KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Lieu, Ryan, and Alberto Campagnolo. 2022. Modelling Linked Data for Conservation: A Call for New Standards. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6(3). https://doi.org/10.18357/kula.232

PROJECT REPORT

Modelling Linked Data for Conservation: A Call for New Standards

Ryan Lieu
Stanford Libraries

Alberto Campagnolo
University of Udine

Conservation documentation serves an invaluable role in the history of cultural property, and conservators are bound by professional ethics to maintain accurate, clear, and permanent documentation about their work. Though many well-documented schemata exist for describing the holdings of memory organizations, none are designed to capture conservation documentation data in a semantically meaningful way. Conservation data often includes deeply detailed observations about the physical structure, materiality, and condition state of an object and how these characteristics change over time. When included with descriptive catalog metadata, these conservation data points typically manifest in seldom-used fields as free-text notes written with inconsistently applied standards and uncontrolled vocabularies. Beyond the traditional scope of descriptive metadata, conservation treatment documentation includes event-oriented data that captures a sequence of steps taken by the conservator, the addition and removal of material, and cause-and-effect relationships between observed conditions and treatment decisions made by a conservator. In 2020, the Linked Conservation Data Consortium conducted a pilot project to transform unstructured conservation data into linked data. Participants examined potential models in the library field and ultimately chose to conform to the Comité International pour la Documentation (CIDOC) Conceptual Reference Model (CRM) for its accommodation of event-oriented data and detailed descriptive attribution. Project technologists worked with real report data from four institutions to create XML data models and map newly structured data to the CRM. The pilot group then imported CRM-modelled datasets into a discovery environment, developed queries to reconcile the divergent datasets, and created knowledge maps and charts in response to a small set of predetermined research questions. Feedback from conservators attending workshop activities revealed a shared need for conservation data standards and guidelines for those developing documentation templates and databases. Project outcomes signalled the necessity of further developing conservation vocabularies and ontologies to link datasets between institutions and from adjacent domains.

Keywords: conservation; data modelling; data standards; linked data; metadata; preservation; tool creation

Introduction

Conservators explore the inner workings and histories of artefacts as part of their work. They seek to understand the preservation condition of the object, recognize possible causes of past or ongoing deterioration and damage, and, to whatever extent possible, stabilize it to preserve its material evidence and its availability to future generations. Natural deterioration is typically inevitable, and not all material evidence can be preserved, even with the utmost care and through a practice of minimum intervention (simply boxing an artefact without conservation treatment). Photographic and written conservation documentation captures a significant amount of information about the history and makeup of cultural property, offering a snapshot of the life of the object at a precise point in time before information loss due to natural deterioration or modification by treatment (Campagnolo 2020).

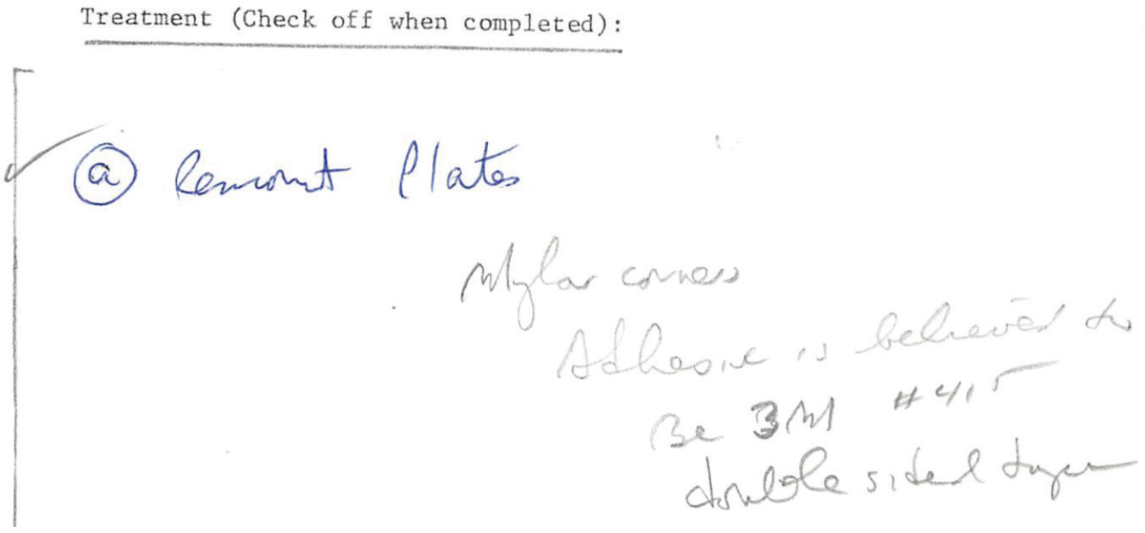
Treatment documentation is a standard practice for modern conservators and is considered a basic requirement and fundamental activity of any conservation program (Aleppo 2003; Scheper 2018, 14–15; Velios and Pickwoad 2020, 159–60). Conservation records document both historical evidence and the preservation condition of objects to assess needs for treatment, digitization, exhibitions, and other activities. Documentation tasks generally involve the careful description of materials, production techniques, and structures to record the physical characteristics of the artefacts examined; statistical analyses of such information; and creation of illustrations, diagrams, and photographs.

