Linked Conservation Data

LCD Pilot report (phase 2) - 2020-2021

Athanasios Velios and Kristen St. John

Linked Conservation Data is funded by:

UKRI Arts and Humanities Research Council

Board reattachment pilot

Summary

The Linked Conservation Data Pilot was one of the main outputs of the second phase of the Linked Conservation Data project. The pilot created Linked Data for a common conservation treatment method from the records of multiple institutions in order to develop workflows, identify challenges and generate recommendations for future Linked Data use in Conservation. The pilot project steps were: developing research questions, selecting records, aligning terminology, modeling the data, and uploading to a portal. This report contains detailed analysis of each step including results, performance and recommendations for future pilots or Linked Data implementations. While the COVID-19 pandemic altered timelines and activities, the pilot was successfully completed within the allocated time frame. Several areas for future efforts have been identified including metadata description of records, closer engagement with non-experts for modelling records and further improvement of software tools to enable wider adoption of Linked Data in the Conservation field.

Introduction

As part of phase 2 of the Linked Conservation Data (LCD) project, the consortium decided to test the practices developed through the work of the first phase of the LCD project by initiating a pilot (https://lcd.researchspace.org/resource/rsp:Start) to integrate conservation data using Linked Data practices. This was one of the objectives for phase 2 intended to help us understand the complexity of the process when multiple organisations are involved with varied levels of familiarity with Linked Data. This text describes the work undertaken to complete the pilot alongside a commentary of the challenges met and recommendations for similar future activities.

Pilot contributors

Contributions from all consortium partners to other project outputs helped with the development of necessary aspects of the pilot (for example the terminology guidelines [https://github.com/linked-conservation-data/conservation-vocabularies/wiki]). Data from four consortium members were included in the pilot. These were:

- the Bodleian Library
- the Library of Congress
- The National Archives (UK)
- Stanford Libraries

Staff from each partner institution had participated in the first phase of the Linked Conservation Data consortium by attending workshops and webinars. Given the richness of documentation held by their labs, their willingness to provide resources (both in staff time and records), and

levels of participation in the Consortium, it was decided to focus the subject of the pilot on collections held by these members.

Subject and research questions

Publishing conservation records as Linked Data allows records from different organisations to be queried together. The pilot was intended to highlight this fact by answering research questions which cannot be answered by querying records from one organisation only. By bringing together sample datasets from four large and important conservation labs, it was possible to provide a **more representative sample** in comparison to each individual dataset. While many Linked Data projects include cataloguing data with descriptions of objects, rarely do they include descriptions of conservation activity.

The LCD pilot focuses primarily on conservation activity. The subject of the pilot is a common problem in book conservation: **board reattachment treatments**. Reattaching boards on historic books has been done using several different approaches and materials over time. Innovations in this treatment have proliferated over the past forty years with selective adoption by different labs. Understanding which labs have used which techniques during which time periods can be efficiently studied through integrated data. The LCD pilot was designed to allow such exploration by trying to answer the following research questions through queries on the linked datasets.

