SILKNOW – Multilingual Text Analysis for Silk Heritage

Dunja Mladenić
Artificial Intelligence Laboratory
J. Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Mar Gaitán
History of Art Department
Universitat de Valencia
Valencia, Spain
Identidaduv.Tres@uv.es

Raphäel Troncy
Data Science Department
EURECOM
Biot, France
raphael.troncy@eurecom.fr

ABSTRACT

We present results of collaborative work bringing together semantic technologies, machine learning and cultural heritage to enable advanced search and visualization of textual descriptions of museum artifacts related to silk fabrics. Proposed is a multilingual txt analysis approach where the developed domain-specific multilingual thesaurus and domain-specific ontology are utilized in data representation and analysis. In addition, a general multilingual semantic annotation tool Wikifier is applied on thesaurus definitions and descriptions of silk-related museum artefacts. The validation on real-world data of several museums confirms suitability of the developed thesaurus and the ontology.

CCS CONCEPTS

• Information systems, Information retrieval, Specialized information retrieval, Structure and multilingual text search • Computing methodologies, Machine learning, Machine learning approaches • Computing methodologies, Artificial intelligence, Knowledge representation and reasoning, Ontology engineering

KEYWORDS

Multilingual text analysis, controlled vocabulary, text annotation, knowledge resources, silk heritage

ACM Reference format:

Dunja Mladenic, FirstName Surname, FirstName Surname and FirstName Surname. 2018. SILKNOW- Multilingual Text Analysis for Silk heritage. In *Proc. of ACM KDD Workshops (KDD'19). ACM, NY, USA, 2 pages.*

1 Introduction

Silk heritage reveals not only the development of silk production itself but also the influence of silk trades along the Silk Road's on exchange of idea and innovations in the society. Although many museums are looking into preservation of silk heritage, interconnection of their collections is limited.

The presented work is in direction of using multilingual text analysis to support development of an intelligent system that will improve understanding of silk heritage. This will be achieved through incorporating semantic data enrichment, rich data visualization and analysis. This work is a part of a bigger collaboration efforts on EU project "SILKNOW - Silk heritage in the Knowledge Society: from punched cards to big data, deep learning and visual / tangible simulations". Partners collaborating on the project cover several complementary areas including cultural heritage, silk production, big data text and image analytics, graphics simulation and data visualization, 3D printing on textile, data interoperability and integration.

Figure 1: The consortium is composed of nine partners as shown in Fig. 1: four Universities, two research institutions, an international organization and two small companies.

2 Multilingual Text Analysis Approach

The proposed multilingual text analysis approach combines multilingual thesaurus, ontology and multilingual semantic annotation of text. The aim is to support prediction of some metadata values, information extraction from textual descriptions provided in the online resources of museum, rich visualization and semantic search.

The SILKNOW multilingual thesaurus [1] was developed by silk heritage domain experts in collaboration with semantic technologies and data analytics experts. Several specialized dictionaries and books were used to define and select specific terms related to historic silk. In addition, several general dictionaries and thesauri were used including Getty AAT [2]. The thesaurus was initially developed in Spanish containing about 500 terms, of which the most relevant have been translated to English, Italian and French. It is worth pointing out that the thesaurus is growing in the sense of new terms being added, additional terms being translated from Spanish and adding more details to the term descriptions. Fig. 2 shows screenshot of a thesaurus definition.

Each term in the thesaurus is described by its unique ID that is the same across all the languages, the term itself as a string, term definition, bibliographic sources, synonyms and associated terms. During the development the thesaurus was validated using textual data in online resources collected by SILKNOW crawler. Validation using two Spanish museums (CERES-MCU and IMATEX), we found that about 60% of the thesaurus terms are present in the museums with the most commonly used terms being *seda, cárcola, calada, jacquard, trama*. We found some terms in the museums that are not included in the thesaurus, these where checked by the domain experts for possible thesaurus extension.

