Creation of a domain ontology in CIDOC CRM OWL format using heterogeneous textual data related to industrial heritage

Eric Kergosien, Kaouther Ben Smida, Rémi Cardon, Natalia Grabar, Mathilde Wybo

Creation of a domain ontology in CIDOC CRM OWL format using heterogeneous textual data related to industrial heritage

Eric Kergosien[1], Kaouther Ben Smida[1], Rémi Cardon[2], Natalia Grabar[2], Mathilde Wybo[3]

1. GERiiCO EA 4073, University of Lille, France, prenom.nom@univ-lille.fr
2. STL, University of Lille, France, prenom.nom@univ-lille.fr
3. IRHIS, University of Lille, France, prenom.nom@univ-lille.fr

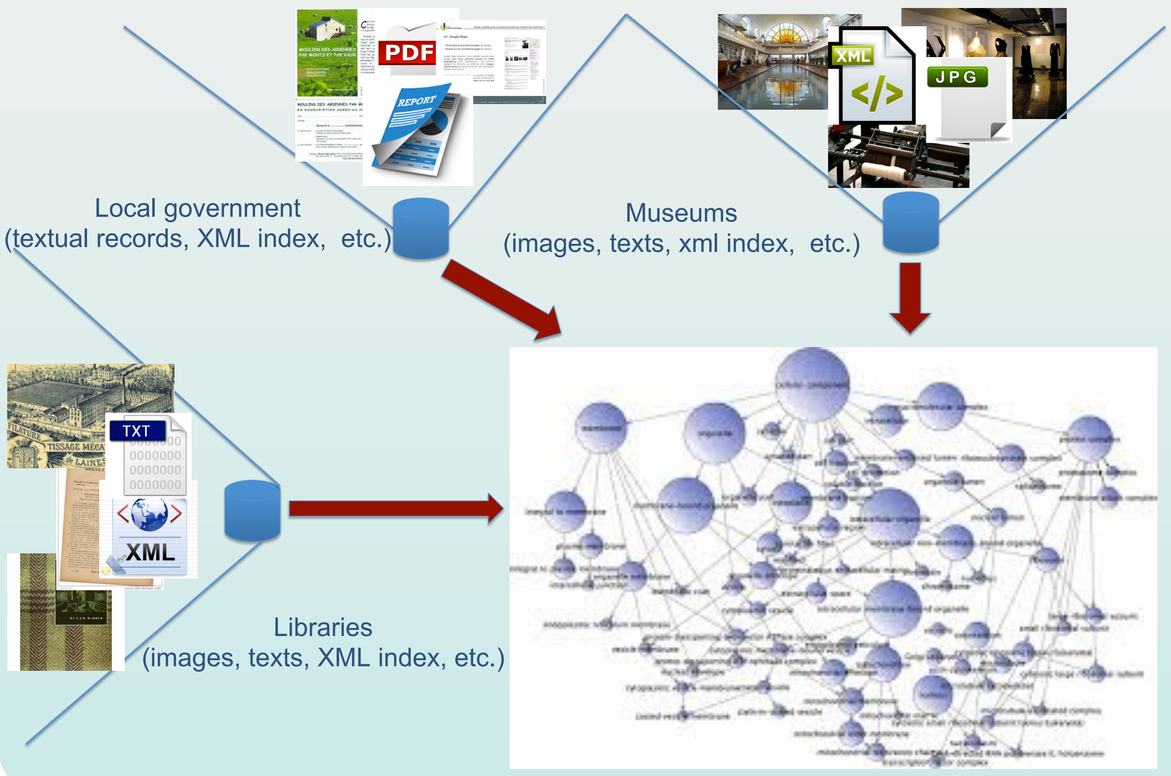
Keywords:, Domain ontology construction, industrial heritage, CIDOC CROM, Text Mining, Document Analysis.