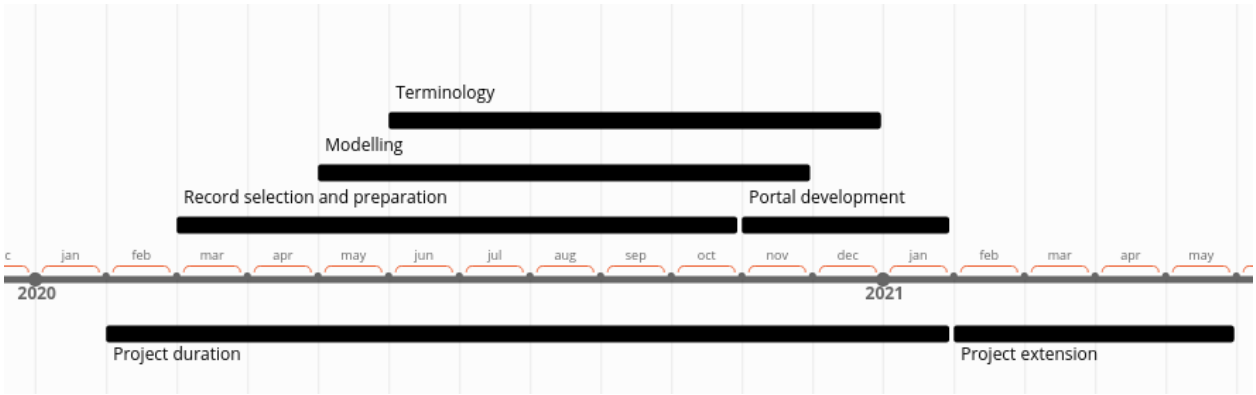
Pilot research questions

- What is the history of board re-attachment techniques over the last 50 years?
- Can we identify the periods that each repair material was being used?
- How do detached boards relate to other book condition types (e.g. spine re-attachments/repairs)?
- What types of records are needed to infer the impact of repairs to the study or history of books?

We aimed at answering these questions from records spanning 40 to 50 years. This would provide sufficient time-span to indicate change of practice over the years and also cover different types of a) vocabularies used and b) documentation practices.

Timeline

A detailed timeline of pilot activities is included as an Appendix to this document. Over here, an overall timeline is shown:



Record selection and preparation	March 2020 - October 2020
Modelling	May 2020 - November 2020
Terminology	June 2020 - December 2020
Portal development	November 2020 - January 2020

The long time-span of each task reflects the fact that some partners were only able to contribute to the pilot later in the project timeline.

Communication

In addition to the regular contact of the PI (Athanasios Velios) and Co/I (Kristen St.John) with the project members for the rest of the outputs, project members for the pilot (pilot working group) had monthly meetings starting in February 2020. The Research Administrator (Brigitte Hart) participated in the group meetings by keeping notes and recording action points. Documents were shared via the project's cloud storage space on Google Drive. Individual meetings with the Post Doctoral Research Fellow (PDRF) (Alberto Campagnolo) and partners contributing to the pilot were held periodically to discuss modelling development and terminology. These included Ryan Lieu (Stanford Libraries), Nicole Gilroy, Andrew Honey and Alice Evans (Bodleian Library), Elmer Eusman, Shelly Smith (Library of Congress), Sonja Schwoll, Holly Smith (The National Archives). The second PDRF (Steve Stead) contributed to specific meetings for additional modelling advice. Partner members from the British Museum (Dominic Oldman, Cristina Giancristofaro, Diana Tanase) contributed to the meetings related to the pilot portal.

Covid-19 pandemic - Challenges/Opportunities

Within a couple of weeks of the first meeting of the pilot, the increasing severity of the Covid-19 pandemic resulted in all participants shifting to working from home instead of in their labs. Some of them were moved to furloughed status, thus were not able to contribute to the project. This

led to delays at times and presented challenges to some partners in securing records for the pilot as well as finding time to work on the project. By the summer and into the fall, many participants were moving back to their labs at least part-time.

Although the pandemic was an unforeseen challenge, in some cases having staff working away from their labs provided an opportunity for greater staff participation. Pilot partners reported that compiling and researching terms as part of the LCD pilot was a productive task to be undertaken from home. They were able to engage several staff who may not have had the capacity to work on this during their previous work days in their labs.

Record selection and pre-processing

Roles

Record providers: Bodleian Library, Library of Congress, The National Archives, Stanford Libraries
Record processing: University of the Arts London, Stanford Libraries

Scope and sequence

Partners were responsible for selecting records of books with reattached boards to meet the criteria for the pilot. In many cases these records also held information about a) the condition of the object, b) the production techniques and c) the materials used, which covered the scope of the research questions posed. In some cases these records were held in free text format as part of conservation report documents, i.e. they were not structured records. In other cases these records were retrieved from in-house databases. Partners were able to also select images to accompany the selected records. Partner data existed in the following formats:

- Library of Congress: structured data, available in spreadsheet format
- Bodleian Library: free text conservation reports, text-based documentation forms on paper
- The National Archives: semi-structured data, available in spreadsheet format
- Stanford Libraries: scanned reports, structured data forms in individual .docx files

Text-based records primarily from the Bodleian Library, The National Archives and Stanford Libraries were manually or semi-automatically converted to structured data. In the case of the Bodleian Library this involved building an XML schema and manually populating XML documents based on that schema after reading the text. Stanford Libraries adapted narrative and semi-structured legacy reports using a transformation routine to convert .docx documents into XML documents. This transformation had been developed for in-house documentation in 2018. Processing the records from Library of Congress and The National Archives involved automatic transformations of data and keyword identification in short textual records in the spreadsheets.

