https://doi.org/10.2352/issn.2168-3204.2021.1.0.7

Digitizing Conservation: Developing data models for preservation data

Ryan Lieu; Stanford Libraries; Stanford, CA

Abstract

Conservation documentation plays a crucial role in preventing misrepresentations about cultural property. Yet conservation records often remain undigitized and unsearchable. As part of efforts to improve access to conservation documentation, members of the Linked Conservation Data Consortium recently embarked on a project to transform paper and born-digital conservation records spanning forty years into linked data. Project team members reviewed existing models for preservation data and found that only the CIDOC Conceptual Reference Model would accommodate documentation of materiality, object structure, and conservation treatment events as prescribed by professional guidelines. Project outcomes revealed meaningful patterns in conservation data that may be useful in future model development as well as shortcomings in the XML technologies employed for transforming the data.

Introduction

Conservation documentation plays a crucial role in preventing misrepresentations about cultural property and the information represented therein. Conservators are ethically bound to create and maintain records of their work to document changes to condition, structure, and materiality in the objects under their care. Unambiguous and accurate conservation documentation is necessary for reviewing the history of collections, and searchable treatment data allows conservators and preservation specialists the ability to research treatment options, evaluate the effectiveness of techniques and the soundness of treatment materials over time, and calculate the average time spent on different types of treatment. Yet conservation records often remain undigitized, siloed, and unsearchable.

Conservators have no formally established data standards or models developed specifically for conservation documentation. In the preservation professions, we use proprietary systems and the practice of openly sharing conservation records is relatively new and divisive. While literature about conservation documentation and guidance from professional associations establish its purpose and broadly define the elements necessary for acceptable recordkeeping, the actual mechanics of reporting and records consultation remain vague and rooted in pre-digital workflows.

In this paper, I will summarize a recent project undertaken by the Linked Conservation Data Consortium, a network of partners working to improve access to conservation records. The project revealed meaningful patterns in conservation treatment records that can inform data models and standards to help establish better digital practices in the preservation field. I will discuss discoveries made while attempting to model complex conservation assessment and treatment data as highly structured database records and potential next steps to further this work.

Conservation Documentation

Conservation documentation consists of textual and pictorial documents that record findings from pretreatment examinations including a description of structure, materials, and condition; results of scientific testing and analysis; treatment plans; and the materials and methods used in treatment [1]. McCann's 2011 survey of conservation documentation practices in academic research libraries [2] enumerates various media used by conservators to capture their documentation, including paper forms, electronic word processing files, and databases. The survey also reports that staff preferences, collection types, and numerous workflow needs all inform the development of divergent documentation systems.

Conservators' written documentation typically includes unstructured notes and narrative summaries. The narrative summary format naturally lends itself to describing a sequence of events, and though professional guidelines loosely define the elements necessary for adequate records of object examination and treatment, reliance on narrative reporting without a rigidly prescribed structure provides the necessary flexibility to address many kinds of projects with unique needs. Conservators may need to document pretreatment examinations, pre-acquisition examinations, extensive treatment undertaken on a single item, or batch treatment performed on a large group of objects—each situation requiring varying levels of detail with emphasis placed more heavily on different aspects of documentation based on the specifications of each project.

The Linked Conservation Data Pilot Project

In 2020, members of the Linked Conservation Data (LCD) Consortium embarked on a pilot project to transform conservation treatment reports into linked data. Conservation labs at the Bodleian Libraries, Library of Congress, the National Archives of the United Kingdom, and Stanford Libraries contributed reports on book board reattachment treatments dating from 1979 through 2019. Raw data formats included PDF scans of paper forms, spreadsheet data, and XML (Extensible Markup Language) documents.

Stanford Libraries had previously undertaken a project to develop and run checkbox analysis and text mining scripts on born-digital Word Document report files to capture data as XML. The resulting XML documents conformed to a relatively shallow hierarchy to capture checkbox data quickly without the complex

scripts necessary to write intertwined branches of data for each true-or-false scenario. As a result, project technologists were able to reuse only the highest levels of Stanford's initial XML model and needed to develop the hierarchy templates necessary for descriptive, condition assessment, and treatment activity data.

