Integrating and Visualizing Humanities and Heritage Science Data

Fenella G. France
Preservation Research and Testing Division
Library of Congress
Washington D.C, United States of America
frfr@loc.gov

***Abstract*—The capture of digital data for cultural heritage includes multidisciplinary data types, analyses and formats from many diverse fields; including materials science, archeology, botany, biology, engineering, physics and chemistry, to name but a few. The continued challenge for digital data in any discipline is sustainable access and the capacity for a more integrated approach to linked data, including high level metadata to enable searchability. Many related fields and disciplines have begun to focus on the need to integrate and assess approaches from other scientific disciplines while also engaging with humanities colleagues who utilize the same information from different perspectives. An initiative for linked scientific data generated from heritage materials has been developed within the Library of Congress Preservation Research and Testing Division. This database integrates multiple scientific analyses all linked back to the original heritage object. For ease of access, a visualization interface integrating humanities and heritage science creates a "digital cultural object" with layers of integrated and linked data. These digital initiatives include, the Center for Linked Analytical Scientific Samples – Digital (CLASS-D) – an infrastructure enabling the unique capability to link a range of types of scientific instrumental analyses back to original source materials, expanding the capability for managing web-accessible access to heritage collections; and the Data Visualization Project (DVP) visual interface.**

Keywords—linked data, heritage science, visualization of scientific data

I. Introduction

The capture of digital data for cultural heritage includes a truly multidisciplinary gamut of data types, analyses and formats. Heritage science data is being captured from many diverse fields; including materials science, archeology, botany, biology, engineering, physics and chemistry, to name but a few. The continued challenge for digital data in any discipline is sustainable access, open source file formats, and the capacity for a more integrated approach to truly linked data, as well as the requirement for high level metadata embedded within datasets to enable searchability. Interoperability, the ability of systems or software to exchange and make use of data and information necessitates a standardized structured approach to data capture. Many related fields and disciplines have begun to focus on the need to integrate and assess approaches from other scientific colleagues – while also recognizing the necessity to engage and learn from the humanities, since the data being generated is simply viewed from different perspectives. An initiative for linked scientific data generated from heritage materials has been developed within the Library of Congress Preservation Research and Testing Division. A relational database integrating all different types of scientific analyses, links these back to the original heritage object, building upon existing standards and authorities. Additionally, for ease of access, a visualization interface integrating humanities and cultural heritage from an object-oriented approach creates a "digital cultural object" with layers of linked data. This initiative, the Center for Linked Analytical Scientific Samples – Digital (CLASS-D) is a structure enabling the unique capability to link a range of types of scientific instrumental analyses back to original source materials, tracking change over time, and expanding the capability for managing web-accessible access to heritage collections and research data [1].

II. Sustainability of Digital Heritage Data

Access and interoperability of data are critical elements for managing any digital heritage database, 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. Critical initial planning needs to include engagement with manufacturers to access non-proprietary file formats, discussion and agreement on high-level metadata for searchability, and a focus on sustainable file formats for long-term access. The establishment of standardized digital protocols for storing and accessing scientific cultural heritage data is vital for interoperability between heritage institutions, within and outside academia, 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, even more so, the desire to share and support research infrastructures without constantly "reinventing the wheel." The Preservation Research and

Testing Division of the Library of Congress (PRTD) focused on developing 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.

Sustainable shared research infrastructures can provide a significant advantage for transdisciplinary integration and the transformation of data to the advancement of all fields of knowledge. This includes adapting new and relevant technologies and applying these tools to increasing our knowledge of heritage institutions. 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. The current data deluge continues to be overwhelming in all fields of science and humanities, especially as we integrate analog and digital systems, primary and secondary sources, and can be daunting when trying to anticipate the needs of all users and 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 United States – Italy Bilateral Agreement on Cultural Heritage and additional 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) has moved forward with discussions about sharing research data [2].

III. Development of this Digital Heritage Initiative

The first step in this initiative, after engaging in lengthy interactions with colleagues, was the recognition that it was easier for potential users to interact with and review a prototype rather than a "concept". Phase one was understanding 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. The second objective was designing a structure to incorporate analyses from a range of research projects, instruments, sample mediums, and institutions. To facilitate the goal 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. 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. For the natural history museums in Europe, successful sharing was accomplished through enforcing standards for managing the scientific metadata.

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 multiple 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 would be sharing, searching, and accessing the same information for each instrument, regardless of the institution or research project. 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 ResearchProject 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 ResearchScope 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 InstrumentAnalysis 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. Ongoing work that will be presented will show how as we have been entering additional data, some of the initial relationships have been modified, for example the relationship between the original source and tests of extracted samples from the original materials.