Each partner contributed about 30-50 records covering the time-span of 50 years with the Library of Congress contributing a large number of records from recent years, reflecting recent staff efforts to organize information on treatments in a spreadsheet.

Results and performance

Records from all partners were selected and encoded in machine-readable formats successfully. Records from the Library of Congress and The National Archives were already available as structured data (in spreadsheets or databases) and therefore machine-readable. Records from Stanford and the Bodleian needed additional work requiring digitization and then encoding to be represented as structured data (as noted above) so this resulted in additional steps before modeling could begin.
Some partner members were able to respond quickly as soon as the pandemic restrictions were announced in the UK and the US. They retrieved paper records in the labs and bulk scanned them just before buildings were shut in February and March 2020. The teams were then able to work online using scanned .pdf documents. Others did not have the capacity to compile records quickly primarily due to lack of access and staff and delayed with selecting and encoding them. While converting the text records to structured records using the XML schema a wide range of detail was included beyond the requirements of the pilot research questions. This meant that the resulting records were rich, providing plenty of opportunity to examine modelling patterns and questions, but perhaps over-stretched our capacity for processing such records.

Scalability and recommendations

Summary points

- Structured records on digital formats allow easy processing
- Choice of documentation fields to be processed needs to be informed by the research questions to limit unnecessary work
- Particularly in legacy data a selection of metadata pointing to the full record seems a practical and feasible alternative to structuring complete records manually
- Objective criteria for identifying information which needs to be included in such metadata needs to be defined

Discussion

Due to the circumstances posed by the pandemic restrictions, partners spent a limited period of time selecting records. Without access restrictions it may have been possible to identify a wider range of paper/text-based records relevant to our research questions. This was not an issue with digital records. Keeping structured records in digital formats makes selection and further processing easier and more efficient (thus scalable).
It is also important to consider the objectives of integration in relation to the data selected. In retrospect, perhaps less effort should have been invested in producing a detailed XML schema and documents. These served a broader scope than that required by the research questions.

The flipside is that the resulting records were rich and illustrated the capacity and scalability of Linked Data methods within the pilot.
Manually converting text-based records from individual conservation reports to structured records was the most time-consuming aspect of this task. This is not scalable for legacy data and the recommendation is that, at least, new records are produced in a digital format following a schema within a database system. For legacy data, such conversion would need to happen at metadata level with enough detail to point to the full reports. In the case of board reattachment such metadata may include the identification of the book, the type of conservation work undertaken (i.e. board re-attachment) and the types of materials used.
A future goal of LCD is to identify systematic and objective criteria for selecting essential metadata fields. Materials and techniques (for both the object and the conservation work) are popular questions which provide significant pointers.

Terminology work

Roles

Terminology providers: Bodleian Library, Library of Congress, The National Archives, Stanford Libraries
Terminology support: University of the Arts London, Stanford Libraries

Scope and sequence

As reports were selected, partners surveyed the records to extract terms to develop local vocabularies. The local vocabularies were to be encoded as SKOS to be integrated into the dataset of the pilot. The expectation was that each vocabulary would need to be aligned to a broader thesaurus through equivalence relationships so that results appear regardless of the terminology used for searching. This task was informed by the terminology guidelines also produced as part of the project.
After extracting terms, synonyms were identified and terms which referred broadly to the same concept were grouped (for example the Bodleian Library records included many variations of Japanese paper but opted to group them all under one heading). Prepared terminology was communicated and reviewed on spreadsheets which also held unique identifiers for concepts/terms in preparation for Linked Data encoding.
As a subsequent step to allow searching across different vocabularies these terms were aligned with terms in the Getty Art & Architecture Thesaurus as well as the Language of Bindings thesaurus. The Getty AAT has been identified as a hub thesaurus for vocabulary alignment already from phase 1 of the project. The intention was that any terms not included in the AAT would be submitted as new entries.

