Alliance in regards to legal interoperability standards for sharing of research data. The distinct difference between European and US digital and heritage science funding and support led to further engagement with US colleagues in RDA. This engagement and interaction with colleagues and interest groups in the Research Data Alliance expanded the potential use of CLASS-D scientific research metadata for integration with related fields of materials science, chemistry, heritage, humanities and other science disciplines. The Research Data Alliance is a research community organization started in 2013 by the European Commission, the American National Science Foundation and National Institute of Standards and Technology, and the Australian Department of Innovation https://www.rd-alliance.org/ . There are more than 4300 members (111 countries in September 2016) and the alliance

provides a neutral space for members through focused global Working and Interest Groups, to develop and adopt infrastructure that promotes data-sharing and data-driven research. This also led to recognition of, and interaction with associated initiatives in the fields of archeology and biomedical infomatics [2]. The latter, through the Center for Expanded Data Annotation and Retrieval is studying the creation of comprehensive and expressive metadata for biomedical datasets to facilitate data discovery, data interpretation, and data reuse, and is engaging in a very similar process of development and discovery.

As outlined in a previous presentation, the first step for PRTD after engaging in lengthy interactions with colleagues, was the recognition that its easier for potential users to interact with and review a prototype rather than a "concept". Phase one of the Center for Library Analytical Scientific Studies – Digital (CLASS-D) was the knowledge that linked subsets of data for each heritage material type and object needed to be established before associated data and user interfaces could be included. The first objective was to more fully organize, catalog, and assign unique identifiers to samples within the CLASS collection, in order to standardize and better manage the collection, and the second objective to continue the development of a robust and accommodating database for the CLASS-D initiative. The design of the database feasibly needs to also incorporate analyses from a range of research projects, instruments, sample mediums, and institutions. To facilitate the goal of the CLASS-D database to have linked open data, adherence to a common vocabulary was critical to ensure commonality and increase interoperability. A barrier to interoperability is a lack of standards, as experienced by natural history museums in Europe, who have been attempting to share scientific data [3]. An in-depth assessment of cultural heritage institutions revealed the lack of capability or focus on linking data from a range of instruments and analytical techniques [4]. Other institutions link datasets across institutions, such as the Museum of Fine Arts, Boston, who created the Conservation and Art Materials Encyclopedia Online (CAMEO) database, but this functions as a reference database of materials cross-institutionally, without incorporating not incorporate the testing on the materials. Other cultural heritage institutions attempt to share scientific metadata. For the natural history museums in Europe, successful sharing is accomplished through enforcing standards for managing the scientific metadata.

Enhanced Database Functionality

Recent adaptations of the CLASS-D organization included an emphasis on tracking database objects and through this focus, specific three-level barcoding of the scientific samples or heritage objects. After assigning unique identifiers, a tri-level identifying schema was utilized, to incorporate the organization of broad and specific locations and items, then added item classification specifications, for standardization of CLASS collection labelling. Each item is classified by the collection it is a part of, as shown in table 1.

The barcode structure was critical to then allow subsets and derivatives from the original, or analyses repeated over time, to be linked and fully incorporated with the original item, as part of an ongoing research project, or multiple projects including different analyses or extended periods of time. The individual sample barcodes are listed in the table **tblSampleBarcode** with the corresponding **SampleID** linking the barcode to the rest of the information contained on the sample. The table **tblStorageLocations** functions as a comprehensive list of possible storage locations with the associated barcodes, as designated through the barcode schema. In order to link the sample barcode with the barcodes of its location, the table **tblStorageHierarchy** connects each sample with preset storage barcodes.

Table 1: Select Barcode Schema Chart

Collection	[NAME]	Barcode Example
Barrow Books	BBL	BBL00001
Modern Media Audio	MMA	MMA00001
Forbes Pigments	FORB	FORB00001
ISR Papers	ISR	ISR00001
TAPPI Fibers	TAP	TAP00001
Herblock	HER	HER00001
Parchment	PAR	PAR00001
Feathers	FEA	FEA00001
Ceramics	CER	CER00001
NIST	NIST	NIST00001

Incorporating research into the database structure required defining the components of the research; the research project; the instruments used for analyses; and file attachments pertaining to any aspect of the research. These sections were added separately, due to the complexity of the task. To add the relationships for the individual instruments and associated scientific analyses, a three table design was created to customize the metadata fields for each instrument. Predetermining the metadata fields for each instrument was important for data standardization, so that each researcher will be sharing, searching, and accessing the same information for each instrument, regardless of the institution or research project. A structure was built where the table **tblInstrumentMetadata** linked specific metadata fields with each individual instrument (table 2). A fourth table, **tblInstrumentAnalysis**, linked the instrument to specific research studies and samples.