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.

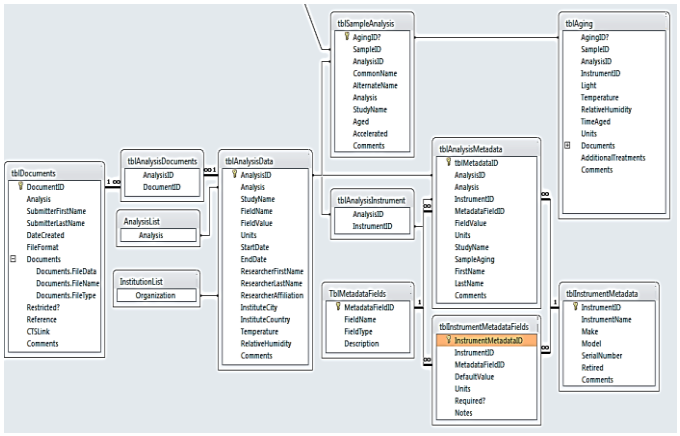

Fig. 1 Database Schema for Inclusion of Scientific Analyses

The interface for allowing scientists to include data was structured to incorporate dropdown selections and reduce the capacity for free text, since this create an enormous challenge for both interoperability and ease of searchability. Administrative controls were included to ensure rigor in data entry and accuracy. This interface is under review and being modified to allow for additional capabilities and extended ease of searchability across the database.

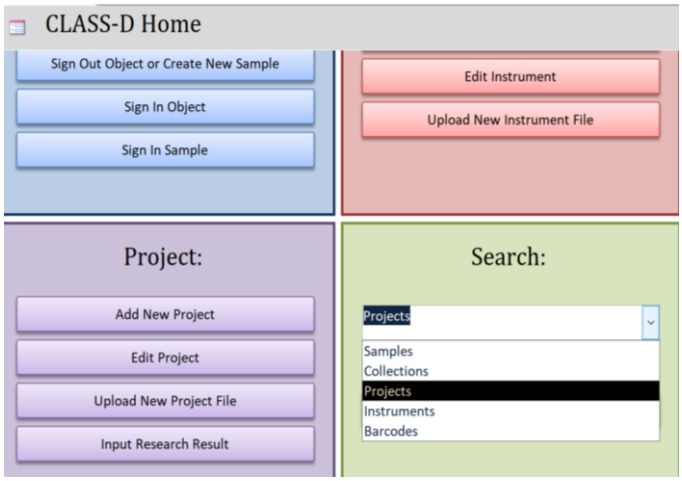

Fig. 2 Database Interactive Interface

The creation of a "User Interface" between the underlying database and a way of visually rendering and linking the data in a more interactive mode was critical for creating a visual capability to link data through a representation of the heritage object. 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), as noted above. 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. Scriptospatial mapping of documents 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 regions on the object. 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 need for this visual interface was to interconnect in a thoughtful and structured manner, scientific data and interpretations that allowed for a richer experience for a range of heritage professional, scholars and researchers. This ultimately was to create a system that could also allow researchers to share, annotate, and build upon the rich layering of information and create knowledge while working in a collaborative manner. Heritage related research questions could be posed and shared while data and perspectives were being examined, connected and potentially vetted, associated and revised. This could include aspects of heritage materiality such as; how has the material changed over time? can the data provide information about construction techniques? are pigments commensurate with the time period or geographic location in relation to provenance of the object, can other historic reports be linked and mined, what can new technologies and/or data mining add to the research?

A further problem with linking scientific data from a heritage object is the three components of spatial, spectral and temporal. Spatial and spectral are more easily rectified on an image or canvas, but the temporal aspect provides additional challenges. For example, mapping and tracking changes over time for analyzing preservation treatment options and impact, layering event-driven data and historical changes such as maps, or the impact of environment (whether storage or exhibit) add to the dimensionality required, as well as the richness of the linked data. An extension of the spectral and spatial is the conundrum of the rich point cloud data generated by 2D and 3D data sets, adding to the need for higher resolution, non-static images [3][4].