Results and performance

In the initial project plan we had estimated a shorter period of time for collecting and analysis of vocabularies from partners. The length of the planned period was under-estimated significantly and the vocabulary analysis was one of the most resource intensive efforts in the project. It often involved large numbers of staff members from the partner organisations and extensive discussions on the nature of the use of words in documentation records. While the terms from each partner were encoded as SKOS and successfully included in the pilot portal, we accepted some limitations:

- Many terms were not included in the Getty AAT and therefore could not be aligned with it. This meant that two different terms in use in two different partners but pointing to the same concept could not be jointly queried if they were not included in AAT. Our plans for submitting these terms to the Getty AAT were not materialised due to the overall delay of this task and the requirement for extensive metadata accompanying newly submitted terms to the Getty AAT. A local hub thesaurus could have been created to accommodate such terms but we decided against it so that we can draw attention to the importance of alignment of local terms.
- Terms used in local vocabularies in previous decades become obscure when members of staff retire and the pattern of use of these terms is no longer recognised. Partners highlighted the value of having such terms documented through local scope notes and alignment with externally maintained reference thesauri.
- The implementation of the pilot on the ResearchSpace platform meant that a set of tools provided out-of-the-box could be customised and used for cross-searching terminology within the limited scope of the pilot.

Partners indicated that the internal discussions undertaken for the analysis and documentation of terminology in records were valuable for the understanding of the state of documentation records within the organisation and gave important pointers for future practices. They also indicated that the terminology analysis was the most complex and time-consuming task of the project for them.

Scalability and recommendations

Summary points

- Consider limiting terminology analysis to reflect modelling requirements (as opposed to complete coverage) to reduce time-consuming terminology work
- Resolve terminology conflicts before modelling is finalised
- A system for documenting the alignment of conservation vocabularies is needed
- Use skos:broadMatch when aligning composite terms from local vocabularies with thesauri which do not allow composite terms

Discussion

While the exercises on the analysis and documentation of terms undertaken by the partners were considered useful, the question around the scale of the effort was posed again. The delay caused during this stage of the project was partly due to the fact that terminology discussions were not focused on the limited number of fields required for the pilot, but were more broad. This allowed partner teams to assess the scale of the task of processing terminology for legacy records. It also flags the importance of prioritisation of the types of fields included in integration projects like this pilot.

Another interesting discussion is whether terminology analysis and encoding should happen before or after modelling activities. The justification of undertaking terminology alignment as a first step in integration is:

- Well modelled datasets are not possible to fully integrate unless terminology is aligned. This is because in expert domains, like conservation, much of the querying is done based on the level of detail offered by vocabularies and thesauri, as opposed to the level of detail offered by a generic ontology such as the CIDOC-CRM.
- Composite conservation terms (e.g. *leather corner turnins* of a quarter binding) correspond to multiple types of observation (e.g. material: *leather* and cover feature: *corner turnin*) and need to be modelled separately.
- Examining the terms used in fields and records allows familiarity with their content which is useful for the process of modelling.

Having said that, undertaking the modelling task before the terminology analysis, would have provided a robust criterion for selecting the groups of terms needed to be analysed and encoded. As a recommendation we still maintain that terminology should be a starting point, but perhaps considered jointly with modelling with the purpose of limiting the amount of work required. This is particularly true for legacy records and historic use of terms. It will also allow the terminology work to become more scalable with maximum effect even for partially analysed vocabularies.

The flexibility offered by ResearchSpace allows planning of complex tools for querying vocabularies and allowing versioned control of their alignment with hub thesauri. A domain expert vocabulary hub for conservation is a significant task that will benefit future projects. Further recommendations for processing vocabularies can be found on the project terminology guidelines.

The work on vocabularies was undertaken by the terminology providers and then discussed in regular pilot working group meetings with the terminology support teams. In retrospect and without any travel limitations in place, it would have been preferable to undertake this engagement as originally planned: i.e. in person within the premises of the terminology provider. Having done so would have helped with communicating the reasoning behind choices made more regularly and possibly speed up the process.