Abstract: The TERRE-ISTEX project aims to provide a knowledge representation that interconnects all of these data, thanks to the semantic web technologies, in order to assist domain experts in producing and providing digital content. The originality of the project is to adopt a multidisciplinary approach to provide stakeholders, experts and non-experts, help them in the discovery of knowledge specific to their heritage, thanks to the extraction, structuring and visualization of knowledge from heterogeneous digital corpora. According to UNESCO, which has contributed significantly to the definition of the heritage (UNESCO, 1954, 1970, 1982), and then to The International Committee for the Conservation of Industrial Heritage (TICCIH, 2003), the industrial heritage can be defined as:

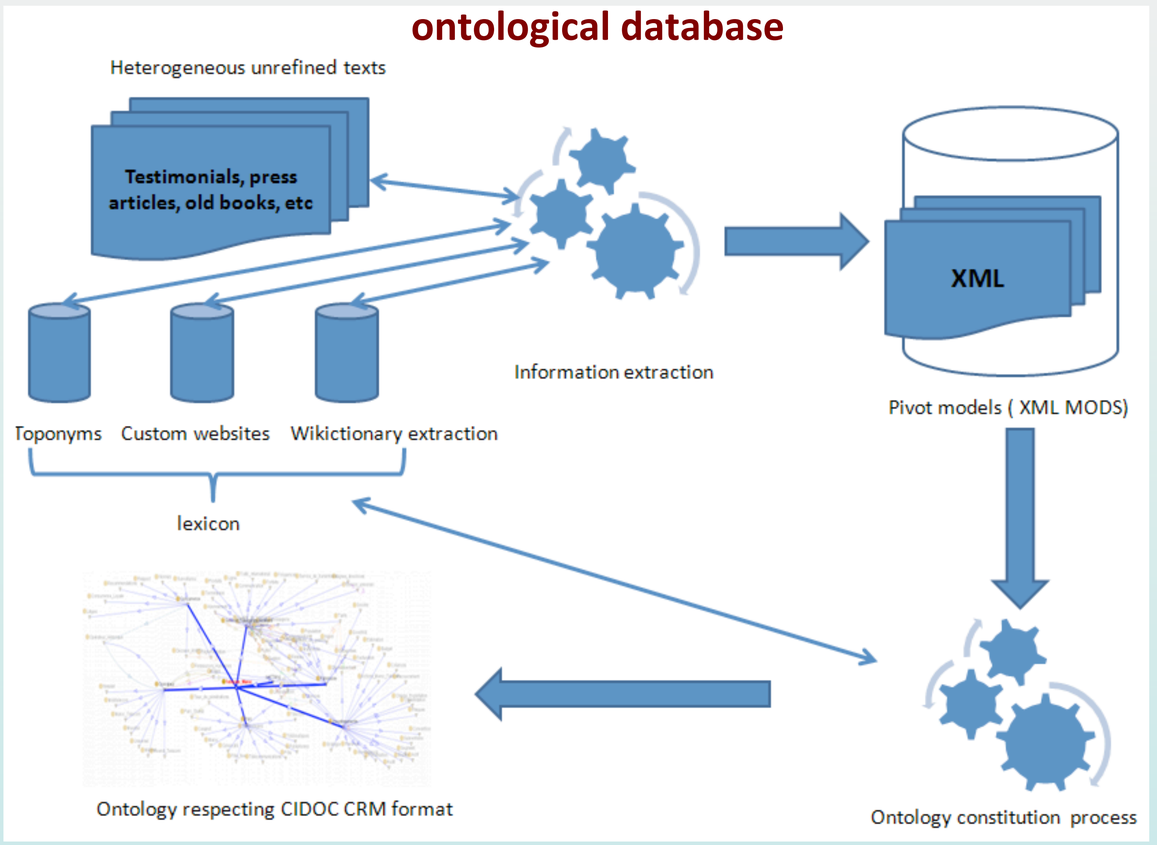
- Material assets: buildings, machinery, equipment, workshops, factories, processing and refining sites, shops, production centers and social activities related to the textile industry;
- Immaterial assets: memories, events, festivals, collective images, intellectual production transmitted by know-how which can be a succession of gestures dictated and displayed in production centers.