Table 2: Instrument Metadata

Table Name	Data
tblInstrument	Comprehensive instrument list. Includes general information about each instrument.
tblMetadataFields	Instrument metadata fields undifferentiated by specific instrument
tblInstrumentMetadata	A union table that links instrument and metadata fields based on architect-identified relationships
tblInstrumentAnalysis	Information on the testing performed on a specific instrument for each sample in a study

Due to the multi-faceted nature of a research project, incorporating and linking research projects into the database required subdividing information into the following levels: overarching research project, sample specific research scope, and instrument specific analysis. The structure was built to accommodate the potential for numerous samples, testing, and instruments within a single research project. The table **tblResearchProject** addressed the broadest level of a research project and contained general but necessary information regarding the research, such as researcher, research affiliation, etc. The table **tblResearchScope** identified the samples used in the research project, allowing multiple samples to be associated with a single research project. This was one of the critical elements of the design and a unique component of the necessary robust yet flexible nature

of the architecture. Finally, the table **tblInstrumentAnalysis** linked each research sample to the instrument that performed the testing. The table structure allowed multiple samples to be linked to one instrument, or inversely, one sample to be linked to multiple instruments. As the database was intended to link samples to research, the design of the linking table was important. The relationship between the tables is shown in figure 1.

The fourth phase of additions incorporated the attachment component. Previous efforts to incorporate file attachments to the database utilized multiple file attachment tables, designating one file table in conjunction with the research project and placing other files directly into the specific tables that related to the file. However, this method resulted in duplicate files across multiple tables, which was undesirable. Instead, to reduce redundancies, we designed a structure that contains all files in one table, **tblFiles**. A secondary linking table, **tblFileLinking**, associates each file with the respective entry or table through relationships.

One consideration in implementing file attachment capabilities to the database was the goal of CLASS-D to enable and enforce standardization for data sharing. With the aim of standardization, file attachments were restricted to internationally-accepted, standard file formats. The design accommodated and enforced the standardization of file formats through permitting only five file formats: PDF/A, XML, TXT, JPG, TIFF.

To communicate the file format options, the table **tblFiles** contained a field "FormatType" with a dropdown list. When a new file record was added, the inputter needed to choose the file format from the predetermined list, without the option for writing in a new or different file format.

The final complexity of the phased infrastructure approach was integrating aging components into the database. The aging aspect of a research project adds an additional unique capability from the more simplified straight analysis and instrument combination. There is the possibility for multiple aging periods during the lifetime of a heritage object, and for predictive testing to assure longevity of our collections, the utilization of accelerated rather than natural aging further complicated the structure, relationships and linked components. The aging aspect being unique could not be treated in the same manner as an instrument or research scope. Tables were allocated solely to associating and integrate the aging component into the functionality of the database. A linking table between the aging table and the research scope table was created to correctly associate the aging with the sample and research.

