CIDOC CRM Negative Properties Test Dataset

Editor(s): Eero Hyvönen, University of Helsinki, Finland
Solicited review(s): Öyvind Eide, University of Cologne, Germany
Open review(s): Robert Sanderson, Yale University, USA

Stephen Steadabc
aPaveprime Ltd, 35 Downs Court Road, Purley, Surrey, CR8 1BF, UK, +44 (0)7802 755 013, steads@paveprime.com
bLigatus, University of the Arts London, Chelsea College of Arts, 16 John Islip Street, London, SW1P 4JU, United Kingdom
cDelving BV, De Ruijterstraat 82, 2518 AW, Den Haag, The Netherlands

Abstract. This data set is intended for the testing of software tools that combine and integrate sets of semantically rich data. It exercises the ability of the software to integrate data that has been recorded at different granularities and using different recording approaches. It allows testing of the ability to detect contradictions when using positive, negative and class level assertions. The data set comes with definitions of the content and the software required to generate additional sets of similar test data.

Keywords: CIDOC CRM, type properties, negative type properties, semantic data, software testing

1. Introduction

The dataset is designed for testing the capabilities and efficiency of software intended to draw conclusions about complementary sets of semantically rich data. Such software should be able to deal with both positive and negative statements about the material documented. It should be able to navigate multi-level, poly-hierarchical thesauri as part of its ability to merge data from different sources and be able to determine if a data set has missing elements that can be inferred by using one or more of the others. Finally, it should be able to detect and identify inconsistencies or contradictions between the different sources.

The Negative Properties Test Dataset is available on GitHub (https://github.com/linked-conservation-data/negative-properties-test-data) under a GNU General Public License v3.0. The dataset includes the semantically marked up data, the source data in CSV format and the program code to generate new versions of the data, together with full documentation. The data was generated in September 2020 and includes the version number (1) in the three output files' names.

The dataset consists of records for 28000 fictional (i.e., artificial generated) binding structures for manuscripts. Each was generated as having been constructed from a set of components and features following a set of reasonably realistic "rules". These complete physical manuscripts were then subjected to two generations of imaginary loss events to give three different states that they could be recorded in: Original, Intermediate (after the first generation of loss event(s)) and Modern (after the second generation of loss event(s)).

Each of these three states was then used to generate records that simulated different recording methodologies. Individual physical manuscripts are identified by a shelf-mark. The shelf-mark consists of a region name (selected from Mediterranean province names) and a four-digit number. There are 28 different region names (available as a CSV file). The four-digit numbers do not form continuous uninterrupted sequences.
The Original state was recorded by a series of seven lists, each recording those physical manuscripts that were missing a particular component or feature set at the time of recording. These lists can be seen as recording this Original state, but of course subsequent development could reveal other intermediate states, if that is desirable. The lists were for Spines, Boards, Clasps, Spine Features, Book-block, Features, Recesses, and Board Features. The lists were converted into an XML file (NilList1) using a proposed extension of the CIDOC Conceptual Reference Model v7.1 [1] that enables the documentation of negative statements about a subject [2,5]. The Original state records identify the physical manuscripts using only their shelf-mark.

The Intermediate state of some of the physical manuscripts (10979 in total) was recorded using an approach that incorporated noting both the presence of particular types of components and features, and the absence of types of features. The resulting data was converted into an XML file (Ant1) using both positive statements about the presence of types of components and features and negative statements about missing feature types. These positive and negative type-level statements also use the extension of the CRM [2]. The Intermediate state records identify the physical manuscripts using only their shelf-mark.

The Modern state of the physical manuscripts is documented using a comprehensive record of all components and features present. The documentation provides unique identifiers for each component and feature and explicitly delineates the relationships between them using the CIDOC Conceptual Reference Model v7.1 [1]. The Modern state records identify the physical manuscripts using modern identifiers as well as their shelf-mark; this ensures that the tested software can correctly navigate across different naming conventions. The resulting XML file (Full1) does not use the type-level statements from the CRM extension.

