Tracking the heterogeneity in the provenance of RDF triples in Cultural Heritage Objects Description

Loic Jeanson, Emilio M. Sanfilippo, Florent Laroche

Tracking the heterogeneity in the provenance of RDF triples in Cultural Heritage Objects Description

Loïc Jeanson[1], Emilio Sanfilippo[2] and Florent Laroche[3]

LS2N, Université de Nantes - Centrale Nantes, 1, rue de la Noë 44321 NANTES
Cedex 3
loic.jeanson@ls2n.fr

Abstract. The variety of sources for data and the different processing applied during extraction or subsequent processing can lead to uncertainties about the data integrity in our data storage system. This paper presents an approach towards tracking data provenance, in the scope of CIDOC-CRM concepts. ...

Keywords: Provenance, Alteration, Tracking, CIDOC CRM, Heterogeneity

1 Introduction of our case study

In the ReSeed project[1], an interdisciplinary team builds up methods and tools for the digital modelling of cultural heritage. The team consist of historians, curators, heritage experts, mechanical engineering, and IT scientists. Our database builds up from various information sources, resulting in data integration and data processing with various techniques. In order to assess the quality of the model we are developing, we needed the model to be as transparent as possible for other researchers or practitioners. The need for a way to track our data and data heterogeneity in several dimensions emerged. In this paper, we present our needs and ways of solving this issue as a first step towards a more robust data tracking system.

The ReSeed project, supported by the ANR (Agence Nationale pour la Recherche - French National Agency for Research), originated after a few projects from two communities: standard heritage practitioners methodologies and mechanical and information engineering [2][3][4]. The aim is to develop usable tools and techniques to support the work of cultural heritage experts dealing with digital sources. Particular attention is given on the 3D representation of heritage objects.

In order to develop efficient tools and techniques, prototypes are first developed and validated against use cases. Our approach is structured around three case studies showing a variety of problems and needs. The first one focuses on the Observatory du Pic du Midi de Bigorre, a science station in the Pyrénées, located at 2877m of altitude. The Observatory is composed of several buildings, which were built during different phases. The second application studies astronomical telescopes that are produced in series even though they partially differ from each other. The telescopes thus share some similarities. Finally, the third case study is a physical model of a metal building frame of a colonial house, belonging to a museum.

We base our modeling approach on CIDOC-CRM, a well known modelling system, like other works in the domain of cultural heritage (among others [5] [6] [7]). In order to meet the specificity of our project, we need however to adapt (a small portion of) this model to our requirements, as we will see. Before presenting our approach in section 3 and more specifically the ontology in section 4, we first present the various types of data heterogeneity we deal with in section 2 .

2 The heterogeneity

2.1 Heterogeneity in the information carrier

According to CIDOC, information carriers, represented by the class *InformationCarrier* (E84), are physical objects carrying information objects (class *InformationObject* E73) in a persistent manner[1].

Example:
A digital plan of a house and its printed version on paper are two different information carriers sharing the same information object.

From the example, it is clear that one and the same information object can be showed on different carriers, analogue physical objects like sheets of papers, or digital physical objects like computer files, just to mention few examples. This heterogeneity plays a relevant role in the context of our project, since depending on the type of object at stake different technologies need to be adopted for their analysis.

2.2 Heterogeneity in information object

CIDOC-CRM (v6.2.3) provides a taxonomy of information objects, among which *LinguisticObject* (E33) and *Image* (E38) are particularly relevant for our purposes.

Linguistic objects are information objects specified in natural languages, e.g., German or French. From the CIDOC documentation, it seems that the identity of linguistic objects is bounded to the languages in which they are specified. For instance, the English version of Lewis Carroll's novel "Alice's Adventures in

[1] The last version of CIDOC-CRM (v 6.2.3) deprecates the use of the class *InformationCarrier* and recommends using *Man-MadeObject* instead. However, we find useful the use of the former class to specifically refer to objects carrying information objects. Also, note that the definition of a *Man-MadeObject* carrying an information object would generate an information carrier anyway.

