https://doi.org/10.2352/issn.2168-3204.2019.1.0.35

Linked Open and Annotated Science and Heritage Data

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

Abstract

The challenges involved in assuring the longevity and validity of heritage science data, and access to that data, require that all scientific data terminology and experimental procedures are not proprietary and have a common meaning across heritage institutions. While the temptation to create a new set of thesauri and definitions is great, it merely exacerbates the "siloed" impact that tends to separate rather than aggregate colleagues and data. Using IIIF and the Mirador viewer to integrate scientific and scholarly data about heritage objects, it became apparent when attempting to create a cohesive structure that terms in common use amongst one group of users were not necessarily familiar to the others. Therefore, easily accessed but rigorous controls on terms needed to be put into effect, with preference deliberately given to reusing existing resources.

Introduction

The Preservation Research and Testing Division at the Library of Congress has been developing an infrastructure for sharing and visualizing scientific and curatorial data relating to cultural heritage objects [1]. The underlying database, the Center for Library Analytical Scientific Samples – Digital (CLASS-D) allows for the inclusion of multiple complementary analyses to be linked back to the original object, while the visualization interface links and annotates the rendering of cultural objects through IIIF and the Mirador viewer with both heritage science and curatorial data.

A critical component of the infrastructure was authoritative linked open data (LOD) that enabled users to quickly understand and interpret the meaning of the scientific analyses. While annotations were compiled to assist curatorial users and viewers, colleagues accessing the scientific data needed to be able to compare what they were viewing with their own instrumentation, and know that, for example, the term "irradiance" was being used in the same manner as they would for their similar instrument, even if they used slightly different software.

An examination of existing databases revealed that they did not in general include this feature, so an investigation was undertaken to see what authoritative sources existed, and how these could be easily integrated without creating yet another in-house thesaurus. It became clear that many supposedly "LOD" sources had significant challenges in the volume and expanse of data the was available as well as whether these data were interoperable and did in fact crosswalk between related heritage, science and humanities ontologies, thesauri and other terminology-related websites and online resources.

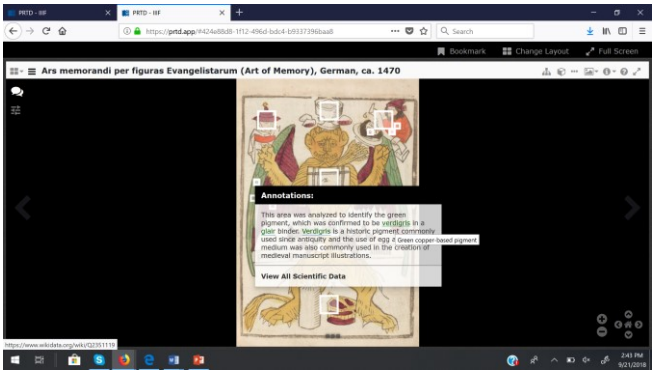

***Figure 1**. IIIF Cultural Heritage Image rendered in Mirador – annotated for scientific data*

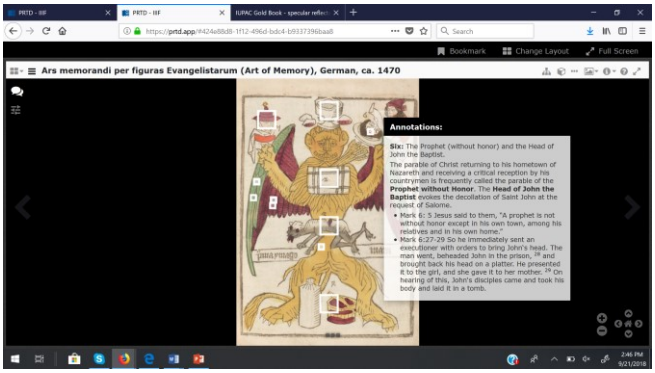

***Figure 2**. IIIF Cultural Heritage Image rendered in Mirador – annotated for curatorial data*

Data Management and Visualization

There is a rich volume of data about the materiality of cultural heritage that gets lost and ignored, and is never made easily accessible to researchers. This data could greatly expand their knowledge of materials; how was the book constructed? What is the parchment made of and therefore what is the country of origin? What are the pigments and date of creation? Who collaborated in the X century to print this? Many of these questions can be analyzed and answered through commonly available scientific techniques, and yet the data is still not being made accessible to scholars. The motivation behind this advancement of visualizing scientific data on a common API such as IIIF was to utilize a platform that was already familiar to scholars, and make it simpler for them to access and understand this additional heritage data.