2. Thesaurus construction

The multi-level, poly-hierarchical thesaurus is inspired by The Language of Bindings Thesaurus (LoB) [4]. The thesaurus is constructed to provide a small set of terms that are related in a poly-hierarchical manner: that is, terms may appear in more than one branch of the hierarchy. In addition, terms may appear at different levels in different branches of the hierarchy (thus also making it multi-level). As an example, the term Clasp-Strap Recess appears at the third level under Feature:Recesses and at the fourth level under Feature:Board Features:Board Clasp Constituent which is also Board:Board Features:Board Clasp Constituent.

The Modern and Intermediate state records use all levels of the hierarchies, but the Original state records do not use the lowest level. This ensures that the tested software has actually navigated the thesaurus when combining the three documentation sources.

The thesaurus is supplied as an XML file using only the CIDOC Conceptual Reference Model v7.1 [1] E55 Type and $P127 \text{ has broader term (has narrower term)}$ constructs.

3. The dataset generation software

The dataset is generated using an R program (Generate Negative Properties Test Data.Rmd). It first loads the region names from a CSV file and then uses them to generate the shelf-marks. The next section of the program creates the physical manuscript records including the Manuscript Identifier (MID), Component Identifiers (CID), and Feature Identifiers (FID). It also identifies the elements that are missing or lost. The Manuscript Identifier (MID) actually encodes Component presence, absence, and missing or loss status as an aid to assessment. However, this encoding should not be considered by the software to be tested and so is not documented in the test briefing, only in the program documentation.

The next two sections of code partition the data. The first partition consists of physical manuscripts that have a Book-block, Spine, two Boards and no Clasps. These are collectively known as "2390" records (from their encoding in the MID) and form the majority of records (over 75%). The second partition consists of all other records. The final three segments of code create the XML output files for the three recording states (Modern, Intermediate and Original).

There is a deliberate error in the implementation of loss documentation. In the Original state, physical manuscripts that have one missing Board and one lost Board are marked as having no Board Features. Obviously, Board loss events do not affect the presence of Board Features and this therefore generates some contradictions between the Original state documentation and the later documentation states. The number of occurrences of such contradictions is low (in version 1 there are 6 such contradictions) and this therefore presents a rigorous test of the software’s capabilities.

4. The XML format

The output files follow a well-understood CIDOC Conceptual Reference Model XML format. In this format following the XML prolog and style-sheet, the root element is named CRMSet. The first level child element contains the record for each documented real-world entity and is delineated with the
E1.CRM_Entity tag. Each instance of a CRM class is declared with two elements: a document-wide unique literal marked as its Identifier and its class name marked as in_class. If the instance occurs more than once in the document, then using the same Identifier will allow the instances to be connected. Class names are formatted following the naming in the CRM standard document with the space between the “E” number and the text replaced with a period (".") and any spaces replaced with underscores ("_"). (for instance E22 Human-Made Object becomes E22.Human-Made_Object).

A property link to an instance of a class is done using an element, named for the property, that encloses the other class instance. The element name consists of the property identifier (Pxx, TPxx, NTPxx etc.) followed by a letter indicating in which direction the property should be read: “F” indicates that the property is acting from Domain to Range and “B” indicates that it is acting from Range to Domain. This first segment is followed by a period (".") and then the name of the property for the indicated direction: so, if “F” then the un-bracketed part of the property name is used and if “B” then the bracketed part of the property name is used. Again, any spaces are replaced with underscores ("_"). So, for example, using the property P1 is identified by (identifies) from Domain to Range would produce an element name of P1F.is_identified_by and from Range to Domain would produce an element name of P1B.identifies.

Software for converting this format into RDF is available from the Institute of Computer Science, Foundation for Research and Technology – Hellas under a Creative Commons Attribution-Share Alike 3.0 license [3]. A copy of this software and documentation (CidocXML2RDF.rar) is included in the 3 Data generation notes and program directory.

5. Physical manuscript construction axioms

The dataset was generated using a set of rules or axioms that determined which components and features were present on each physical manuscript as well as what happened to the components during its simulated lifetime. The axioms are documented alongside the program that implements them and are summarized in this section.