Review of Existing Models

Members of the project team evaluated existing shared data models to assess their suitability for conservation data. Amongst publicly available preservation metadata and conservation database models reviewed, the project technologists chose the International Committee for Documentation's Conceptual Reference Model (CIDOC CRM) [3] as it would accommodate both documentation of materiality in detail and event-oriented treatment data.

PREMIS

The highest levels of Stanford's initial XML model were drawn from PREMIS, a preservation metadata standard developed by the Library of Congress for digital preservation [4]. Stanford's model reused three top-level PREMIS entities—*Objects*, *Agents* (i.e. conservators), and *Events* (i.e. treatment activities).

The fourth PREMIS top-level *Rights* entity is of particular importance for digitization and managing access to digital surrogates. This entity was not reused since Stanford's conservation data does not include rights metadata about the items conserved, and such metadata is available from authoritative sources elsewhere within Stanford Libraries. As a result, the *Rights* entity was not included in the LCD pilot project, but this entity must be reconsidered and incorporated in future models to address concerns throughout the conservation community regarding privacy, potentially sensitive data recorded in conservation records, and release permissions granted (or not granted) by the object owners.

The Database of the St. Catherine's Library Conservation Project

To support the St. Catherine's Library Conservation Project, Velios and Pickwoad developed a database for maintaining conservators' records about the monastery's collection of manuscripts [5]. The database documents observations of structures, materials, and condition states to plan for future conservation treatment. Since its primary use has been for examination and planning, the database does not yet record any actual undertaking of conservation treatment activities. Each manuscript in the database is modeled as the sum of its component parts (e.g. binding, cover, text block) to which conservators assign structure types, materials, and preservation condition states. The database includes libraries of pre-populated term lists of controlled vocabularies, which are filtered to suit the scope of each form field.

MARC 21 583 Field and BIBFRAME

Following McCann's survey of conservation documentation practices and assessment of the usefulness of MARC data for conservators, Hobart described new practices undertaken in 2016 at Pennsylvania State University (Penn State) using the MARC 583 field to record conservation actions in bibliographic or holdings records for special collections [6]. The goals in implementing these processes included noting an item's condition, recording materials used in conservation activities, documenting treatment decisions, and identifying items treated with specific methods or techniques. A conservation "Action Note" in the MARC 583 field would be constructed as follows:

> 583 $3 [Collection name] $a [action] $c [time/date of action] $i [method of action] $k [initials] $l [status] $z [public note] $2 [source of terminology] $5 [institution to which field applies]

This model allows for documentation of condition states as [status], conservation actions and plans as [action] and [method of action] with controlled vocabularies, and the [initials] of the agent or conservator. Entry of conservation data into the bibliographic record provides the additional benefits of collocating all of a library's information about an object and subjecting conservation data to regular backups. The model does not provide for structured documentation of detailed descriptive attributes identified by a conservator during examination, nor does it accommodate component-level documentation of condition assessments and treatment activities. This data must be recorded in public or private notes as free text. Hobart identifies limitations of "Action Notes" for documenting conservation treatments of special collections where detailed information is crucial to capturing the complexity of decisions made and steps taken by the conservator.

The model's ties to bibliographic data present other perceived drawbacks for conservators. Though it may provide a relatively useful template for the museum community, it can only be applied directly for library materials described with MARC 21. Moreover, since library conservators are not typically allowed privileges to edit MARC records, they must work in constant collaboration with catalogers in order to create and maintain this data.

As of 2019, published mappings from the MARC 583 field to its successor ontology, BIBFRAME, convert most "Action Note" data to free text notes, generalizing the data and making it less specific to the preservation professions [7]. There has been no attempt to convert the [method of action] subfield $i where conservation techniques and methods are documented in Field 583, so it is unclear how this data will migrate to BIBFRAME.