Wonderland" is a different linguistic object when compared with the French or German versions. Also, CIDOC specifies that "formal languages such as computer code or mathematical formulae are not treated as instances of E33 *LinguisticObject*. These should be modelled as instances of E73 *InformationObject*".

Images are information objects which are meant to be specified by using forms, colors, etc (note that, as information objects, images do not have to be confused with their carriers, e.g., photos, paintings, or prints). Differently from *LinguisticObject*, however, CIDOC does not model explicitly the relationship between images and the non-verbal languages in which they are represented[2]. Also, differently from *LinguisticObject*, CIDOC does not take a position on the identity of images with respect to their representations. Indeed, according to the documentation, the original painting of Da Vinci's Mona Lisa and all its reproductions may be said to share the same image.

From our perspective, we find limited CIDOC's reference to *Language* (E56) in the sense of natural languages. As we saw for images, it seems reasonable to talk about non-verbal or graphical languages. For example, in the case of Computer Aided Design (CAD) systems, it is common to use graphical geometric elements such as surfaces or solids to create the desired shapes. This limitation affects also the manner in which information objects represented in formal languages are represented. As noted above, indeed, they are not treated as instances of *LinguisticObject*.

We propose in the following a more general approach to represent linguistic objects that is built on CIDOC while revising it. Figure 1 presents the proposed taxonomy. Note that the *LinguisticObject* class directly subsumed by (CIDOC) *InformationObject* is now understood as an information object specified in a language, where the latter may be a natural language, a formal language (including computer code) or a (non-verbal) visual language. By looking at the taxonomy, *LinguisticObjectNaturalLanguage* corresponds to CIDOC *LinguisticObject*. Note that the subclasses of *LinguisticObject* are neither disjoint, nor they form a complete specialization of the parent class. In this manner, one can define further classes mixing, e.g., formal and visual languages (see example below).

Example:
A digital 3D CAD model of a house. In this case, we have to distinguish between (at least) two linguistic objects (in the sense of Fig. 1). First, we have a linguistic object, call it lob_1, expressed in a formal, computer-based, language. Depending on the level of generality, one can identify either a series of bits, or a higher-level code. Second, we have a linguistic object, lob_2, that is expressed in a visual, geometric-like, language. Since lob_2 is carried on a digital support, the language in which it is encoded is a formal, computer-based, language. In this sense, lob_2 instantiates both *LinguisticObjectFormalLanguage* and *LinguisticObjectVisualLanguage*[3].

[2] Note that CIDOC does not use the expression non-verbal languages.

[3] It should be clear that lob_1 and lob_2 cannot be reduced to each other. In the former case, the language is purely verbal, in the second one the language is graphical.

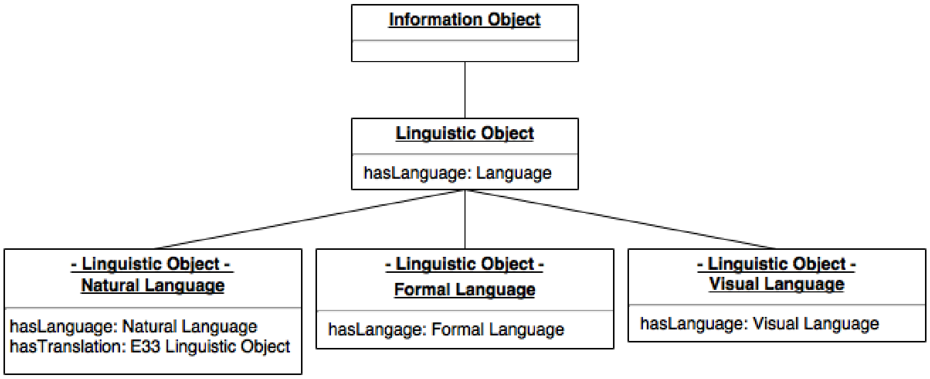
Information Object
Linguistic Object
hasLanguage: Language
- Linguistic Object - Natural Language
hasLanguage: Natural Language
hasTranslation: E33 Linguistic Object
- Linguistic Object - Formal Language
hasLangage: Formal Language
- Linguistic Object - Visual Language
hasLanguage: Visual Language

Fig. 1. Linguistic Object taxonomy

2.3 Heterogeneity coming from data alteration processes

The last heterogeneity type that we identify comes from alteration of data. Consider the following example.