IV. Utilizing an Existing Infrastructure

While yet another structure could be created, it seemed imperative to begin by assessing the viability of investigating the capability of an existing framework to integrate scientific data. Engagement with the International Image Interoperability Framework (IIIF) had the function of starting the process working with an existing internationally recognized image sharing protocol that many institutions were using in some capacity [5]. While the existing capabilities of IIIF for including visualization of scientific data did not fully meet the needs of the data, it allowed for rich engagement with colleagues to begin the work, and look at what was needed to

adapt the canvas for the inclusion of not just heritage, but potentially many other scientific data layers. IIIF has the mission that the access to image-based resources is fundamental for the diffusion of cultural knowledge, as well as research and scholarly communications. So many web interfaces are based upon the use of digital images as a surrogate for heritage and other research materials, including photos, books, newspapers, manuscripts, maps, music and archival materials. Through IIIF, an expanding community of the world's leading research libraries and image repositories has embarked on an effort to collaboratively produce an interoperable technology and community framework for image delivery. Since this is being widely used in many national and international repositories such as the National Gallery of Art, Yale University, Europeana, DARIAH, etc. this was a good framework to build upon, rather than trying to create something entirely new. IIIF states the goals to "*give scholars an unprecedented level of uniform and rich access to image-based resources hosted around the world; ...define a set of common application programming interfaces that support interoperability between image repositories, ... and to develop, cultivate and document shared technologies, such as image servers and web clients, that provide a world-class user experience in viewing, comparing, manipulating and annotating images"*. These goals aligned with the challenges and needs of linking and visualizing the connections between scientific and humanities data. At the IIIF international conference in June 2017 a group formed to discuss and expand the incorporation of scientific data into IIIF applications and in March 2018 further engagement with the Research Data Alliance (RDA) community highlighted the need for interoperability between not only digital heritage data, but expanding this to other scientific disciplines, an easy extension for heritage science since the field is already multidisciplinary [6].

Utilizing IIIF with the Mirador viewer gave the options of integrating plugins for other programs, expanding the existing capabilities of Mirador 2.0. For example, progress to date allowed for stacks of images taken at multiple and different wavelength (multispectral imaging) to be incorporated into a IIIF layer slider function (created using Seadragon) that allows the user and researcher to engage with the data, and easily navigate between multiple renderings of the same data, exposing the spectral response of different materials, construction techniques, and in some cases uncovering previously redacted text.

Additional functionalities have been incorporated for this slider. These functionalities allow the user to move through and compare only the wavelengths they are interested in. A continued focus on the user experience has been a large component of this development. The user community will encompass scientists, preservation specialists, curators, researchers and potentially other groups, since we are crating a dashboard that the user can customize for their experience.

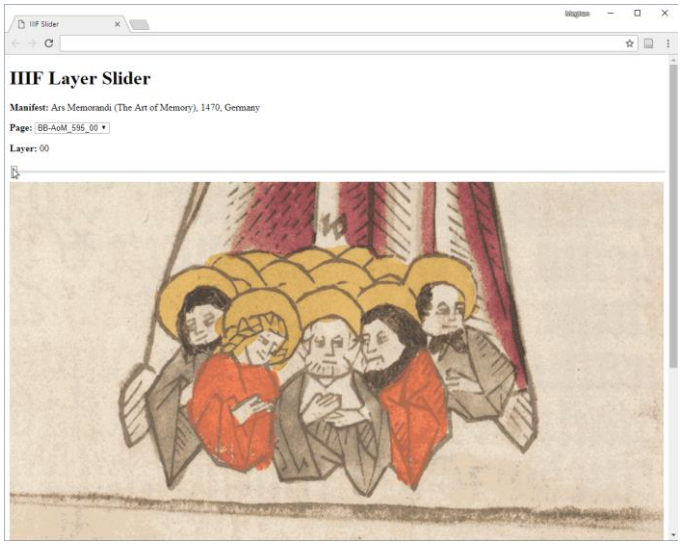

Fig. 3 Interactive Interface for Multispectral Datasets using IIIF

The use of layers was more unwieldy for interacting with multiple types of analyses done on the same document, and in this case, scientific analyses were added as annotations to a color image of the document, page or object. As can be seen in figure 4, colored boxes are layered onto the canvas image of a fifteenth century block book illustration. As the user hovers over that box the text explanation of the analysis appears, with this explanation referring to the data in a manner that intends to create interest for a non-scientist as to why it is important for the materiality of the document. Clicking on that hyperlink brings up the actual data as another page. In the example below both a high magnification microscope image and X-ray fluorescence spectra are included as data points that help understand and reflect the questions from a curatorial perspective about how the block book illustrations were created, and the palette of the pigments and colorants used.