CIDOC CRM

The primary purpose of the CIDOC CRM is to facilitate integration between heterogeneous sources of cultural heritage data [8]. The model's broad scope and event-centric nature allow for object description at many degrees of granularity, causality in preservation decision-making, and event-based documentation of conservation treatment activities. The model supports both highly structured data and unstructured free text notes simultaneously.

Developed to model countless scenarios, the breadth of the CRM may overwhelm many newcomers as the exhaustively detailed list of classes and properties can quickly entrench new users in the work of determining differences between similar entities. Those working specifically with preservation and conservation data may initially struggle to identify the simple patterns and relatively small subset of classes and properties necessary to model their own records. For example, one may be confused as to whether their collection item should be classified as an *E18 Physical Thing*, an *E19 Physical Object*, an *E22 Human-made Object*, or an *E24 Physical Human-made Thing*. Similarly, one might struggle to

remember whether to use the *P5 consists of*, *P9 consists of*, or *P45 consists of* property for their data.

Data Preparation

The new XML model devised for preparing data for transformation into linked data incorporated three top-level PREMIS entities—*Object(s)*, *Agent(s)*, and *Event(s)*—with enriched hierarchies for describing objects and documenting treatment activities. Objects conserved were modeled as *components* with *structure types*, consisting of *materials*, and exhibiting *condition states*. Conservation treatments were modeled as *activities* employing *techniques* and *materials* with each activity further classified as a *part addition*, *part removal*, or *repair* as appropriate.

Project technologists transformed raw data into documents conforming to the new data model via two pipelines. One technologist converted spreadsheet data into XML documents by importing Microsoft Excel files into the Oxygen XML Editor, an application for working with XML documents and working with related technologies. The technologist then converted the intermediate-model XML data into RDF/XML (the XML format expression of Resource Description Framework linked data) with Extensible Stylesheet Language Transformations (XSLT). XSLT is an XML technology developed for transforming one XML document into another. An XSL Transformation script steps node-by-node through an existing XML document and provides instructions to a processor to write new data according to parameters written by the XSLT author. The most common use of XSLT is to transform XML records into HTML to view as web pages [9].

Stanford participants converted undigitized paper records to initially "flat" XML documents by cataloging them with a Microsoft Word form template bound to custom XML markup [10][11] that matched the XML documents already supplied from Stanford's more recent born-digital documentation and previous data mining projects. The technologist then transformed the resulting "flat" XML documents into the new intermediate CRM-aligned XML model with another XSLT script. To illustrate the nature of the conversions that occurred, the Stanford XSLT script transformed the following two nodes regarding condition of book boards—

```
<boardsDetachedFront>true</boardsDetachedFront>
<boardsLooseBack>true</boardsLooseBack>
```

—into the newly developed intermediate XML template—

```
<component name="boards">
    <condition>
        <remark>
            <label>detached</label>
            <places>
                <place>front</place>
            </places>
            </remark>
        <remark>
            <label>loose</label>
            <places>
                <place>back</place>
            </places>
        </remark>
    </condition>
</component>
```

—so that each node in the new hierarchy would map to a single concept in linked data.

These intermediate XML documents from Stanford's data were then processed further using the Mapping Memory Manager (3M) [12], a free tool developed by FORTH Institute of Computer Science's Information Systems Laboratory. The 3M tool provides a user-friendly graphical interface that enables users to create complex mappings from their XML input without needing to know how to write XSLT scripts. The user must understand the basics of XPath, the language used to express the hierarchical paths down the branches in an XML tree [11]. The 3M user inputs XPath expressions into tables to identify nodes in their XML input documents, identifies the desired CRM classes and properties to define each node and relationships between nodes, and provides URIs from preferred terminology to enhance linked data output with meaning from the user's specific knowledge domain. Compared to the XML tree structure of the input data, the resulting linked data has a web-like structure due to links created between nodes throughout the data (1).