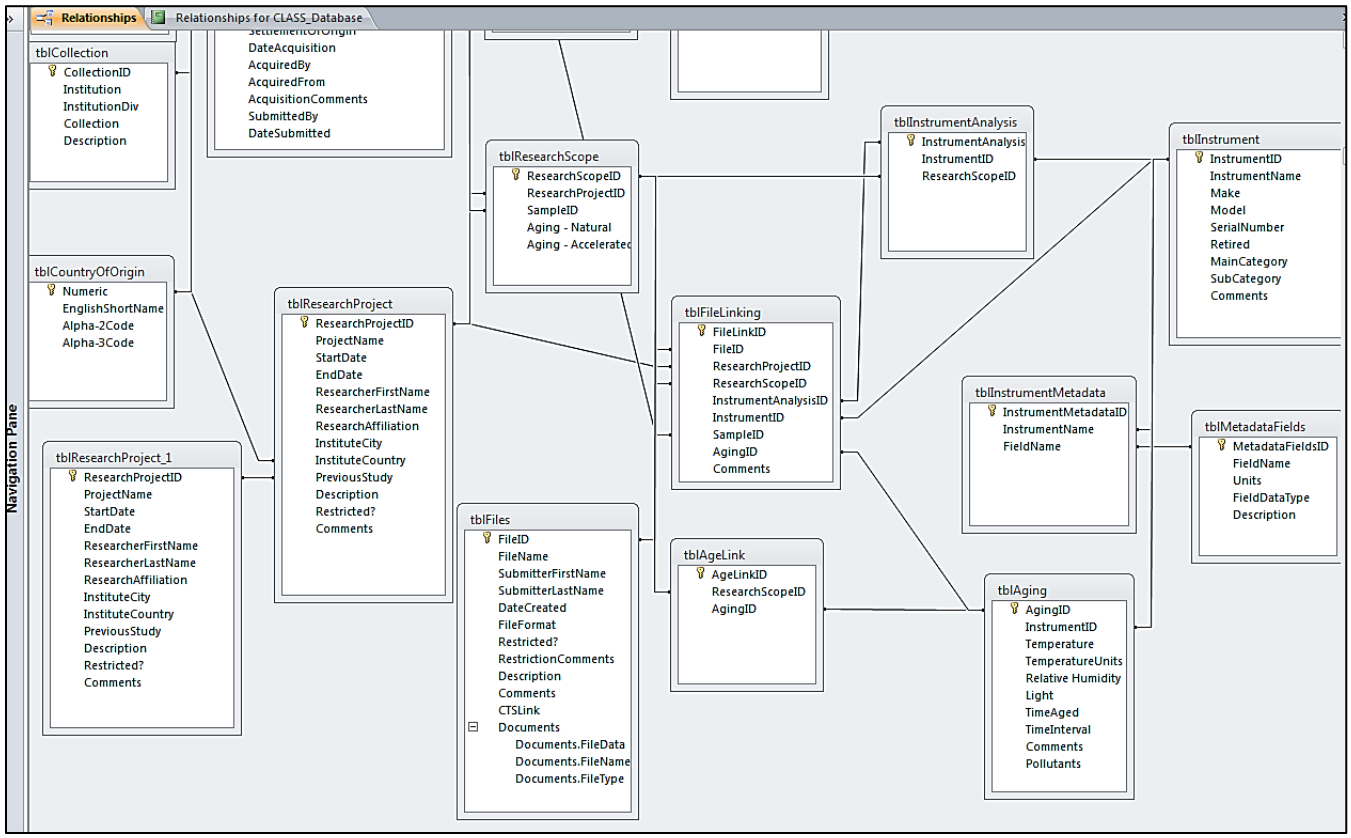

Figure 1. Relationship of Research, Instruments, Analyses, Files and Aging Data)

Figure 2. Examples of CLASS Reference Materials: (Clockwise from Upper Left: The Barrow Book Collection, Magnetic Tapes, Painted Samples, Tidelines of Books. Forbes Collection, Traditional Parchment Skin

The User Interface

Creating a visual capability for linking data allows for greater usability and interpretation of data, since trends can be more easily represented and envisioned. Scriptospatial representations of digital data refer to geospatially locating where specific analyses were undertaken on heritage objects. Through a rendering of the original object from using a "google map" rendering approach/view, documents can utilize an accurate coordinate system that links scientific and scholarly analyses to the creation of a new digital cultural object (DCO). The approach to viewing digital cultural materials in multiple layers applies an archaeological approach toward uncovering and interconnecting information strata of historic and modern documents [5]. 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 (or areas). This approach to viewing the DCO applies a GIS methodology toward uncovering and interconnecting information layers of cultural heritage artefacts, just as in the case of archaeological 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 [6].

In its Scriptospatial Visualization Initiative, the Library of Congress PRTD has capitalized on developments with geospatial systems to apply Thermopylae "i-spatial" support and Google Map tiling and data formatting to the integration of large and complex visual scriptospatial datasets populated with scientific data from various instruments, research topics or objects. This provides data access in "one shared layer" of scientific data. This is an important first step in capitalizing on the three decades of technology development by the GIS community to advance preservation science and cultural heritage data sharing and research. The additional unique component will be the layering of scholarly research interpretations and publications, enabling ease of access to a rich resource of data directly linked to the original object. This will reduce the challenges faced with searching through data that is not well catalogued, or yet searchable without this expanded document interpretation and linking of scholarly knowledge.