The pilot working group initially focused on the SKOS properties skos:exactMatch and skos:closeMatch to align vocabularies in particular in connection with the AAT. Thesauri like the AAT avoid including composite terms in their hierarchies when the same meaning can be communicated by combining two or more terms. This is not the case for local vocabularies in conservation where composite terms are often used to populate database fields, thus making the use of the above properties difficult. Therefore the pilot working group adopted the additional

skos:broadMatch as a good way of addressing this problem: positioning composite terms of local vocabularies as narrower terms under multiple parents in the AAT (typically as many as the components of the composite term). This meant that searching with a local known composite term will retrieve results from the local database and indicate steps up in the AAT hierarchies to attempt and broaden the search results. We note that an alternative solution to this problem would be to split the composite terms to separate database fields. For example, *leather corner turnins* would be accommodated into two fields instead of one: a) turnin type: corner turnin and b) turnin material: leather. However, we did not want to interfere with the local schemas in partner databases.

Modelling

Scope and sequence

The CIDOC-CRM was chosen as the ontology to guide modeling of the conservation data provided by the partners. In this ontology, classes are used to describe items (e.g. this board, this spine) and properties connect these items through relationships (e.g. item board *consists of* material wood). Modeling is the conversion of the conservation data into these patterns and is critical so the data from all pilot partners conform to the same schema, to enable integration and can be searched across.

These models were primarily produced by members of the consortium from the University of the Arts London (Alberto Campagnolo, Athanasios Velios, Stephen Stead) and Stanford Libraries (Ryan Lieu). Other consortium members had limited contribution to the modelling task. The outcome of this task was a number of transformation scripts which could receive contributed datasets and produce Linked Data datasets ready to be inserted to the pilot web application/portal.

Two different paths were chosen to model data. Datasets from Bodleian were manually transcribed and encoded into XML documents which followed an XML schema built to store the variety of the information in the records. The schema was used to undertake initial modelling work using the 3M software (https://isl.ics.forth.gr/3M/) which validates modelling choices based on the CIDOC-CRM ontological rules. Following this validation stage and in order to maintain complete control over the data, an XSLT script was built to transform the XML documents into RDF/XML.

A similar process was used for the records from The National Archives and the Library of Congress. In these cases the transformation script worked from tabular records and no manual inputting was necessary. Tabular records were transformed into XML documents by importing spreadsheet files into the Oxygen XML Editor application. These XML documents were then transformed with XSLT to an ad hoc schema similar to that used for rekeying the Bodleian's dataset and transformed once more into RDF/XML with the XSLT script written for the Bodleian dataset.

Some of Stanford Libraries' records were already encoded as XML documents with a different XML schema structure. Stanford's handwritten paper documents were rekeyed into XML-integrated electronic forms to match the other XML documents in Stanford's dataset.

Another XSLT script was built to transform these initial XML documents from Stanford into an ad hoc intermediary schema optimized for better alignment with the CIDOC-CRM and similar to the ad hoc schema used for records from the other three contributing institutions. The mapping of these documents to CIDOC-CRM was done on 3M. This resulted in a transformation script provided by 3M which was then used to convert the XML documents into RDF triples. Discussions around the accuracy of the semantics for the mappings were done online by rendering the models using the CRMVIZ tool (https://github.com/natuk/crmviz), which was specifically developed for this project and which allows the visualisation of sample mapped records for easier comprehension.

Results and performance

Several iterations of data modeling were required before the two models could be queried effectively. Some of the issues that arose during modeling include:

- Not all components needed to answer the research questions were included in the initial selection of treatment reports. For example the exact date of treatment was sometimes not given, but could be deduced from the date of the documentation.
- Provision for keeping term identifiers consistent was not initially made across datasets. This meant that although both tasks of terminology analysis and modelling were completed successfully, they could not work with each other due to the lack of common identifiers until they were rectified.
- The scale of the modeling tasks was ambitious due to the level of data to be modelled. A more restrictive selection of records, focussing only on the research questions, would have limited the amount of time and effort needed at modelling stage. Such economy would have also been the case if metadata was considered as part of an exercise in answering broader research questions. A rough estimate would indicate that the research questions could have been answered in some cases with less than half the available fields.