Though most conservation documentation has historically resided in paper archives as handwritten notes on cards (Figure 1), most conservation departments today have switched to digital documents and ad hoc databases. The Linked Conservation Data (LCD) Consortium, co-led by Athanasios Velios (Ligatus, University of the Arts London) and Kristen St. John (Stanford Libraries), is a network of partners seeking to improve access, use/reuse, and dissemination of conservation documentation records through linked data. LCD's current work focuses specifically on improving terminology in specific thesauri (e.g., the Getty's Arts and Architecture Thesaurus and Ligatus's Language of Bindings Thesaurus), modelling conservation data through the Comité International pour la Documentation (CIDOC)-Conceptual Reference Model (CRM) mapping (see below), and dissemination through linked data–ready tools (LCD Network 2020). In 2020, LCD conducted a pilot project focused on book conservation data to investigate and showcase the potential of linked data for conservators. Consortium participants from the Bodleian Library, Library of Congress (LC), the National Archives of the United Kingdom (TNA), and Stanford Libraries each submitted data for the pilot project. Alberto Campagnolo joined the LCD team as a research fellow to lead development of data models and transformation scripts, and Ryan Lieu had already supported a previous LCD terminology development project as the data steward for Stanford's conservation lab.

Book Conservation Records and Existing Metadata Models

As a supplement to core library and museum catalog data, conservation data is highly useful for capturing preservation condition, treatment modification decisions, and other changes to an item in a collection over time. However, it does not fit well within established descriptive metadata models used in the library world, such as Bibliographic Framework Initiative (BIBFRAME), a model and vocabulary for recording bibliographic description (Library of Congress 2021), or Preservation Metadata: Implementation Strategies (PREMIS), a standard for implementing preservation metadata in digital preservation systems (PREMIS Editorial Committee 2015).

As noted previously, legacy data resides in a variety of formats ranging from text-based documents, including handwritten reports with checkboxes and somewhat searchable but similarly structured electronic word-processing files, to structured collections of information, including spreadsheets and databases (Ravenberg 2012). However, even within highly organized spreadsheets and databases, users typically enter conservation data as free text supplemented with photographic documentation. While traditional free-text



Treatment (Check off when completed):

✓ (a) Remount Plates

Mylar corners

Adhesive is believed to
Be 3M #415
double sided tape

Figure 1: An excerpt from a handwritten conservation report. Courtesy of Stanford Libraries.

documentation may be suitable for single-object record systems, where an information seeker opens a document to read documentation with a specific item of interest in mind, free text limits complex information retrieval, multi-language access, and the integration of different databases. More structured models and controlled vocabularies provide better long-term solutions to document large collections.

Besides descriptive records and the issues related to unstructured free-text information, photographic documentation and diagrams are also a vital source of information on the condition state and treatment of artefacts. These are generally annotated by hand, if at all, and therefore automated information retrieval remains difficult. Photographic documentation captures data that would be otherwise lost in written documentation alone. Spatial information is rarely represented properly by written documentation; natural language alone is not capable of communicating specific and detailed descriptions of spatial relationships because spatiality is challenging to articulate through linear and sequential verbal communication (Campagnolo 2015, 69–81, 105–7). However, structured data can capture and record spatial relationships up to a degree (Campagnolo 2015, 109–11; Porter, Campagnolo, and Connelly 2017). Similarly, photographs are better at capturing colour and discolouration. Though colourimeters are a superior tool for precise measurements of these properties, they are not generally found in conservation laboratories. Advanced imaging techniques (e.g., spectral imaging, X-ray fluorescence imaging) can be used to evaluate discolouration, track changes over time, assess the efficacy of conservation treatments, characterize and identify inks and pigments, and archive data on objects' materiality (France 2016, 2020, 172–76). For this reason, in 2017, the Committee for Conservation of the International Council of Museums (ICOM-CC) established a triennial working group on documentation (2017–20), concerned specifically with the use of digital technologies in the documentation of objects, to promote imaging techniques as a tool for examination (ICOM-CC 2017).