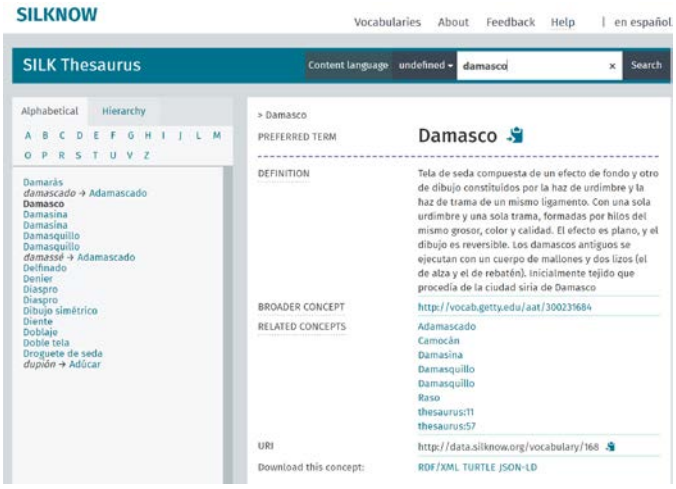

Figure 2: Thesaurus screenshot showing *damask* definition.

The SILKNOW ontology [3] is based on CIDOC-CRM that is commonly used for describing concepts and relations in cultural heritage. The mapping rules were defined that enable validating coverage of the ontology based on the metadata records from on-line resources collected by SILKNOW crawler. Validation using three museums (GARIN, Joconde and IMATEX) has shown that all the current fields can be represented using classes and properties from the SILKNOW ontology. As future work, we will develop the mapping for additional museums: Museos estatales del MEC, Musée des Arts Décoratifs, Metropolitan Museum, Boston Museum of Fine Arts, Musée des Tissus de Lyon, Rhode Island School of Design, Victoria and Albert Museum.

The multilingual annotation of text is performed using publicly available Wikifier service [4] that enables annotation of document with concepts from Wikipedia. This enables us to obtain semantic representation of textual fields available in online resources of the museums, as well as the textual definition of terms provided in SILKNOW thesaurus. To understand better the content of different textual fields in the museums, we have calculated representative words and phrases for each field. For instance, in IMATEX descriptions of decoration the most frequent phrases are *floral pattern*, *plant motifs*, *stylized plant motifs*.

Fig. 3 provides an illustrative example of the annotated Italian definition of term "damask" from SILKNOW thesaurus. We can notice a number of terms that are domain-specific and were not annotated. This naturally suggest to extend the existing Wikifier to include terms from SILKNOW thesaurus, which is ongoing work. As future work, these semantically enriched texts will be used in data analysis and rich visualization.

3 Discussion

The presented work brings KDD technologies to the domain of cultural heritage to enable multilingual and semantically enriched access to existing digital data on silk heritage. The developed multilingual thesaurus http://skosmos.silknow.org/thesaurus/en/ and the ontology http://data.silknow.org/ are both publicly available. Presented work is a part of a larger collaborative efforts within SILKNOW research project http://silknow.eu/

Figure 3: Wikifier annotation of *damask* definition in Italian (translation by Georgia Lo Cicero)**.**

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and SILKNOW European Union Horizon 2020 project under grant agreement No 769504.

REFERENCES

[1] Mar Gaitán, Dunja Mladenic, Raphaël Troncy, 2018, D3.1 Historical silk multilingual thesaurus, Technical Report SILKNOW project.
[2] The Getty Research Institute. Art & Architecture Thesaurus Online, 2017, http://vocab.getty.edu/aat/
[3] Raphaël Troncy, D3.2: Design of the SILKNOW Ontology and the Ontology, Technical Report SILKNOW project.
[4] Janez Brank, Gregor Leban, Marko Grobelnik, 2017, Annotating documents with relevant Wikipedia concepts. Proc. of the SiKDD-2017, pp 9-12.