Elements that were lost in the first generation of loss events are termed “missing” and those lost in the second generation of events are termed “lost”.

A Book-block (a collection of leaves) is always initially present. It may be missing (0.5%) or have been lost (0.5%). A Spine may never have existed (4%) but, if it did exist, it may be missing (0.5%) or have been lost (0.5%). If the Book-block is missing, then the Spine may never have existed (4%) or be missing (96%). If the Book-block is lost then the Spine may never have existed (4%), be missing (0.5%) or be lost (95.5%).

The physical manuscript may have two boards (either side of the Book-block), that are individually identified. It may never have had boards (10%). If it had boards then each board may be present (87%), missing (1.5%) or lost (1.5%). This means that they could both be missing (0.5%), lost (0.5%) or present (86%); or Board 1 could be present with Board 2 missing (0.5%) or lost (0.5%); or Board 2 could be present with Board 1 missing (0.5%) or lost (0.5%) and finally Board 1 could be missing with Board 2 lost (0.5%) or Board 2 could be missing with Board 1 lost (0.5%). These proportions are true if the Book-block is either present or lost but if the Book-block is missing then the only evidence for the physical manuscript is the Boards. Therefore at least one must have survived the first generation of loss events. Consequently, the proportion of both Boards being present rises to 96.5% while the chances of never having Boards or both being missing falls to 0%.

Boards may have one (3.5%) or two (1.5%) Clasps. When Boards where never present there could be no clasps and if they are both missing there was never any evidence for clasps. If both Boards have been lost, then the clasp or clasps will also have been lost. Single clasps may be missing (0.1%), lost (0.1%) or present (3.3%). Where there are two clasps, they can both be present (1%), missing (0.1%) or lost (0.1%). In addition, they can each have had different fates with one present and the other missing (0.1%), one present and the other lost (0.1%) or, finally, one missing and one lost (0.1%).

A Book-block may feature zero (60%) or two (40%) Knife-Cut Recesses. Book-blocks that also have a Spine, may also feature either, Sawn-In Recesses (20%) or, V-Shaped Recesses (20%), but never both. These are also recorded on the Spine using the same unique identifiers in Modern records. Both Sawn-In Recesses or V-Shaped Recesses are either absent, or present in pairs.

In combination this means that a Book-block may have zero, two or four features.

In addition to the Sawn-In Recesses or V-Shaped Recesses from the Book-block, Spines may also have Adhesive Recesses and Sewing Recesses. There may be zero (70%), two (20%) or three (10%) Adhesive Recesses. If there are two or more Adhesive Recesses, then there are never more than four Sewing Recesses. Three Adhesive Recesses are never found on Spines with Sawn-In Recesses or V-Shaped Recesses.

There may be zero (50%), four (30%) or six (20%) Sewing Recesses. If there are four Sewing Recesses, then there are never more than two Adhesive Recesses and if there are six Sewing Recesses there are never any Adhesive Recesses.
In combination this means that a Spine may have zero, two, three, four, six or eight features.

Boards that have Clasps have Board Clasp Constituent (BDCLC) features. In addition, Board 2 may also have a Clasp-Strap Recess for each Clasp. About one in ten Clasp(s) have Clasp-Strap Recesses and if two Clasps are present then both will have Clasp-Strap Recesses or neither will.

6. Conclusion

This is the first dataset to provide software implementors with a bench-markable, semantically rich dataset for testing both capability (i.e., functionality) and capacity (i.e., performance). It includes the software to create new instances (versions) of the dataset, so it is easy to ensure that solutions are not over-fitted. While it is supplied as vanilla XML, the provided utilities allow the generation of corresponding RDF in a form that suits the local test environment. The thesaurus is provided as explicit, machine-readable axioms, making it a 2* resource. Its nature as a self-contained artificial dataset means that it should not aspire to link to the rest of the Web as it will cause wide-spread confusion.

7. Acknowledgements

This work was partly funded by the Arts and Humanities Research Council in the UK as part of the Linked Conservation Data project.

8. Bibliography

[3] FORTH, CIDOC CRM Xml to RDF Converter, FORTH, Crete, 2010