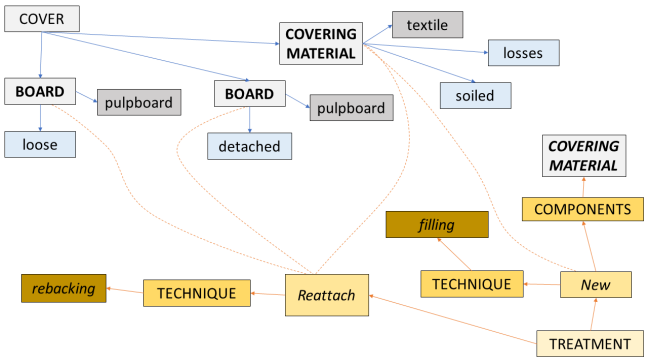

***Figure 1**. The web-like nature of linked data removes ambiguity in relationships between conservation examination and treatment data.*

Results

Mapping to the CRM tested the initial XML model based on Stanford's conservation treatment report template. Both Stanford's initial model and the intermediate model devised for the project separated repository metadata (e.g. bibliographic information, object identifiers) from descriptive attributes and data about object condition, though all three categories of data are information about the object addressed in the conservation report and thus should all be considered attributes of the *Object* entity defined at the highest level of the models. The project technologists discussed whether it was necessary or not to distinguish between the conservator's observations and attributions made at the time of the report and attributions made by others. One assertion posited that there may be dissenting opinions about any collection object. For example, a cataloger may have documented the covering material of a volume as pigskin based on a dealer's records, while the conservator believes it to be another kind of animal skin. Likewise, three conservators examining the same binding of a single volume may interpret what they see as three completely different sewing structures. Though arbitrarily separating cataloging data from conservators' observations in the document structure only served to make mapping and scripting considerably more difficult, project participants agreed that it is important to denote where possible the source and date of each attribution recorded in the linked data output

to help reconcile conflicting data created as a result of such differences in interpretation.

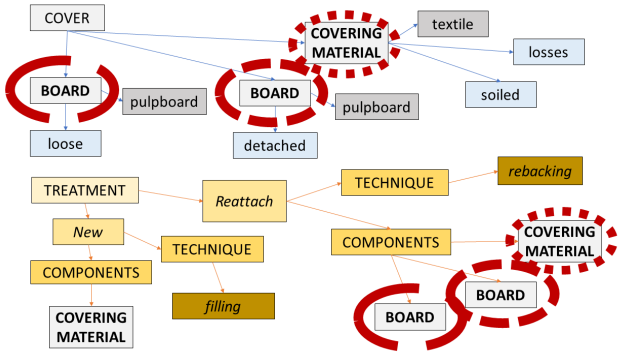

***Figure 2**. Repeating references to the same or similarly named components on different branches of data creates ambiguity regarding events as recorded.*

The hierarchical structure of the XML tree is not conducive to neatly expressing causality between data in separate branches. In the LCD pilot project's data model, this issue manifested as an inability to draw a direct relationship between a condition state observed and the activities undertaken to treat the condition. Since observations made by the conservator during pretreatment examination reside within the *Objects* entity and treatment activities responding to those observations reside within the separate *Events* entity, relationships between these branches can only be inferred by human reading, and references to the same or similarly named components in different sections of the data may lead to confusion (2). One strategy for expressing a relationship between a condition state and a treatment activity more directly is to model data with the activities descending directly from a condition state node. However, treatment activities often affect multiple parts of an object and may remedy multiple conditions at once. Modeling data in this manner would create a confusing repetition of data (3).