The two different pathways to the modeled data provided information on potential approaches that different institutions might follow for their data. Due to the complexity of the CIDOC-CRM, a streamlined model or pre-developed profile would have made the process of modeling more efficient. By profile we refer to a selection of classes and properties of the CIDOC-CRM relevant to conservation (as opposed to using the full model). A profile would have limited the number of revisions required and ensured that all components needed to answer the research questions were included in the data selection.

A pre-developed profile might have also enabled more partner engagement in modelling. The plan of the project made provisions for partner engagement at all stages of the transformation of the datasets to Linked Data datasets. Such engagement was limited during mapping the datasets to the CIDOC-CRM. The main factor for this was that some partners found the ongoing terminology tasks challenging or time-consuming and did not have the resources available to take on additional complexity around modelling. Some partner feedback points to the lack of engagement during this part of the project. Also, keeping the modelling team small among project members with significant experience with modelling meant that we could undertake the work faster but without communicating our experience to partners as inclusively.

Scalability and recommendations

Summary points

- Develop a conservation-specific CIDOC-CRM profile and related didactic material.
- Identify criteria for selecting and sharing metadata when resources for modelling full datasets are not available.
- Focus on questions of interest when modelling data and do not model for the sake of model completeness.
- Assess the extra cost of deducting data from existing data, i.e. whether questions of interest can be answered indirectly through associated data.
- Engage with conservators who do not have modelling experience when mapping conservation datasets.

Discussion

A pre-developed conservation profile would make modeling more efficient, accessible and scalable for conservators. In order for conservators to make the first steps in comprehending the CIDOC-CRM easier, a pre-selected subset of classes and properties should be made available as a conservation-related profile of the CIDOC-CRM (also see the recommendations from the modelling working group survey: https://www.ligatus.org.uk/lcd/output/248). The CIDOC-CRM offers classes and properties, the number of which often overwhelms newcomers. CIDOC-CRM classes and properties alongside those from its extensions cover a seemingly limitless number of scenarios, and this proves challenging to those getting started in modelling as well those introducing them through didactic material. The profile should be accompanied by concise, non-technical, and domain-specific training material to help lower barriers to entry for those wishing to adopt systems that use the CRM.

A minimal metadata schema pointing to conservation documentation as opposed to a fully encoded dataset would allow sharing conservation records with relatively little effort. Specifying the criteria for establishing such metadata schema requires more discussion and risks being fragmented depending on the conservation specialisation considered.

Engagement during the modelling task between project members with modelling expertise and project members with expertise in the datasets would have benefitted the team. The engagement could take the form of a dialogue between the modeller (with expertise in CIDOC-CRM) and the conservator (with expertise on the data to be modelled). Such dialogue ensures more accurate modelling decisions and also sharing of modelling expertise.

Portal implementation

Roles

Infrastructure and support provider: British Museum
Template and querying: University of the Arts London, Stanford Libraries

Scope and sequence of this work

This task included uploading modelled datasets and associated vocabularies on an instance of the ResearchSpace platform (http://researchspage.org). ResearchSpace is a Linked Data platform which has been used extensively for Linked Data produced based on the CIDOC-CRM ontology. It offers pre-built templates for the entities encoded in our project and a range of flexible tools for querying and presenting the data. The main objective was to reflect the pilot research questions into encoded queries that the platform could answer based on the underlying data. Another objective was to show how conservation reports can be enriched with queries on the encoded datasets pointing to the observations which support the narrative of the reports.
The steps undertaken to implement the pilot on ResearchSpace are presented here sequentially, but in some cases iterations of these steps were required:

- Imported transformed data
- Modified ResearchSpace templates to add extra details
- Imported photographs and linked them to items
- Built search pages for the research questions
- Built narratives based on queries in the data

The above consisted of technical work which depends on familiarity with the ResearchSpace architecture and a querying language for RDF called SPARQL (https://www.w3.org/TR/sparql11-query/).