The problem that became apparent, was not just creating the additional information on the annotated API to share this data, but ensuring that scholars not accustomed with reading and interpreting scientific data could feel comfortable engaging with it. To this end we developed an automated glossing tool that would clarify what specific scientific terms meant and link

to authoritative definitions, thus clarifying how these analyses related to data within their knowledge domains.

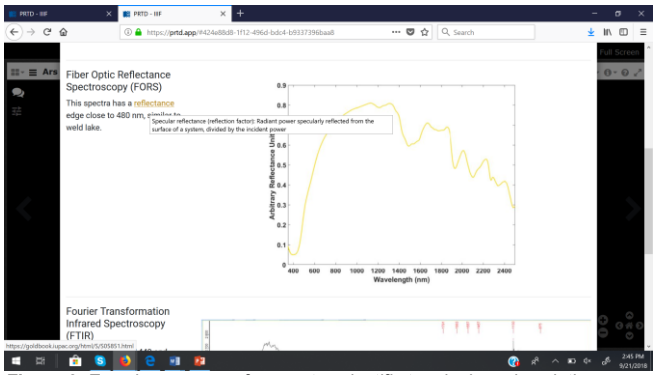

***Figure 3**. Easy hover-over reference to scientific terminology description*

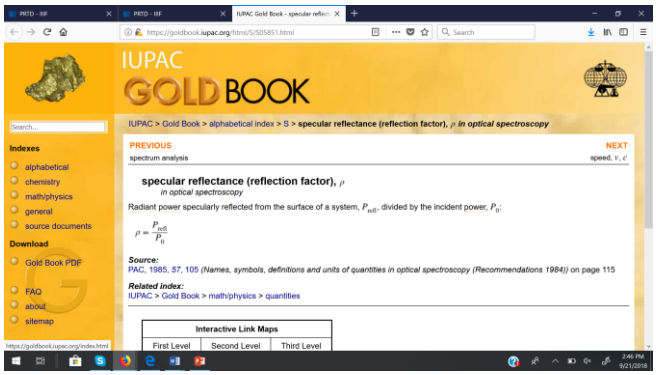

***Figure 4**. Linked reference to the original term in IUPAC book for chemistry, which is not itself directly usable as a LOD resource.*

While working with this data and their research, scholars and researchers could also be assured that there were authoritative sources through linked open data (LOD) – they would have direct access to the scientific terms, and could be confident that the terms in use were consistent within those in authoritative vocabularies. In addition, this visualization crosses disciplines so not only are we dealing with humanists, we are also attempting to include scientific nomenclature with the commensurate structure for linking scientific research processes. Discussions with national and international colleagues confirmed that the existing CIDOC CRM [2] structure would be outside the scope of a transferable, mobile approach.

Approach

Having a controlled vocabulary is critical to the success of linked open data because the structure can promote consistency of multidisciplinary and multilingual metadata, both within collections and across institutions, and just as importantly, it can increase external search linkages and capabilities. The many fields involved with cultural heritage objects have good reason for their own discipline-specific terminologies, and a high degree of variation within disciplines is perfectly normal, but collaboration and sharing of data fails if the many synonyms involved in a project cannot each be resolved to a single LOD term in a controlled vocabulary.

In order to share cultural heritage data via our web-accessible interface, it was necessary to control and standardize all terms and reconcile them with existing ontologies. Two ontologies were initially explored as part of working towards a linked open data platform: the Getty Art & Architecture Thesaurus Online (AAT) [3] and the International Union of Pure and Applied Chemistry Gold Book (IUPAC) [4].

The *Art & Architecture Thesaurus* (*AAT)* is a structured vocabulary, including terms, descriptions, and other metadata for generic concepts related to *art*, *architecture*, conservation, archaeology, and other cultural heritage topics. Included are work types, styles, materials, techniques, regions, periods, and others. The AAT is structured such that the terms for describing the ancient ceramics of West Mexico exist in a hierarchy that can also describe sculpture from the early or late Nara period in Japan. Using the Getty's AAT not only allows cultural heritage institutions to describe their works and processes in a consistent way, it enables reasoning about the relationships between those described works, materials, techniques, and so on.