The peculiarity of conservation data stems from conservators' need to record detailed pictorial and textual information in two parallel data streams—one for physically descriptive data about structures and materials and another for event data about step-by-step treatment activities. In book conservation, the volume is examined and described structure by structure (e.g., sewing, endpapers, endbands, cover) to document preservation condition following the hierarchy suggested by the physical makeup of the book. Responding to these observations, the conservator records event-oriented data about sequential modifications to the collection object deemed necessary to stabilize or improve its condition. For these reasons, while metadata models developed for bibliographic description may suit the first reporting stream to some degree, they are not well suited to record conservation treatment event data. Instead of adapting existing book-specific data models, the LCD team sought models better suited to events and activities. Structured textual data can then be mapped onto photographic documentation through International Image Interoperability Framwork (IIIF) annotations (IIIF 2021).

The Project

Data Modelling

The authors chose the CIDOC-CRM for the model's event-centric orientation, which would accommodate modelling of conservation treatment activities and allow for conservators' description of objects at many levels of detail as both free text and through controlled vocabularies. The CRM was already familiar to some project participants, who also knew of existing technology optimized for mapping data to the ontology.

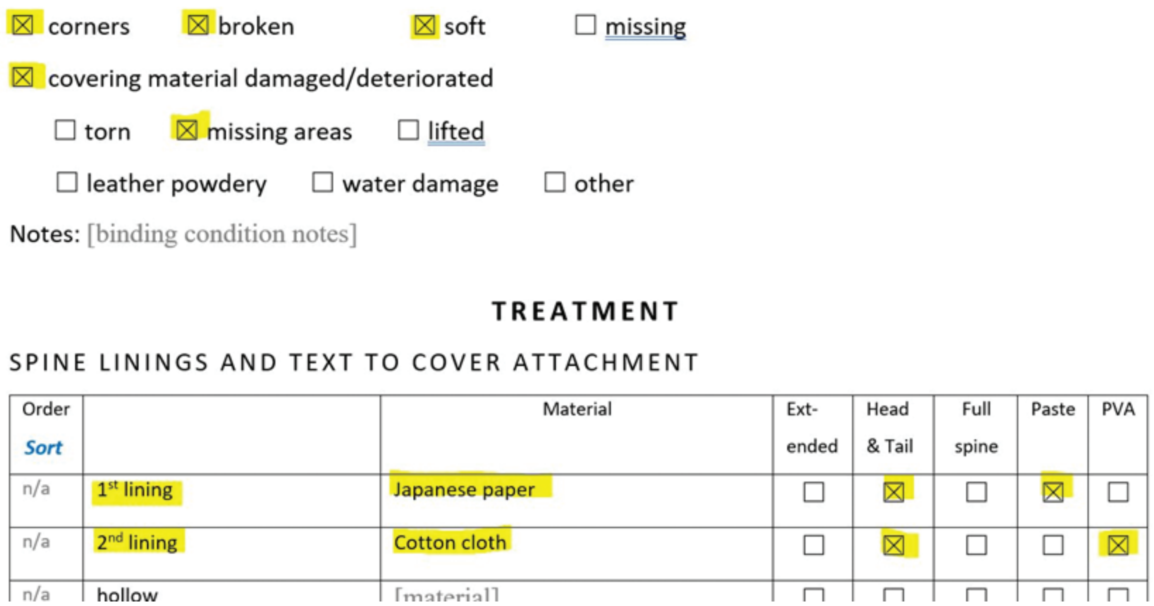
The data model for the project evolved from a relatively flat hierarchy into a layered model suited to expression as linked data. The initially flat hierarchy emerged from Stanford's existing XML data, generated by born-digital checkbox forms, which captured true or false statements regarding the description, condition, and treatment of collection objects (Figure 2). The highest levels of the initial model reflected this document structure, with a complex element to hold identifying information about an object and respective complex elements for description, condition, and treatment report data.

The Conservation Division of LC contributed spreadsheet data that followed a one-node-to-one-concept schema, which would map more easily to the CRM than Stanford's data. Based on professional guidelines for conservation documentation (AIC 1994; CAC and CAPC 2000; ICON 2020), the authors developed a new data model to act as an intermediary XML model for aligning spreadsheets and Stanford's flat data with the CRM. This intermediary model represents objects as aggregates of *components* composed of *materials* and exhibiting *condition states*. The objects are modified by conservation treatment *activities* employing *techniques* and *materials* (Figure 3). The model allows for recursive *component* and *activity* node hierarchies to accommodate varying levels of reporting detail employed throughout the sampled conservation records.