Results and performance

Importing records and building queries can be considered as the ultimate test for integration as querying and returning results can prove the success of the project. It was possible to encode the pilot research questions using SPARQL queries and present the results in the form of timelines and diagrams which assisted with the comprehension of the answers. Existing familiarity with the ResearchSpace platform and support from the British Museum as part of their role in the project, meant that this work was put in place efficiently. Narratives from conservation reports enriched with results from the encoded data were also produced using ResearchSpace. Apart from the British Museum, the University of the Arts London and Stanford Libraries, other partners made limited contributions during this task due to the technical nature of the work. The implementation was demonstrated to the consortium with explanations on its functionality. While the ResearchSpace team aims to make the software easy to use for users without expertise on Linked Data, customising the system and building the required queries remains a job for a technical expert.

Scalability and recommendations

Systems like ResearchSpace are designed to be scalable. Therefore the volume of work depends on the required queries and customisations that are needed for the purposes of a specific implementation. They do not depend on the size of data. Some important observations and recommendations follow:

- Clarify biases in underlying data
- Building complex queries is a good way of discovering faults on the underlying datasets.

Discussion

The results of queries reflect biases of the underlying data. The risk of interpreting the results as representative information that can be generalised needs to be mitigated by clear sign-posting about data sources. For example, the Library of Congress dataset was relatively large and focused in the period between 2010 and 2020. This may be misunderstood to indicate that there was significantly more conservation activity for board reattachment during this period as opposed to the 40-50 year timeframe covered by other pilot participants which is not necessarily true. In reality, it is the skewed sample which creates this peak of activity.

Building the pilot portal was a good exercise for validating the quality of our data transformations from the original sources to the Linked Data datasets. Querying the data following integration offers the opportunity to consider the way that it has been modelled. If the results of queries are surprising then it is possible that mistakes have been introduced during data modelling. As such, one could consider the process of building queries on the integrated data as an evaluation of modelling. It is likely that overlapping modelling tasks with implementing queries on a Linked Data platform would lead to fewer model iterations and more efficient development. We recommend that modelling and establishing queries in the portal should be done in parallel. This follows the experience of reviewing our models through querying and discovering the underlying problems in the transformed datasets.

ResearchSpace is a significant project in the field of Linked Data which has enabled communities to share data and build new connections and knowledge with them. While the flexibility of the system is evident, allowing great variety of customisations, at the moment the default configuration did not cover all our requirements and we did not have the time and resources to implement them as part of this pilot. These requirements are summarised here:

- The default advanced search tool does not automatically scan the models in the available data. This means that a user cannot build their own queries, but has to rely on pre-built queries provided by an administrator. Configuring advanced search to enable custom queries is possible, but required more time than we had available.
- The custom querying tool also does not observe the hierarchy logic of the CIDOC-CRM. For example features like class and property hierarchies and property inheritance need to be implemented manually although they are the core of the theory of integration with the CIDOC-CRM.
- The default vocabulary manager does not allow querying reconciled concepts across vocabularies. For example, our local vocabularies were reconciled with the Getty AAT, but it was not possible to use that reconciliation automatically when searching terms.

Resources

Members of the working group who contributed to the delivery of the pilot are listed in the Appendix: List of people. It is difficult to quantify the amount of time spent in pilot related activities in the different partner institutions. Pilot work initiated in these institutions resulted in additional discussions around documentation practices which, although not strictly related to the pilot, were very much within the broad scope of Linked Conservation Data. For the two main partners the committed time is reflected in the project posts who undertook the bulk of the work:

- Athanasios Velios (PI, University of the Arts London): 0.2FTE
- Kristen St.John (Co/I, Stanford Libraries): 0.1FTE
- Alberto Campagnolo (PDRF, University of the Arts London): 0.5FTE
- Stephen Stead (PDRF, University of the Arts London): 0.2FTE
- Ryal Lieu (0.2FTE, Stanford Libraries):

Tools for producing Linked Data are becoming more mature, but still require further development before they can be used independently by working conservators. During the pilot, we observed that conservators were able to contribute more easily when working with familiar tools. Improvements to tools such as spreadsheet editors could also increase participation. For example, Google Sheets was used extensively during the project as it allows collaborative working remotely which was a necessity during the pandemic. The templates provided for vocabulary alignment are utilising Google Sheets. Productivity would increase if it was possible to lookup terms in external thesauri (such as Language of Bindings and Getty Arts and Architecture Thesaurus) from within Google Sheets. An add-on for Google Sheets enabling such activity should be fairly straight-forward to produce and may be worth considering in the future.