In our work, the main efforts are focused on modeling of the domain stakeholders, the spatial entitiesand thematic, which belong to both of the assets.

Main goal: to provide a knowledge representation based on heterogeneous data related to the industrial heritage



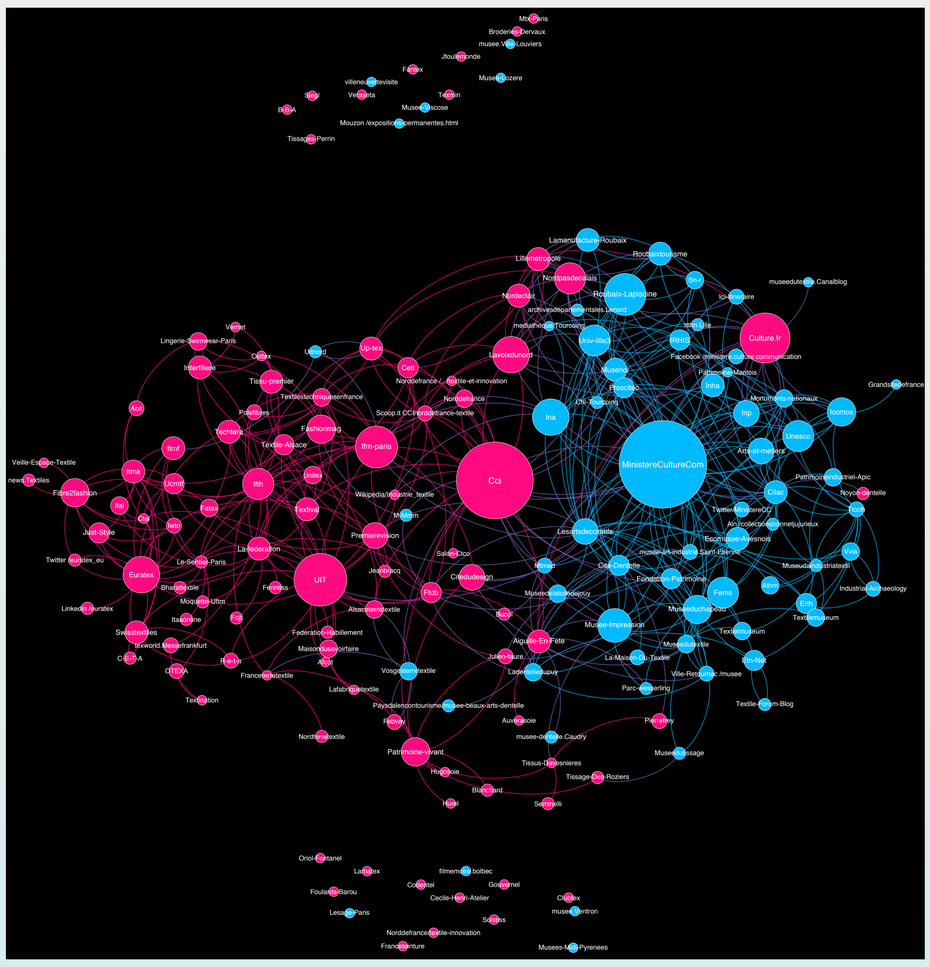
Method: Information extraction method for creation of the ontological database



Experiments

A three step methodology for semi-automatic building of semantic representation of the studied domain from thousands heterogeneous documents