The data from Stanford required additional transformation steps to parse information into the discrete CRM-aligned nodes of the intermediary XML model. In Stanford's conservation report form, one checkbox often represented observations made or treatment steps taken that involved several conservation terms at once. As a result, each Boolean data node in the flat XML model represented several vocabulary concepts in the true-or-false checkbox forms (Figure 4). For example, a single checkbox for "leather treated with microcrystalline wax" meant that the "covering material" component composed of the material "leather" was modified by a treatment activity employing the material "microcrystalline wax."

Data Transformation

Transformation of data from initial XML input into Resource Description Framework (RDF)/XML occurred via two pipelines (Figure 5), each devised and managed by one of the authors. To transform datasets from the Bodleian, TNA, and LC through Pipeline A, Campagnolo rekeyed scanned paper records contributed by the Bodleian into intermediary XML files based on an ad hoc schema. TNA submitted data to the project in a series of spreadsheets similar to the LC spreadsheets. Campagnolo then imported the spreadsheet data into the Oxygen XML Editor application to convert it into XML data based on the same ad hoc schema developed



☒ corners ☒ broken ☒ soft □ missing

☒ covering material damaged/deteriorated

□ torn ☒ missing areas □ lifted

□ leather powdery □ water damage □ other

Notes: [binding condition notes]

TREATMENT

SPINE LININGS AND TEXT TO COVER ATTACHMENT

Order *Sort*		Material	Ext- ended	Head & Tail	Full spine	Paste	PVA
n/a	1st lining	Japanese paper	□	☒	□	☒	□
n/a	2nd lining	Cotton cloth	□	☒	□	□	☒
n/a	hollow	[material]	□	□	□	□	□

Figure 2: An excerpt from a Stanford conservation report Word document file. Courtesy of Stanford Libraries.

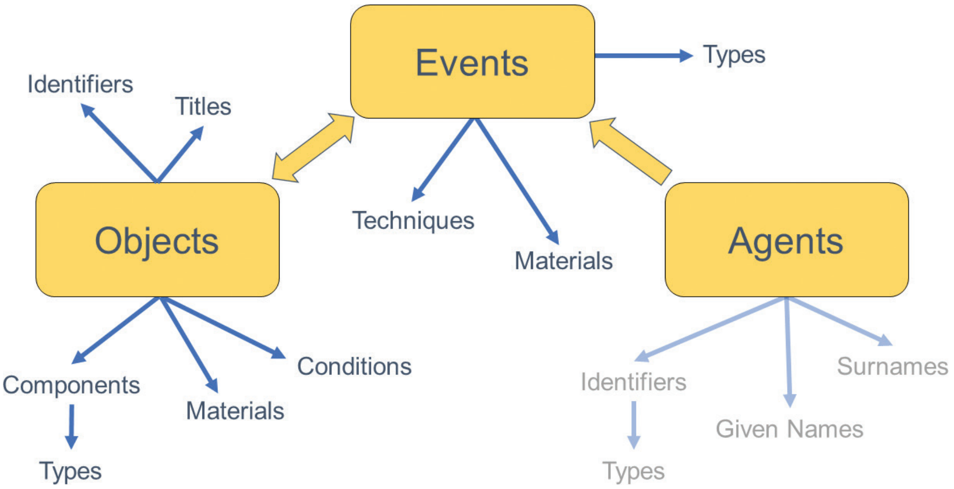
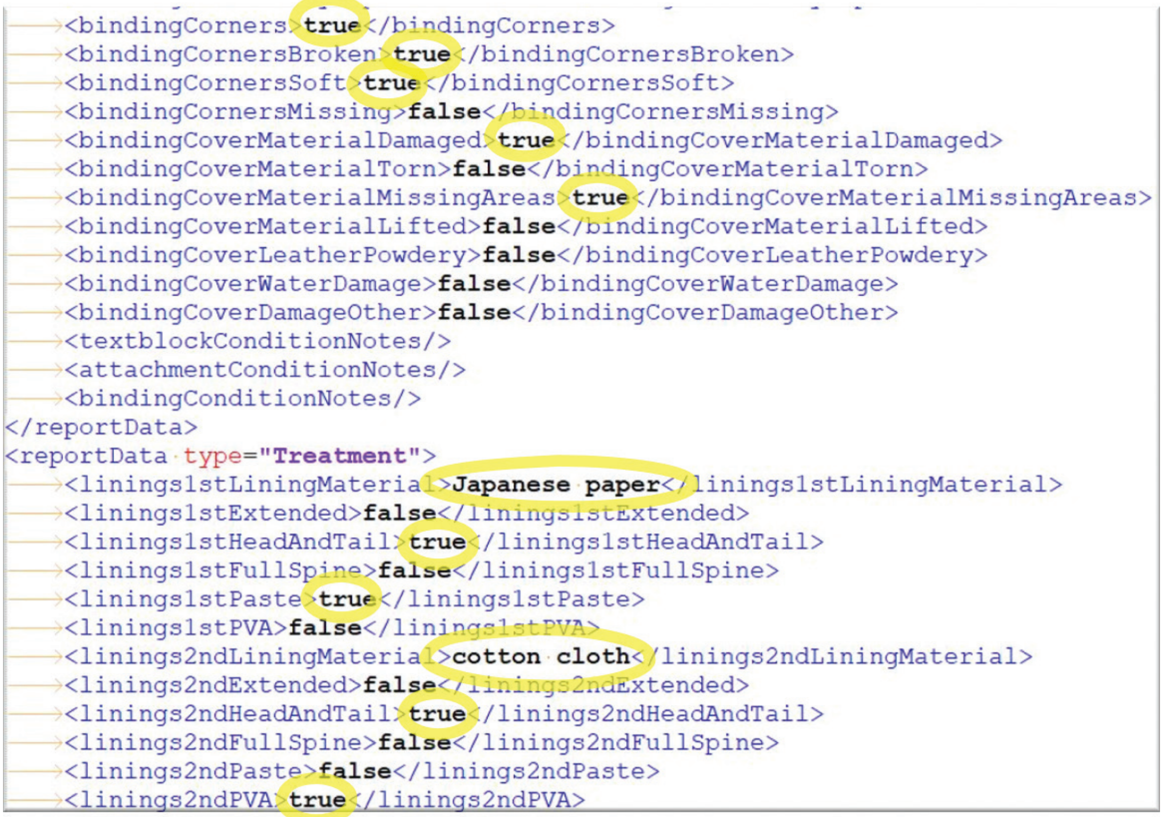