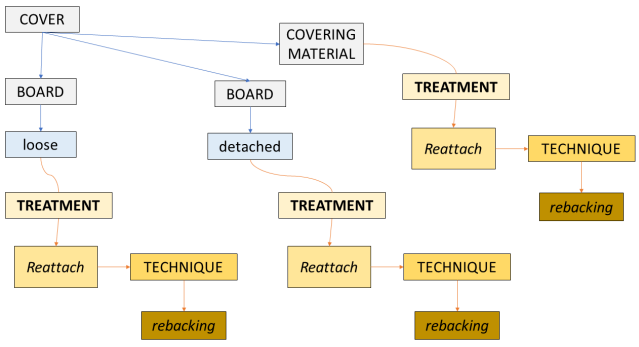

***Figure 3.** Repeating branches representing the same event affecting different components implies similar events may have occurred multiple times.*

Another problem encountered during mapping to the CIDOC CRM involved the limitations of the XPath language. Project technologists could not find a way to map from the intermediate XML model to express a precise sequence of events using the 3M tool. Users of 3M input XPath expressions to identify nodes within their XML input and additional XPath expressions as parameters for the 3M processor to define relationships between the identified node and neighboring nodes. However, the nature of XPath processing does not allow for comparison between two unpredictable values because there is no means for retaining the value held in one node while traversing the XML tree to another node. This can be accomplished with the more complex XML technologies XSLT and XQuery using variables, but the 3M application, which was essential for the complex operations necessary to conform to the CRM ontology, only accepts user-supplied XPath expressions.

Conclusion

Bringing together divergent datasets from four labs revealed how a lack of cohesive data standards for documentation over decades hindered analysis as it was difficult to reconcile differences in levels of reporting detail and disagreements over terminology. To optimize data for searchability, relational and linked graph data models suit the complexity of conservation data, but the community must determine levels of granularity for documentation vis-à-vis competing collection concerns, workflow needs, and style preferences held by conservators. In the Linked Conservation Data pilot project, this was presented to workshop participants as "signposting" for the level of detail to model from conservation records given resource limitations, whereas catalogers employ similar systems of cataloging levels, and software developers typically share related concerns about scalability.

A workshop held in January 2021 at the close of the project garnered positive interest from the international conservation community with requests for us to share models and suggest common database templates and calls to develop and publish data standards for the field [13]. Sharing data models and developing standards for conservation should be a high priority within the field to guide practitioners in modernizing practices. This work is essential for conservators to have searchable data suitable for computer analysis and if data interoperability is to ever become a reality. Best practices, guidelines, and thoughtful arguments in favor of model and ontology development have been discussed elsewhere in more detail [14][15].

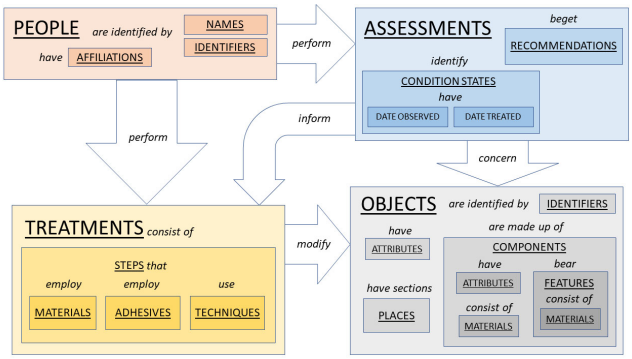

***Figure 4**. A sketch for modeling a new conservation database for Stanford Libraries Conservation Services unit.*

Stanford's conservation lab has continued to review and refine the LCD pilot project's model in preparation to build a new local database (4). To supplement our research, we have begun conducting interviews with colleagues from other labs to discuss their documentation systems. One potential next step the preservation community can take to improve data standards would be to ask institutions with existing conservation and preservation databases to share insight into backend system structures. Past calls

within the conservation field to share report templates have served an adjacent purpose of comparing documentation practices. A call to gather otherwise hidden database models would further expose common patterns throughout our datasets without any need to share sensitive data.