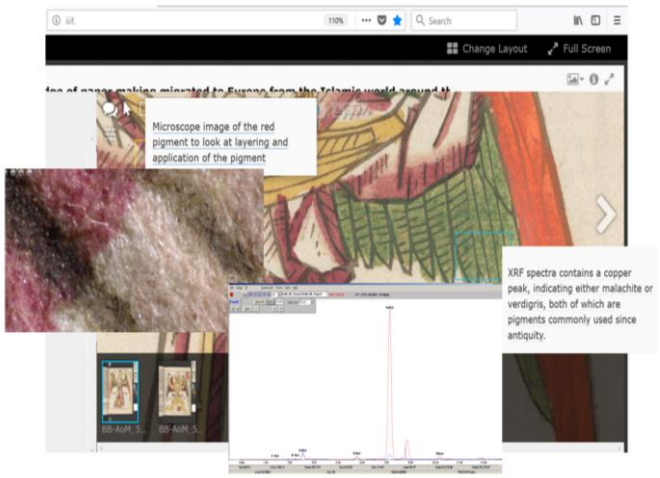

Fig. 4 Integration of Data as Annotations to the IIIF Canvas

As we added additional layers of information it became apparent that there would be difficulties with the "deep learning" – initial spectral mapping followed by multiple

analyses captured at the same location – to build up knowledge about the materials, for example inorganic, organic and morphological information, and/or answer the specific research questions.

Fig. 5 Expanded Data Annotations

Clicking on the annotation box (fig 5) brings up the description (with linked authoritative data descriptions for e.g. pigments and analytical techniques) and the additional option to "View all Scientific Data" opens a further interactive layer with the descriptions, and analytical spectra included (fig 6).

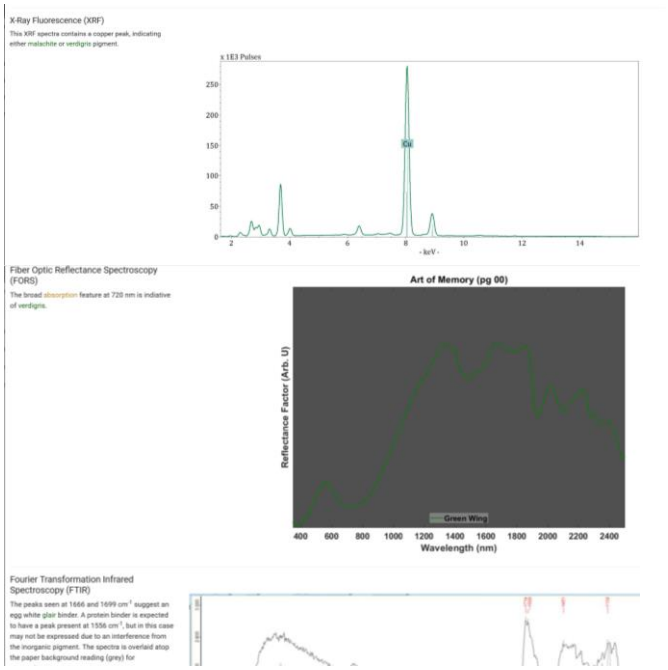

Fig. 6 Integrated Scientific Data Portal

V. Conclusions

Continued progress has exposed a number of challenges with data integration, but the initial template has proved useful and informative for creating a dialogue between heritage scientists and humanities scholars and researchers. It is imperative that this expands the capabilities for digital heritage to engage and inform all heritage professionals.

Acknowledgment

F.G. France acknowledges the support of Glen Robson, IIIF and PRTD staff Meghan Wilson and Chris Bolser.

References

[1] F.G. France, "Online Scientific Reference Sample Collections and Shared Linked Data for Heritage Science and Related Disciplines", Proceedings of the Coalition of Networked Information (CNI), Albuquerque, NM, April 2017.

[2] France, F.G., "Advances in Integrated Research Infrastructures for Science and Humanities Linked Data", Imaging Science and Technology, Riga, Latvia, May 2017.

[3] R. Pillay, R., "IIPImage and an Analysis of JPEG2000 Encoding Parameters", Wellcome Trust, London, 10th November 2014 http://www.dpconline.org/docs/miscellaneous/events/1358-2014-nov-jp2k-ruven/file.

[4] E. Bertin, R. Pillay, and C. Marmo, "Web-based visualization of very large scientific astronomy imagery", Journal of Astronomy and Computing, Elsevier, Vol. 10, p43-53, 2015.

[5] International Image Interoperability Framework http://iiif.io/

[6] IIPImage http://iipimage.sourceforge.net.