ResearchSpace offers huge potential for demonstrating, creating and querying conservation data especially for conservation documentation which does not always conform to strict schemas. Current ResearchSpace tools such as the Knowledge Map editor are not familiar to conservators and present a steep learning curve for adoption. It is possible, and this project supports the direction that conservation records as Linked Data can be produced in ResearchSpace, but further usability studies and development of additional functionality are needed to lower the barrier which exists currently.

Appendix: List of people

Consortium members directly engaged in the delivery of the pilot are:
Alberto Campagnolo (University of the Arts London)
Elmer Eusman (Library of Congress)
Alice Evans (Bodleian Library)
Cristina Giancristofaro (British Museum)
Nicole Gilroy (Bodleian Library)
Brigitte Hart (University of the Arts London)

Andrew Honey (Bodleian Library)
Artem Kozlov (British Museum)
Ryan Lieu (Stanford Libraries)
Dominic Oldman (British Museum)
Sonja Schwoll (The National Archives UK)
Holly Smith (The National Archives UK)
Shelly Smith (Library of Congress)
Kristen St.John (Stanford Libraries)
Stephen Stead (University of the Arts London)
Diana Tanase (British Museum)
Athanasios Velios (University of the Arts London)

Partner members who contributed to tasks within partner organisations are:
Jennifer Evers (Library of Congress)
Alan Haley (Library of Congress)
Linnea Vegh (Library of Congress)

Appendix: Pilot timeline

November 8, 2019	Partners meet to scope the pilot. Decision made about the type of treatment to focus on.
February 2020	Phase II of AHRC grant period begins.
February 25, 2020	CIDOC-CRM SIG meeting (Athanasios Velios presenting the Model for Plans and the proposal for Negative Properties extension).
March 6, 2020	First pilot call, calls follow every 4-6 weeks on Fridays.
mid-March, 2020	Due to the COVID-19 pandemic, all institutions transition to some form of remote working or furlough staff. Access to records is a challenge for some partners.
March 31, 2020	Modelling working group meets.
April 10, 2020	Partners assess what records are available and how to share information.
May 11, 2020	CIDOC-CRM SIG meeting - (Athanasios Velios and Stephen Stead report on progress of Negative Properties extension).
May 22, 2020	Alberto Campagnolo develops a schema for Bodleian free-text records and Stanford Libraries discuss vocabulary sprint concept for group terminology work.
June 12, 2020	Terminology work: Alberto Campagnolo generates term list for Bodleian, discussion with Dominic Oldman about ResearchSpace.
July 15, 2020	3M meeting (introduction to the 3M mapping tool).
July 17, 2020	Updates on vocabulary work, modeling work continues.
Aug 14, 2020	Updates on vocabulary work, modeling work continues.
September 4, 2020	Modeling work continues with Alberto Campagnolo working on Bodleian model and Ryan Lieu using 3M; Library of Congress team discusses vocabularies; Dominic Oldman discusses removal of personal data from records.

September 25, 2020	Pilot meeting - progress update.
October 16, 2020	Francesca Whymark expresses the interest of the British Library to contribute; the Library of Congress and the Bodleian teams continue work on terms; Sonja Schwoll gets sample for the National Archives records; Modeling: Bodleian model evaluated positively; agreement to apply similar model to the Library of Congress records, Stanford Libraries fix errors on identifiers, remodel data in 3M.
October 31, 2020	Final versions of pilot vocabularies ready.
November 13, 2020	Records to ResearchSpace uploaded. Discussion of vocabularies alignment. Alberto Campagnolo and Ryan Lieu discuss how their models will align. Partners look for images.
December 4, 2020	Athanasios Velios uploads data to Github and ResearchSpace - work on queries begins. Questions on vocabulary alignment options within ResearchSpace. Library of Congress work on additional terms.
January 8, 2021	Demo of ResearchSpace by Athanasios Velios. Work on pages with narratives in progress.
January 29, 2021	Final pilot meeting and evaluation of effort.