Example:
The scanning of a paper photo results in a digital picture that is different from the original, analogue, picture, e.g., because it was poorly positioned, or the scanning machine stopped working in the course of the process.

In this case, by looking back at our taxonomy, we clearly have to distinguish between two linguistic objects carried in different carriers, paper and computer file. However, since we assume that the digitalisation of the paper photo contains some noise with respect to the original one, the challenging question is: Do the paper photo and the digital one carry the same information object? Otherwise said, how much of the original photo is preserved in its faulty digitalisation? In the context of cultural heritage this is a very challenging question, which experts need to face. From our point of view, the approach we propose allows to track alterations, while it is left to domain experts to agree on the identity and persistence conditions of the entities they manipulate. Other works have proposed alternative approaches, such as [8][9]

Clearly, in the case of a CAD system, there is a way to establish a correspondence between the verbal code and its graphical visualisation.

3 Tracking approach

Tracking the sources and dealing with all the heterogeneity has led us to develop a specific methodology. We distinguish the data collection from the data generation based upon rules and previously gathered data. We use Semantic Web technologies and approaches, e.g., rely on triple stores databases to store and manage data.

During a data collection process, we treat sequentially every information object and create a specific description file. At this step, the triple store is the alteration of the selected information object source content through the filter of our ontology. At the level of the triple, this implies tracking the information object coming along with an information carrier. Another need is to track the relevant segment/part in the Information Object the triple comes from. This second tracking requires to have the Information Object segmented and to have every segment positioned one to another. Furthermore, out of these segments, radical elements are identified that will be then interpreted, and possibly aligned to an ontology. This last step enables the translation from a segment of an Information Object into single triplets. This is very similar to any ETL process (as in [ref]) or other data collection or transfer from heterogeneous sources.

In order to be more complete, we also needed to track meta-data. They are of two types in our context: meta-data of the Information Object and meta-data of the collection process. To sum up, any collected triple in our triple store must be linked to an Information Object, and if possible a precise segment of it, to its Information Object metadata, and to the collection process metadata.

In order to be consistent, we developed a specific file structuration. Every RDF file describes a collection process over a specific Information Object in five different sections:

- The collection meta-data (containing the destination ontology)
- The Information Object meta-data
- The radical elements
- The collected triples
- The association between the triples and the Information Objects segments

In the case of a data generation process (by combination of data and possibly rules), the methodology follows the same principle: we need to be able to track the origin of every generated triple, “to go up the river”. In order to keep track of our work, we store generated data in separated RDF-files and every triple is linked to the set of collected triples and rules whose combination generated it.

A data-generation RDF-file is structured into 5 sections:

- The generation metadata (the set of generative rules)
- The collection stores metadata
- The radical elements
- The generated data
- The association between collected triples and rules, and the generated triples.

4 The ontology

In order to build our RDF files following th approach previously presented, we need to develop a specific vocabulary. The collection and generation metadata is a vast topic in itself and would bring more confusion than understanding if briefly addressed. It will not be covered in this paper. The four remaining sections are hereby more detailed.

4.1 Radical elements

Radical elements are all the objects at play in the content of the described source. They are all the objects and subjects of our triples. In this section of the file, they are associated with authority IRI. This section aims for disambiguation and precise identification of radical elements.

Two relations are present in this section:

- Has for litteral | *hasLiteral*
    - Domain: the radical element identifier (internal IRI)
    - Range: its literal expression (a string of characters)

  This property is close the SKOS property *prefLabel*, only aims to explicit literal labelling of the relevant radical element identifier.
- Refers To | *refersTo*
    - Domain: the radical element identifier (internal IRI)
    - Range: a relevant authority object (external IRI)

  This property is semantically near of the SIO, the CIDOC, and a lot of other *refersTo* properties. Only, we strictly use it with an authority resource for disambiguation and stronger cross linking.