Figure 3: A diagram of the project's data model developed as an intermediary to the CIDOC-CRM. Agent's attributes were not fully modelled for the project per privacy regulations. Courtesy of the authors.



```
    <bindingCorners>true</bindingCorners>
    <bindingCornersBroken>true</bindingCornersBroken>
    <bindingCornersSoft>true</bindingCornersSoft>
    <bindingCornersMissing>false</bindingCornersMissing>
    <bindingCoverMaterialDamaged>true</bindingCoverMaterialDamaged>
    <bindingCoverMaterialTorn>false</bindingCoverMaterialTorn>
    <bindingCoverMaterialMissingAreas>true</bindingCoverMaterialMissingAreas>
    <bindingCoverMaterialLifted>false</bindingCoverMaterialLifted>
    <bindingCoverLeatherPowdery>false</bindingCoverLeatherPowdery>
    <bindingCoverWaterDamage>false</bindingCoverWaterDamage>
    <bindingCoverDamageOther>false</bindingCoverDamageOther>
    <textblockConditionNotes/>
    <attachmentConditionNotes/>
    <bindingConditionNotes/>
</reportData>
<reportData type="Treatment">
    <linings1stLiningMaterial>Japanese paper</linings1stLiningMaterial>
    <linings1stExtended>false</linings1stExtended>
    <linings1stHeadAndTail>true</linings1stHeadAndTail>
    <linings1stFullSpine>false</linings1stFullSpine>
    <linings1stPaste>true</linings1stPaste>
    <linings1stPVA>false</linings1stPVA>
    <linings2ndLiningMaterial>cotton cloth</linings2ndLiningMaterial>
    <linings2ndExtended>false</linings2ndExtended>
    <linings2ndHeadAndTail>true</linings2ndHeadAndTail>
    <linings2ndFullSpine>false</linings2ndFullSpine>
    <linings2ndPaste>false</linings2ndPaste>
    <linings2ndPVA>true</linings2ndPVA>
```

Figure 4: XML data mined from Stanford's born-digital conservation reports conformed to a flat structure optimized for integration with Word document form content controls. Courtesy of the authors.