We must also carefully consider the reach and influence of our computing practices and how that should inform the scope of our data modeling activities. The collaborative character of our work demands that we consider the needs of conservators, preservation administrators, catalogers, curators, and scientists. As we develop standards, we should look to existing ontologies for guidance, but we should be wary of developing tunnel vision—a probable side effect in detail-oriented ontology projects. Overcomplicating the initial work of model development with overdependence upon any single ontology may obscure unique patterns and use cases in our data that we might otherwise miss.

References

[1] American Institute for Conservation of Historic and Artistic Works (AIC), "Code of Ethics and Guidelines for Practice," *American Institute for Conservation of Historic and Artistic Works*, 1994. [Online]. Available: https://www.culturalheritage.org/about-conservation/code-of-ethics [Accessed May 21, 2021].

[2] L. McCann, "Conservation documentation in research libraries: Making the Link with MARC Data," *Library Resources Technical Services*, vol. 57, no. 1, January 2013. [Online serial]. Available: https://journals.ala.org/index.php/lrts/article/viewFile/5220/6339. [Accessed: May 21, 2021].

[3] Information and documentation — A reference ontology for the interchange of cultural heritage information, ISO 21127, 2006.

[4] Library of Congress, "Understanding PREMIS," *Library of Congress*, 2017. [Online]. Available: https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf. [Accessed: May 21, 2021].

[5] A. Velios, N. Pickwoad, "Current use and future development of the database of the St. Catherine's Library Conservation Project," *The Paper Conservator*, vol. 29, pp. 39-53, 2005.

[6] E. Hobart, "Recording conservation information: The MARC 583 field in practice" *Library Resources Technical Services*, vol. 62, no. 3, July 2018. [Online serial]. Available: https://www.journals.ala.org/index.php/lrts/article/view/6730/9057. [Accessed: May 21, 2021].

[7] Library of Congress, "MARC 21 to BIBFRAME 2.0 conversion specifications," Library of Congress, Fields 5XX – Notes, 2019. [Online]. Available: https://www.loc.gov/bibframe/mtbf/ConvSpec-5XX-v1.5p.xlsx. [Accessed: May 21, 2021].

[8] CIDOC CRM Special Interest Group, "Definition of the CIDOC Conceptual Reference Model," *International Council of Museums International Committee for Documentation*, 2020. [Online]. Available: http://www.cidoc-crm.org. [Accessed: May. 21, 2021].

[9] M. Kay, *XSLT 2.0 and XPath 2.0 Programmer's Reference*, 4th Ed., Indianapolis, IN: Wiley Publishing, Inc., 2008. [E-book] Available: O'Reilly Media, Inc. e-book.

[10] G. Maxey, "Tinkering with custom XML parts," March 9, 2021. [Online]. Available: https://gregmaxey.com/word_tip_pages/tinkering_with_CustomXML Parts.html. [Accessed May. 26, 2021].

[11] Office Open XML File Formats, ECMA Standard 376, 2006.

[12] FORTH Institute of Computer Science, "X3ML Toolkit," Foundation for Research and Technology - Hellas Institute of Computer Science. [Online]. Available: https://www.ics.forth.gr/isl/x3ml-toolkit. [Accessed: May 26, 2021].

[13] University of the Arts London, Modelling Conservation Data: A workshop with Linked Conservation Data – Day 2, 2021. Accessed on: May 26, 2021. [Streaming video]. Available: https://youtu.be/M9kshqOmkM0.

[14] R. Sanderson, P. Ciccarese, H. Van de Sompel, "Designing the W3C Open Annotation Data Model," in WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference, Web Science 2013, Paris, France, May 2-4, 2013, H. Davis, H. Halpin, Eds. New York: Association for Computing Machinery, 2013. pp. 366-75.

[15] N. Noy and D. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems Laboratory, March 2001. [Online]. Available: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf.

Author Biography

Ryan Lieu is the Operations and Technology Specialist for Stanford Libraries Conservation Services where he manages documentation practices, digital workflows, and collection logistics. He holds a Master of Fine Arts from the School of the Art Institute of Chicago and is currently pursuing a Master of Library Information Science at the University of Wisconsin-Madison.