The IUPAC Color *Books* are the *world's* authoritative resource for *chemical* nomenclature, terminology, and symbols. However, an API for using the Gold Book as an LOD ontology is one of several updates IUPAC have planned in order to modernize the resource, and it is not known when that might be [5]. The current online implementation of the GoldBook raised other concerns, given our goals. Most of the definitions assume a technical audience, while many defer viewers to other definitions which not infrequently defer them to yet further definitions. The terms rely almost exclusively upon images for formulae rather than machine-readable (and so searchable) formats. It should be added that the first two of those issues are perfectly understandable, given what IUPAC can reasonably expect from their intended audience.

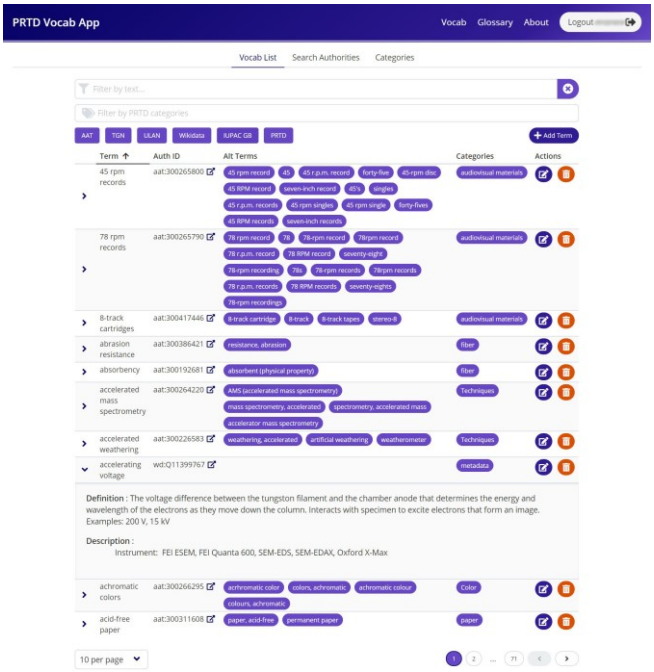

***Figure 5**. Initially we collated and evaluated suitable terms in the Getty vocabularies, Wikidata, and the IUPAC GoldBook online.*

This initiative explored using Wikidata's ongoing efforts to aggregate IUPAC GoldBook data, which by Wikidata's calculations at the time of writing represents approximately 19% of the GoldBook's terms [6]. Where there are terms available in Wikidata the preceding IUPAC GoldBook online concerns did not apply: users can find general and technical details, and cross-references to other LOD resources, while the entries themselves have been manually edited to avoid relying on images for symbols and equations. Wikidata's SPARQL interface also allowed us to query other aggregated chemical authorities, including PubChem, ChEMBL, and the FDA's

UNII [7]. That said, some Wikidata terms collapse multiple terms under one more general term [8], which rather defeats the purpose of using Wikidata as a proxy for IUPAC's GoldBook, and aggregation progress has slowed significantly since a coordinated effort in January 2018, averaging a dozen terms per month.

While the AAT in concert with conservation-oriented LOD initiatives resolve very nearly all of our non-scientific terminology requirements, and the Getty are responsive to receiving suggestions for additional terms, the results were far less satisfactory for preservation science needs. The terms available were too general or too specific in ways that would be misleading, or there simply were not any appropriate equivalent terms that met our needs. We turned to the well-established and ever-growing network of BioMedical LOD resources for accurate and internally coherent sets of terms for describing our techniques, instrumentation, and data sets. The Open Biological and Biomedical Ontology (OBO) Foundry [9] maintains and publishes over 160 active ontologies from developers committed to collaboration and adhering to shared principles. The ontologies address discrete areas of scientific research, but are designed to complement one another, avoiding internal conflicts.