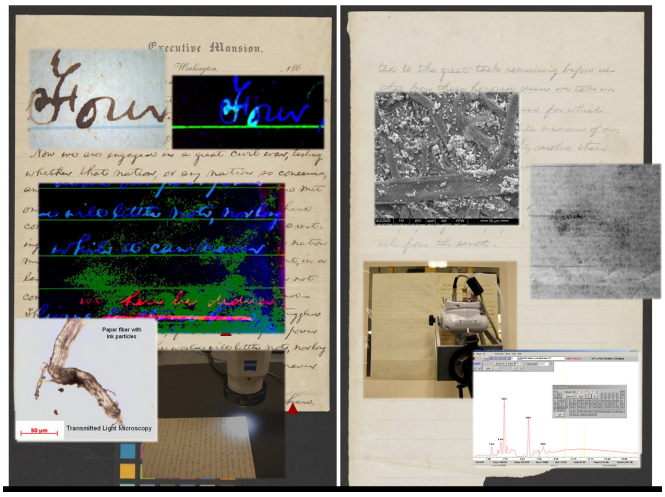

Figure 3. "Google-like Mapping" of the Document with Annotated Analyses

One of the interesting ways to move this user interface forward is also looking at the types of interfaces gaining widespread use such as the International Image Interoperability Framework (IIIF) http://iiif.io/ . The initiative notes that: "*Access to image-based resources is fundamental to research, scholarship and the transmission of cultural knowledge. Digital images are a container for much of the information content in the Web-based delivery of images, books, newspapers, manuscripts, maps, scrolls, single sheet collections, and archival materials. Yet much of the Internet's image-based resources are locked up in silos, with access restricted to bespoke, locally built applications".* This awareness of the need to support interoperability and standardization of access to information through an image-based interface focuses on the image and scholarly annotations. The shared canvas data model and image and presentation application

programming interfaces (APIs), are being assessed to see how far the extended required metadata for linked scientific data can be taken to use this increasingly commonly used framework for integrating scientific and scholarly data.

Conclusions

This multi-dimensional approach to the future of linked and integrated heritage and scientific data allows for greater interoperability of previously un-linked but related data while greatly expanding the current visualization of data. The object oriented approach through the "scriptospatial" system allows representation of the original heritage object, while supporting effective integration through the commonality of 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 the original documents. Continued collaborations and engagement with EU, US and RDA colleagues and initiatives will assist with greater integration of a standardized terminology, shared and coordinated metadata and authority definitions and procedures, and a truly integrated approach to linked data. The creation of a research infrastructure architecture that can incorporate the needs of diverse organizations and user needs allows for expanded utilization of library, archives and heritage institution information and data, while encouraging trans-disciplinary engagement and collaboration between scholars and scientists.

References

[1] France, F.G., Emery, D., and Toth, M.B., "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).

[2] Mark A. Musen, Carol A. Bean,1 Kei-Hoi Cheung, Michel Dumontier, Kim A. Durante, Olivier Gevaert, Alejandra Gonzalez-Beltran, Purvesh Khatri, Steven H. Kleinstein, Martin J. O'Connor, Yannick Pouliot, Philippe Rocca-Serra,Susanna-Assunta Sansone, Jeffrey A. Wiser, and the CEDAR team, "The center for expanded data annotation and retrieval", Journal of the American Medical Infomatics Organization, 2015;0:1–6. doi:10.1093/jamia/ocv048, Brief Communication, June 2015.

[3] G. Skevakis, K. Makris, V. Kalokyri, P. Arapi, and S. Christodoulakis, S. "Metadata, management, interoperability and Linked Data: publishing support for Natural History Museums", 2014.. *International Journal on Digital Libraries*, *14*(3/4), 127-140 2014. doi:10.1007/s00799-014-0114-2

[4] R. D. Emery, "CLASS-DB Prototype and Analysis" (internal LC report) pp55, 2011.

[5] France, Fenella G. and Toth, Michael B. "Integrating science and art: the scriptospatial visualization interface", IFLA WLIC 2014 – Lyon - Libraries, Citizens, Societies: Confluence for Knowledge, Session 149 - Art with Science and Technology Libraries (2014) In: IFLA WLIC 2014, 16-22 August, (2014, Lyon, France, http://library.ifla.org/id/eprint/763.

[6] France, F.G., Toth, M.B, and Hansen, E.F., "Challenges of Linking Digital Heritage Scientific Data with Scholarly Research: From Navigation to Politics", Digital Humanities, Kings College, London, July (2010).

[7] Mark A. Musen, Carol A. Bean,1 Kei-Hoi Cheung, Michel Dumontier, Kim A. Durante, Olivier Gevaert, Alejandra Gonzalez-Beltran, Purvesh Khatri, Steven H. Kleinstein, Martin J. O'Connor, Yannick Pouliot, Philippe Rocca-Serra,Susanna-Assunta Sansone, Jeffrey A. Wiser, and the CEDAR team, "The center for expanded data annotation and retrieval", Journal of the American Medical Infomatics Organization, 2015;0:1–6. doi:10.1093/jamia/ocv048, Brief Communication, June 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.