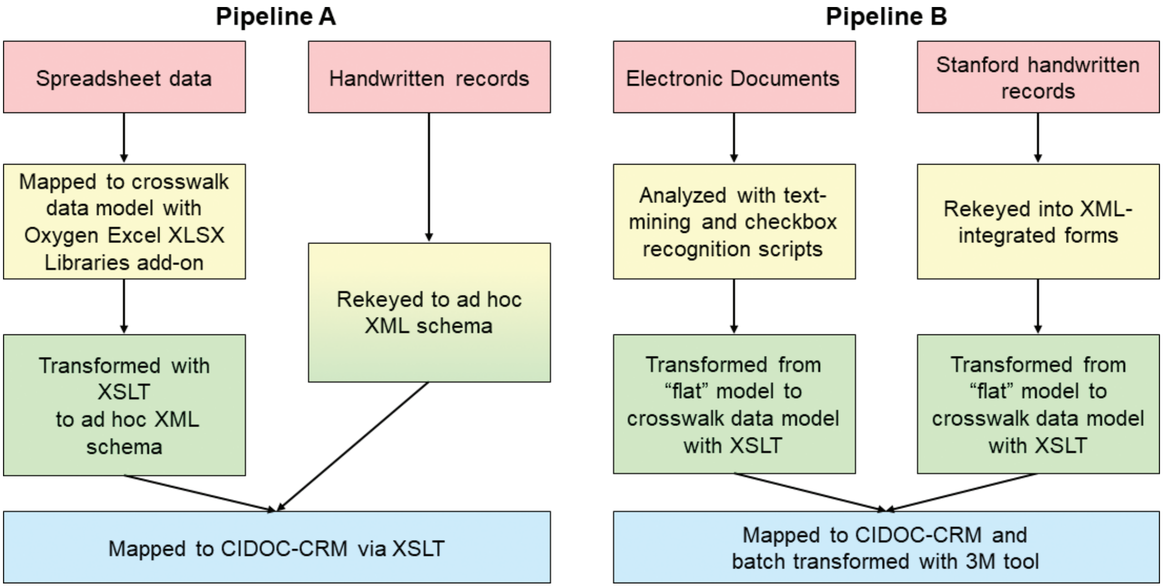

Pipeline A
Spreadsheet data
Handwritten records
Mapped to crosswalk data model with Oxygen Excel XLSX Libraries add-on
Transformed with XSLT to ad hoc XML schema
Rekeyed to ad hoc XML schema
Mapped to CIDOC-CRM via XSLT
Pipeline B
Electronic Documents
Stanford handwritten records
Analyzed with text-mining and checkbox recognition scripts
Rekeyed into XML-integrated forms
Transformed from "flat" model to crosswalk data model with XSLT
Transformed from "flat" model to crosswalk data model with XSLT
Mapped to CIDOC-CRM and batch transformed with 3M tool

Figure 5: The two transformation workflows employed in the project. Courtesy of the authors.

for Bodleian data, and subsequently applied XSLT scripts to transform the imported data into CRM-aligned RDF/XML.

To transform Stanford's data through Pipeline B, project co-leader St. John rekeyed scanned paper records into an XML-bound report template to match the flat XML records extracted from born-digital report files, and Lieu wrote and applied an XSLT script to transform the flat Stanford XML into the intermediary XML model. Lieu then transformed intermediary-model XML records further with a mapping created in the

Memory Mapping Manager (known more commonly to users as 3M), a free online tool developed by the Foundation for Research and Technology (FORTH 2019).

3M is optimized to work with the CRM ontology to create data mappings from XML data to linked data. After uploading an XML Schema file or a sample XML record, the 3M user enters XPath expressions, chooses entities and properties from the pre-loaded ontology, and drafts URI and label-generation rules to map relationships between XML nodes. The resulting mapping files provide processing instructions for the 3M engine to draw data from the nodes identified in local XML data to serialize an RDF dataset in a variety of encodings.

Data Review and Revision

Due to the complexity of the network of links in RDF data, RDF serializations are not typically human-readable. Human reading of RDF data generally requires interpretation and rendering by a machine to illustrate relationships between points in a dataset. The authors reviewed the first generation of RDF output with a Python command line tool called CRMVIZ, developed by project co-leader Velios, that generates graphs from CRM-mapped datasets, providing the user with image files to review linked data as diagrams (Velios 2020). Initial review revealed a few minor label-generation issues where long entity titles required truncation or incorrect language attributes had been used.

The authors also discovered a few major mapping issues in need of revision at this early stage. The diagrams revealed unexpected results from the 3M tool where broad XPath expressions linked more nodes than were accurate in reality. Some data nodes were missing from the initial RDF files due to invalid use of the CRM. For example, one mapping called classes from an unreleased version of the ontology, and the other mapped conservators' activities as employing *specific* rather than *general techniques*. The authors also determined it necessary to add *attribute assignment* and *condition assessment* events to the data model to explicitly distinguish observations as data collected before the treatment *modification* events occurred.