This research found that between the OBO Foundry's Chemical Methods and Mass Spectrometry (ChMO and MS) ontologies we could accurately describe lab techniques and instrumentation to a high degree of specificity [10]. The intricately cross-referenced structure of the Chemical Entities of Biological Interest (ChEBI, European Bioinformatics Institute) ontology allowed us to use their terms to assign classes and functional groups to PubChem compounds for more targeted real time visualization queries on data sets. A quick perusal of ChEBI's GitHub open *and* closed issues page [11] will testify to how responsive the development team is to community requests for new and modified entries. Further, while AAT offers many contemporary and historical measurement units, and QUDT [12] is a W3C member with a second release in process that could well standardize LOD units for scientific disciplines, OBO Foundry's Units of measurement Ontology (UO) simply supplies the units we use in labs now [13], and with the same straightforward pragmatism as the other OBO Foundry ontologies serve their areas of interest.

Finally, the NCI Thesaurus OBO Edition (NCIT) [14] has been invaluable for providing a curated list of qualitative terms, which are necessary for a wide range of categorical data types, and which few of the other ontologies we looked at included. A GC-MS peak's compound identification may well be 'uncertain' (NCIT_C47944), and it is responsible to note it as such (fig. 6). An Oddy Test coupon might indicate the sample is appropriate for 'permanent' or 'temporary' use with collections, or that it is 'unsuitable' for use around cultural heritage objects (fig. 7). With subjective data of this kind, even terms like 'permanent' or 'temporary' are problematic, implying as they do that there is something permanent or temporary about the sample itself. Accordingly, for now we have opted to assign LOD terms that reflect what the test is designed to ascertain rather than use LOD equivalents for those conventional words themselves. Namely, did the Oddy Test suggest that the sample material is *acceptable*, *conditional*(ly acceptable), or *unacceptable* (NCIT_C63350, NCIT_C63905, or NCIT_C63354) for use around heritage objects?

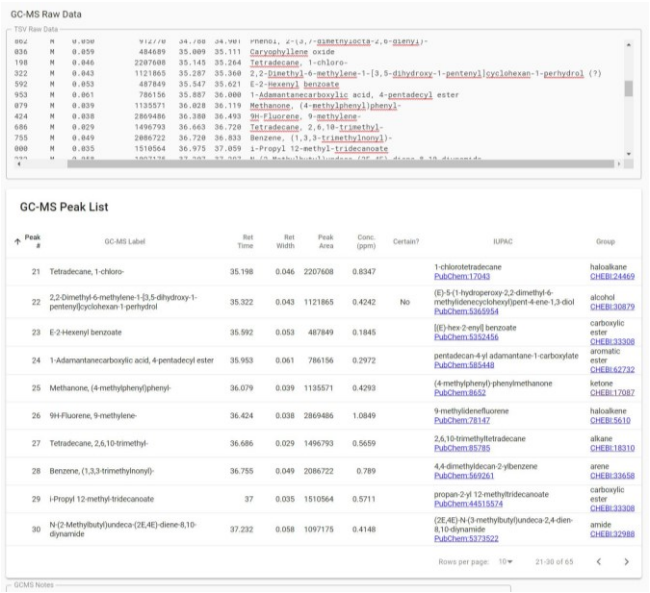

***Figure 6**. Parsing raw GC-MS data to LOD for use via Mirador annotations requires qualitative terms to represent categorical data for a GC-MS peak's confidence level (here, 'certain' vs not 'certain') as well as a variety of discrete and continuous quantitative values (time in minutes and peak area).*

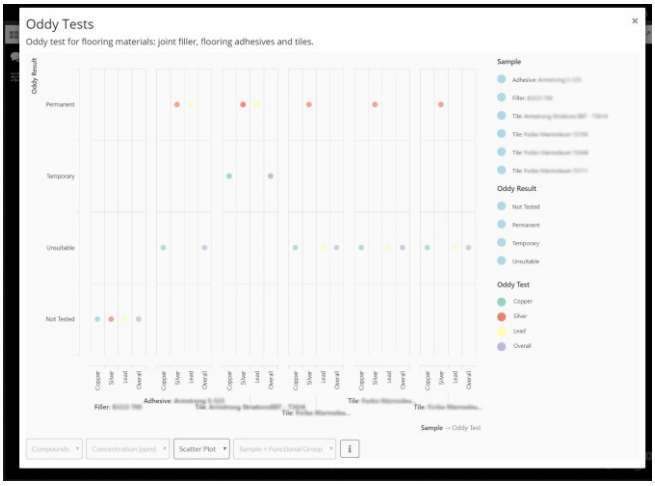

***Figure 7**. Categorical data used in the visualization of Oddy Test results via a Mirador Annotation – not all data can be represented numerically, we also need LOD qualitative terms.*