Example:

```
:PicDuMidi a cidoc:place ;
              voc:hasLiteral "Pic du Midi de Bigorre";
              voc:refersTo dbr:Pic_du_Midi_de_Bigorre .
```

If needed, according to the corresponding ontology, this section can also include the typology of the radical elements.

4.2 Generated data

Generated data Each extracted data triple is designed in a named subgraph. Inside of which, the data reflects the analysis/generation ontology. No specific relation needed, the named subgraph identifier is associated, de facto with the contained data.

4.3 Provenance of generated data

Thanks to the previous named-subgraphs, we can directly associate our data triples with their respective source sub-segments. Only relation is therefore needed:

- Has for origin | *hasOrigin*
  - Domain: a named subgraph identifier (internal IRI)
  - Range: a sub-segment of a Information Object (external IRI)

  Close to other hasOrigin properties (like the dbpedia-owl one or the bevon origin), only it associates the IRI of one of the data collected triple with the external resource location of the source sub-segment.

Example:

```
:Description1 = {:BenjaminBaillaud cidoc:isIdentifiedBy
                              :directorOfToulAstroObserv} .
:Description1 voc:hasOrigin src:Segment01 .
```

4.4 Metadata of the Information Object

This section is composed of all possible metadata for the Information Object: e.g. the list of all known information carriers for it, its segmentation and the relative positioning of all of its segments, etc. Because of the specific structuration of digital documents we are working with, we needed to address the inclusion of information of structuration. This is not, in itself a direct tracking indicator of the alteration process of an information object. It can nevertheless be an indirect tracking tool: from the file structure it is possible to make assumptions on the typology of the alteration process. For example, from a

- File Contains | *contains*
  - Domain: Digital file IRI
  - Range: Information Object IRI

  This associates a digital file to an information object.
- Has structure of / Is structured after | *hasStructureOf*
  - Domain: Digital file IRI (external or internal)
  - Range: Structure type (external IRI)

  Associates a source document with a structure type. A structure type is the result of the use of specific syntax within a file format. It is especially useful to ease data comparison and data aggregation. *hasStructureOf* could be compared to the cube *structure* property, only it allows for a less strict range. The structure type would need a more complete definition to make this property externally usable. For now, we only defined structures types fitted to the project needs and have no general definition. This will need further work to correctly explicit the range and have authority data on the already most comonly used structures.
- Shares structure with / Has same structure than | *shareStructureWith*

  - Domain: Digital file IRI (external or internal)
  - Range: Information Object IRI (external or internal)

  Associates two source document through their structuring without specifying it. We couldn't find an equivalent to this. It is an impoverished derivative of hasStructureOf, only to be used to keep file association if the structure type has not been formalized.

- Is a direct measurement of type | *directMeasurementOfType*
  - Domain: Digital file IRI
  - Range: Measurement Type (External IRI)

  Relevant for Information object produced by a standardised measurement process, this property enables the specification of the type of measurement.
- Has measurement conditions | *measurementConditions*
  - Domain: Digital file
  - Range: Measurement Metadatafile (URL)

  This property allows the association in our model between a metadata file and the respective measurement file.
- Derives From | *derivesFrom*
  - Domain: Digital file IRI
  - Range: Digital file IRI

  This enables the tracking of various steps in a digital processing pipeline. This property can be seen as a purl version of derivesFrom for digital files.