Further mapping and data revisions occurred iteratively throughout the remainder of the project as project co-lead Velios tested the data via SPARQL queries in the ResearchSpace discovery environment. ResearchSpace is a digital scholarship platform that leverages linked data to allow researchers to connect, communicate, and represent knowledge in context through a graphical user interface (Oldman and Tanase 2018, 325–40). While the system's capacity for analysis and visualization allowed the team to answer the project's research questions, setting up and using the ResearchSpace instance to create charts and knowledge maps from the project data emphasized a divergence in conservation documentation practices at different institutions, particularly in regard to preferred terminology and levels of detail captured.

Workshop Presentations

The authors presented their work to an international audience of conservators during a modelling workshop held at the close of the project. Throughout the presentation, several attendees asked if the LCD Consortium would publish the model. Others called for professional data standards. One participant specifically requested any kind of database template upon which they could base their own documentation system. Rather than seeking more specific knowledge about the technology or ontology described in the workshop, conservators in the audience sought guidance of a much more general nature—a shared model based on our common practices so that each conservation lab need not reinvent their own wheel to update documentation systems.

Conclusion

Within the field of preservation, shared data models and linked data would enable conservators and other practitioners to compare notes on specific treatments, certain materials, and historical structures. Beyond the conservation studio, sharing conservation data through a standardized model as offered by LCD's mapping to the CIDOC-CRM would allow access to information generally not included in library or museum catalogues, exposing previously siloed data for scholarship and integration into colleagues' processes. Although usually ignored due to its inaccessibility, conservation data is an essential additional information resource that can complement, integrate, and corroborate evidence collected by other practitioners, including scholars, curators, and scientists. In libraries and museums, interoperable conservation data could enhance metadata developed by catalogers and archivists, supplementing and updating physical descriptions after close examination and modification of a collection object by a conservator.

Linking conservation data generated through divergent documentation practices at different labs revealed the consequences of vague data standards applied across the field. The profession's broad guidelines for documentation provide much flexibility for conservators to adapt their practices to the unique needs of each object or project. However, the looseness of guidance on documentation and a lack of data standards have consequently left many conservators in the dark about the best practices for and the benefits of creating structured, searchable, and analyzable data.

While the project revealed useful insights into the modelling of conservation treatment data, much work remains before truly interoperable linked conservation data becomes viable. The community must continue to develop terminology for the conservation domain and determine strategies for vocabulary alignment between similar terms to accommodate divergent terminology preferences. New data standards should take into consideration the differences in granularity of detail captured between conservation records, varying over decades from project to project, and based on conservators' habits and preferences. For conversion of legacy data to linked data, flexibility in modelling granularity will be necessary to work around resource limitations and varying data sharing preferences. Finally, to realize the true benefits of linked conservation data, research in modularization of ontologies is essential for integrating this data with linked data from catalogers, scientists, and scholars.

Given the lack of controlled vocabularies to cover many essential conservation terms, the challenges encountered by the LCD Consortium in reaching consensus regarding terminology within even small groups of conservators, and the wide range of digital fluency across the field, it may take a generation or longer to develop conservation data structure and content standards in the conventional sense of an internationally agreed-upon expression of information. International standards may never come to being, but common methodologies can help. This work must begin with more consistent communication among practitioners about conservation data, and success will likely require an agile data model and flexible querying strategies to meet the needs of many preservation professionals.

Acknowledgements

This project was supported by a grant from the AHRC, the United Kingdom's Arts and Humanities Research Council (Grant Reference Number AH/T013125/1). The authors would like to thank conservators from the Bodleian Libraries, the British Library, the John Rylands Research Institute and Library, the Library of Congress, and the National Archives of the UK for their invaluable participation and feedback as well as the ResearchSpace team at the British Museum. Our thanks also go to Linked Conservation Data project leaders Kristen St. John and Athanasios Velios for their unwavering support.

Competing Interests

The authors declare that they have no competing interests.

References

Aleppo, Mario. 2003. "160 Years of Conservation Documentation at The National Archives, UK." *The Paper Conservator* 27 (1): 97–99. https://doi.org/10.1080/03094227.2003.9638635.