Modeling Scientific Heritage Data

IIIF APIs and Mirador form the backbone for publishing our humanities and scientific data, so we looked to Linked Art [15], a IIIF sister community for modeling cultural heritage data. Using the same development and design principles that have guided IIIF, and supported by many of the same institutions [16], Linked Art offers cultural heritage institutions a subset of the CIDOC's CRM, 'a functional and robust baseline to cover 90% of the use cases of 90% of the organizations, with only 10% of the complexity of the full CRM ontology with all of its approved extensions' [17]. As with IIIF, decisions have been made based on what is genuinely useful, provides interoperability with other data sets, and lowers the technological barriers for adoption (for example, using JSON-LD [18] as the primary data format) [19]. For these same reasons, it would be a mistake for Linked Art to expand its

scope to account for scientific heritage data. We have instead begun applying Linked Art's selection criteria and conventions to the CIDOC CRMsci [20], creating a simple, yet fully functional and compatible, scientific observation model expansion for Linked Art.

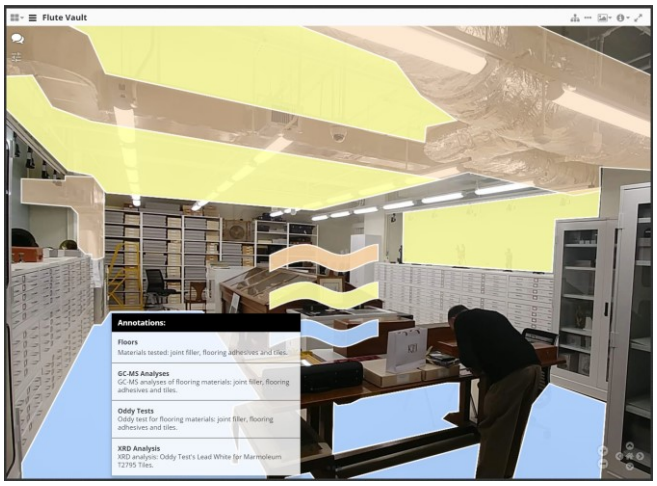

***Figure 8**. Using Mirador Annotations to group and link visualizations of scientific data for, in this case, materials tested for use on the HVAC system, floors, walls and ceiling, as well as ongoing environmental testing in this vault.*

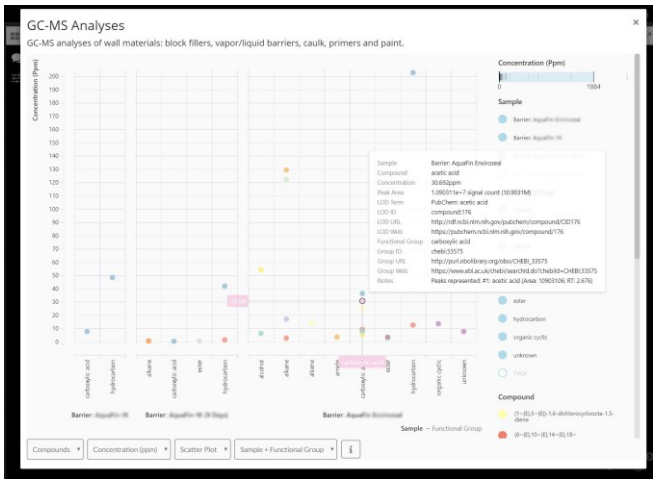

***Figure 9**. Gas Chromatography – Mass Spectrometry Data Rendered to allow multiple uses.*

The resulting structured data can properly account for events such as describing materials, taking samples, making observations, analyzing data, making predictions, and using simulations. And, significantly, we can not only fully describe analytical parameters without resorting to exhaustively complex data models, but we can employ the same straightforward Linked Art classes and properties as a curator might use to describe the dimensions of a manuscript, and these same classes and properties resolve to the standards developed by the CIDOC CRM.