Example:

```
:3DScanRaw voc:contains :3DScanCoupoleBaillaud .
:3DScanRaw voc:hasStructure struct:Faro3DPointcloud .
:3DScanClean voc:shareStructureWith :3DScanRaw .
:3DScanRaw voc:directMeasurementOfType meas:lasergrammetry .
:3DScanRaw voc:measurementConditions
http:reseed.ls2n.fr/scan_data/20180101_3DScanRaw_Metadata.xml .
:3DScanClean voc:derivesFrom :3DScanRaw .
```

In case of recusion

In a case of data collection, the content of the Information Object may refer to

- another Information Object (e.g. An excerpt from a text stating that one of the protagonist reads Les trois mousquetaires of Alexandre Dumas).
- a segment of another Information Object (e.g. A digital picture or a text containing a quote from Les trois mousquetaire)

Depending on the collection ontology, if this segment is relevant information to the collect, the database would be richer if it linked collected data and the origin Information Object. Tracking through Information is simple: the origin Information is considered a radical object in our file. At the content level of the origin Information Object, it is case dependant. The triple of data collected associates the segment of the origin Information Object as radical element, with another radical element in accordance with the extraction ontology. Then in the radical element section, the segment is linked to its original Information Object. 1). Either at segment level if the original source has been properly described by an analogous methodology. Then in the radical elements description section, it is only a matter of associating the original segment identifier with the collected one by using refersTo and hasOrigin as previously described. 2). Or at source level, if the original Information Object has not been described, in the radical element description section, by associating the segment to the origin Information Object. This association is done with the help of:

- Comes from a section of / *comesFromSectionOf*
  - Domain: a radical element identifier (internal IRI)
  - Range: an Information Object IRI

  We have found no equivalent property to compare to this one. It creates an association between an excerpt of the source document with another source, more or less literally related. This association has a broad acceptance, from the slight reference to the direct citation.

5 Perspective and limitations

This approach enables us to track the provenance of the data with regard to our needs. For us, it seems that, the validity of the data stored can be criticized at the level of the Information Object and of its segments, while collecting of generating data. It is a first step towards the implementation of data integrity indicators. It also enables a monitoring of biases/efficiency between data extraction/generation processes.

This approach has been developed with the aim of building a more global data integrity assessment process when building a complex digital model of an object. The needed various alterations, inherent in the course of working with digital data can lead to undocumented data loss or data creation. Since we do not think it is possible or preferable to avoid them, we are needing a tool to track and quantify the data alteration, in order to allow the critic of the final model.

The work is complicated by this approach, but when working with cultural heritage objects, such integrity assessment is a common practice that we try to formalise and reproduce when working in digital environment, especially with 2D/3D data.

References

1. ReSeed Project. http://www.reseed.fr
2. Hervy, B., Laroche, F., Bernard, A., Kerouanton,J.L.: Framework for historical knowledge management in museology. In: International Journal of Product Lifecycle Management, Inderscience, 2017, 10 (1), pp.44-68. 10.1504/IJPLM.2017.10003813.

3. Laroche, F. Bernard, A., Hervy, B.: DHRM, A new model for PLM dedicated to product design heritage. In: CIRP Annals - Manufacturing Technology, Elsevier, 2015, CIRP Annals, 64, 4p. 10.1016/j.cirp.2015.04.027.
4. Ouamer Ali, M., Laroche, F., Bernard, A., Rémy S.: Toward a methodological knowledge based approach for partial automation of reverse engineering. CirpDesign April 14th 2014, Milano, Italy
5. German Digital Library https://ercim-news.ercim.eu/en86/special/preparing-the-ground-for-the-german-digital-library
6. Carlisle, P. K., Avramides, I., Dalgity, A., Myers., D. The Arches Heritage Inventory and Management System: A Standards-Based Approach to the Management of Cultural Heritage Information. Paper presented at the CIDOC (International Committee for Documentation of the International Council of Museums) Conference: Access and Understanding – Networking in the Digital Era, Dresden, Germany, 6-11 September 2014.
7. Claros Project http://explore.clarosnet.org/XDB/ASP/clarosHome/about.html
8. Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., Melessanakis, V.. Modeling and Querying Provenance using CIDOC CRM. 2018.
9. Frank, M., Zander, S. Pushing the CIDOC-Conceptual Reference Model towards LOD by Open Annotations. 2016.