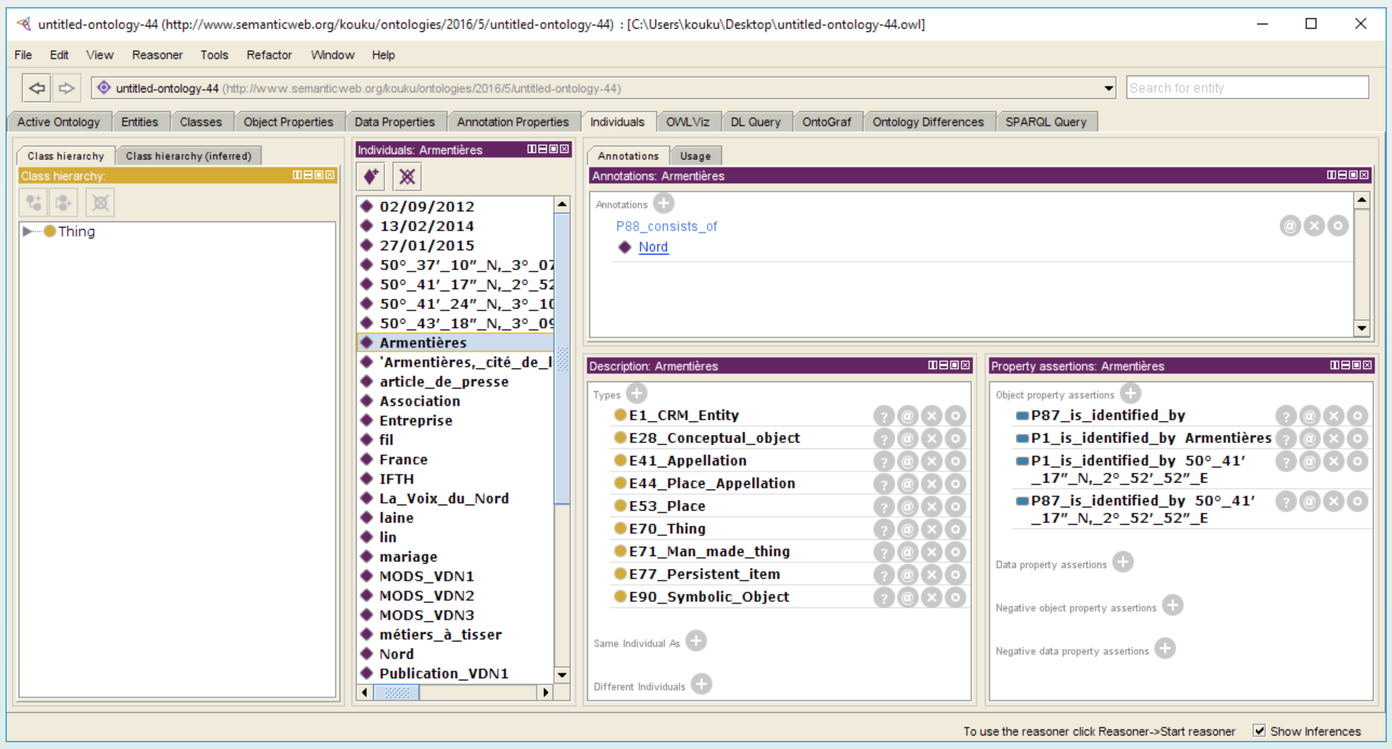
1. We collect and formalize the history through interviews with stakeholders. In addition to the collected information, we also exploit the Gephi tool to analyse stakeholders relations



2. identification and extraction of information related to industrial cultural heritage from heterogeneous textual documents : → Combining lexicon projection with text mining methods to improve the identification of relevant data.
   - Lexicon of spatial Entities (regional municipalities)
   - Lexicon of the domain's stakeholders (step1)
   - Thematic lexicon: combines (1) several existing specialized resources (Joconde created by French museums, Rameau created by the National Library of France, Wiktionnary) and a Text mining approach based on the Word2vec algorithm in order to identify of new terms from the processed corpus

3. Automatic ontology construction using the OWL CIDOC CRM format to merge together all our lexica. In this phase, it is important to filter the CIDOC CRM model to obtain a sub-model with the relevant concepts and properties

Ontology instantiation



Extract of the domain ontology based on four heterogeneous documents using the Protege Software (Musen et al., 1995)

Example:

- Xml document *IRHIS_FL1269145.xml* speaks about the French President involvement to the textile international exhibition in 1911
- Pdf document MEL_Roubaix_AVA .pdf states that the textile international exhibition took place in the Parc Barbieux in Roubaix, city in Northern France