Results

The chemistry terminology was extensive, so we began a structured overview, to see where there were gaps in existing terminology for cultural heritage, so that we could work collaboratively with others to fill these gaps and create a shared, and not simply internal, institutional terminology. This exploration was instructive to see what heritage terms had been created, and where specific areas of research lacked identifying information, therefore pushing institutions towards creating their own terminology. We wanted to reuse existing resources, believing that the problems we faced must have been faced and resolved before, even if that was by those in adjacent or wholly different fields. And to date that has been the case, but we also feel the OBO Foundry's success is an example of how to go about meeting discipline-specific needs the right way. The tight scope of each ontology makes for a tidy and manageable resource, while the rules and collaborative spirit that holds the collective together allows users to mix-and-match between ontologies where it makes sense to without fear of creating inconsistencies in terminology, and terms that must necessarily appear in multiple ontologies are clearly marked and cross-referenced. Perhaps cultural heritage fields could be well served by a similar collective approach, where a number of smaller, targeted, ontologies *worked together* in a coordinated way. The approach seems to be more flexible and resilient over time than the traditional monolithic resource.

It was obvious from our early work visualizing humanities and scientific data with IIIF that for truly useful data interaction, and especially so for data reuse, we needed to move from static data files to dynamically generated LOD accessed via our own APIs, with the commensurate challenges and benefits this transition brings. For a properly functional architecture, the interface needs to be simple for the user to engage with critical information – whether as a researcher uploading analytical results, or a curator searching for, say, information on a specific pigment. However, it became apparent that there were two significant functions that complement one another in institutions that work with research data, and yet have conflicting data model requirements. "Administrative" functions require a specific outcome, a file upload that becomes the record for that project/study for instance, and these can be naturally stored in and retrieved from named fields in normalized RDBMS tables. However, "research" functions require access to data that may or may not even be summarized in such records, and would be non-trivial to extract from, for instance, a dozen PDF files, even if there were a way to identify those specific files from the (tens or even hundreds of) thousands of administrative project records. A report will have identified several peaks deemed significant in various analyses, and the file might have embedded spectra images. But, at a later date, when a researcher wants to see if a new pattern can be discerned in the results of earlier analyses, it is impossible to extract information about unidentified peaks, or garner any useful data from the spectra's low-resolution images. To make the data accessible, (re)usable, and able to be restructured to relate to multiple current (and potential future) queries, then the data requires a completely different structure. Apache CouchDB [21], a JSON document-oriented key-value store, allows us to store, index, filter, and create views of heterogeneous analytical data in structures that make sense for each kind of analysis (for example, GC-MS, Oddy Tests, XRD, FTIR, pH, SEC, FTIR), and enables views of disparate analyses (look up analyses that revealed compounds belonging to a given functional group, for instance). Raw data can also be stored, then filtered as required for future lines of inquiry. Unlike the traditional RDBMS approach, this is achieved *without* having to create tables for each type of analysis or add many new columns to a single table, almost all of which would be largely

redundant for any given record. Both RDBMS-oriented solutions also limit discovery options across analyses, where creating customized filters and views is a non-trivial task. CouchDB appeals especially for its replication protocol – a web browser on a mobile phone, in a lab environment where there is no WiFi or wired network access, can continue working using its own copy of the data, which reconciles with servers once the researcher is back in a WiFi zone. Triple stores, and SPARQL interfaces for them, such as Jena Fuseki [22], provide yet another way to handle research data, in this case natively in LOD triples (subject, predicate, object). That said, while *publishing* reusable LOD is one of our project goals, *storing and managing* analytical data in a triple store seems unnecessarily complex and cumbersome compared with the other available options, such as a lightweight, efficient, JSON-based key-value store like CouchDB.

For administrative purposes, we cannot dispense with a RDBMS, so we are exploring ways to complement project data held there with analytical and API-related data held in key-value and triple store databases, *without* having to manage the same data in multiple places.