American Institute for Conservation of Historic and Artistic Works. 1994. *Code of Ethics and Guidelines for Practice*. Washington, DC: AIC. https://www.culturalheritage.org/docs/default-source/administration/governance/code-of-ethics-and-guidelines-for-practice.pdf. Archived at: https://perma.cc/9YFT-N6B4.

Campagnolo, Alberto. 2015. "Transforming Structured Descriptions to Visual Representations. An Automated Visualization of Historical Bookbinding Structures." PhD diss., University of the Arts London. http://ualresearchonline.arts.ac.uk/8749/.

Campagnolo, Alberto. 2020. "Conservation and Digitization: A Difficult Balance?" In *Book Conservation and Digitization: The Challenges of Dialogue and Collaboration*, edited by Alberto Campagnolo, 49–82. Collection Development, Cultural Heritage, and Digital Humanities. Leeds: Arc Humanities Press.

Canadian Association for Conservation of Cultural Property (CAC) and of the Canadian Association of Professional Conservators (CAPC). 2000. *Code of Ethics and Guidance for Practice*. 3rd ed. Ottawa: CAC and CAPC. https://www.cac-accr.ca/download/code-of-ethics/.

France, Fenella G. 2016. "Spectral Imaging: Capturing and Retrieving Information You Didn't Know Your Library Collections Contained." In *What Do We Lose When We Lose a Library?*, edited by Lieve Watteeuw and Mel Collier, 189–97. Leuven: KU Leuven University Library. https://www.goethe.de/resources/files/pdf94/streamgate.pdf.

France, Fenella G. 2020. "Spectral Imaging to Aid Preservation and Conservation of Cultural Heritage." In *Book Conservation and Digitization: The Challenges of Dialogue and Collaboration*, edited by Alberto Campagnolo, 169–78. Collection Development, Cultural Heritage, and Digital Humanities. Leeds: Arc Humanities Press.

Foundation for Research and Technology - Hellas Institute of Computer Science (FORTH). 2019. *X3ML Toolkit*. Heraklion, Crete, GR: FORTH ICS. https://www.ics.forth.gr/isl/x3ml-toolkit.

International Council of Museums – Committee for Conservation (ICOM-CC). 2017. “Documentation Working Group. Triennial Programme 2017-2020.” https://web.archive.org/web/20210417015245/ https://www.icom-cc.org/73/Triennial%20programme/#.YHo_defP2Uc.

International Image Interoperability Framework (IIIF). 2021. http://iiif.io.

The Institute of Conservation (ICON). 2020. “Conservation Reports.” https://www.icon.org.uk/resource/intro-to-conservation-reports.html.

Linked Conservation Data Network. 2020. “Linked Conservation Data.” http://web.archive.org/web/20201124092727/ https://www.ligatus.org.uk/lcd/.

Library of Congress. 2021. “BIBFRAME Model, Vocabulary, Guidelines, Examples, Notes, Analyses.” http://web.archive.org/web/20210424013138/ https://www.loc.gov/bibframe/docs/index.html.

Oldman, Dominic, and Diana Tanase. 2018. “Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace.” In *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference*, *Monterrey, CA, USA*, *October 8–12, 2018*, edited by Denny Vrandečić et al., 325–40. Cham: Springer.

PREMIS Editorial Committee. 2015. *PREMIS Data Dictionary for Preservation Metadata, Version 3.0*. Library of Congress. http://web.archive.org/web/20210415051612/ https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

Ravenberg, Heather. 2012. “A Data Model to Describe Book Conservation Treatment Activity.” MPhil thesis, University of the Arts London.

Scheper, Karin. 2018. *The Technique of Islamic Bookbinding: Methods, Materials and Regional Varieties*. 2nd ed. Leiden: Brill. https://brill.com/view/title/31508.

Velios, Athanasios. 2020. CRMVIZ. GitHub. https://github.com/natuk/crmviz.

Velios, Athanasios, and Nicholas Pickwoad. 2020. “The Development of the Language of Bindings Thesaurus.” In *Book Conservation and Digitization: The Challenges of Dialogue and Collaboration*, edited by Alberto Campagnolo, 157–68. Collection Development, Cultural Heritage, and Digital Humanities. Leeds: Arc Humanities Press.

How to cite this article: Lieu, Ryan, and Alberto Campagnolo. 2022. Modelling Linked Data for Conservation: A Call for New Standards. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6(3). https://doi.org/10.18357/kula.232

Submitted: 28 June 2021 **Accepted:** 22 October 2021 **Published:** 27 July 2022

KULA: Knowledge Creation, Dissemination, and Preservation Studies is a peer-reviewed open access journal published by University of Victoria Libraries.

OPEN ACCESS