Conclusions

As has been apparent from the above discussions, creating an infrastructure that enables effective linking of data with interactive annotations requires considerable coordination and careful assessment of authoritative sources for data descriptors. The project has worked to aggregate thesauri and vocabularies, crosswalking them to assure shared terminology. Implications of this approach are to encourage and engage colleagues to allow shared and updated thesauri and terminologies to be utilized, and truly follow the linked open data approach. Until we embrace this approach we will continue to confound and frustrate potential users of heritage science data, and maintain the farcical separation between heritage science and scholarly humanities data and research. Further, maintaining only a few updated comprehensive heritage vocabularies will be of significant benefit to heritage institutions showing true collaboration and engagement across libraries, archives, galleries and museums. For like reasons, we promote using IIIF and Linked Art principles in the modelling of heritage scientific data, where the aim is to deliver lightweight workable solutions for the community early, then iterate through versions that resolve community use cases as they arise, over 'perfect,' convoluted, and exhaustively comprehensive solutions that will be usable someday, perhaps.

There is a strong need in the heritage community (and the sciences) for a more coordinated approach to the sharing, storing for longevity, and effective reuse of data. The desire to create a unified approach is challenging, but until we start working together to achieve this goal, multiple platforms with limited interoperability will continue to be created and recreated. In a world with a plethora of data, there are often new questions to be asked. Having the capacity to interrogate datasets for more effective mining of data will greatly advance our capacity to answer preservation questions and link together related data sets.

References

[1] France, F.G., Wilson, M and Bolser, C, "Crosswalking or jaywalking: the visualization of linked scientific and humanities data", Imaging Science and Technology Archiving Conference, Washington DC, April 2018.

[2] CIDOC Conceptual Reference Model (CRM) for cultural heritage documentation: http://cidoc-crm.org/

[3] http://vocab.getty.edu/

[4] https://iupac.org/what-we-do/books/color-books/

[5] https://iupac.org/projects/project-details/?project_nr=2016-046-1-024

[6] https://tools.wmflabs.org/mix-n-match/#/catalog/908

[7] PubChem (National Center for Biotechnology Information): https://pubchem.ncbi.nlm.nih.gov/search/, ChEMBL (European Bioinformatics Institute): https://www.ebi.ac.uk/chembl/beta/, Unique Ingredient Identifier (UNII, FDA): https://fdasis.nlm.nih.gov/srs/

[8] For example, see: 'absorbed dose': https://www.wikidata.org/wiki/Q215313, which incorporates the GoldBook's absorbed dose of radiation and of a substance (A00031 and A00030, respectively).

[9] http://www.obofoundry.org/

[10] For example, 'thermal desorption – gas chromatography mass spectrometry': CHMO_00026640, http://purl.obolibrary.org/obo/CHMO_0002640.

[11] https://github.com/ebi-chebi/ChEBI/issues?q=is%3Aissue

[12] Quantity kinds, Units of measure, Dimensions, and (Data) Types (QUDT, W3C): http://www.qudt.org/

[13] For example, 'micrograms per milliliter' (UO_0000274) and 'parts per million' (UO_0000169) for a 'concentration unit' (UO_0000051).

[14] http://www.obofoundry.org/ontology/ncit.html

[15] https://linked.art/model/

[16] https://linked.art/community/index.html

[17] https://linked.art/model/profile/

[18] https://json-ld.org/

[19] See, for instance, their proposal and Rob Sanderson's Keynote for EuropeanaTech (May 2018) on LOUD: Linked Open Usable Data, https://linked.art/loud/index.html. In terms of intent, there are clear parallels with FAIR Data Principles: https://www.force11.org/group/fairgroup/fairprinciples.

[20] CIDOC CRM Scientific Observation Model (CRMsci): http://cidoc-crm.org/crmsci/

[21] http://couchdb.apache.org/#about

[22] https://jena.apache.org/documentation/fuseki2/index.html

Author Biography

Dr. France, Chief of the Preservation Research and Testing Division at the Library of Congress, researches non-invasive techniques and 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. She maintains collaborations with colleagues from academic, cultural, forensic and federal institutions through her service on a number of international bodies. In February 2016 Dr. France was appointed as a CLIR Distinguished Presidential Fellow.

Dr Forsberg, a Preservation Researcher in the Preservation Research and Testing Division at the Library of Congress, previously a CLIR/DLF/Mellon Postdoctoral Fellow in Data Curation for Medieval Studies, researches using internet-based technologies to improve data sharing and collaboration between the sciences and humanities in cultural heritage institutions. He has been a professional in the web development industry since the mid-1990s, and an academic researcher and lecturer in Medieval and Early Modern literature and theory.