The SILKNOW Knowledge Graph

Thomas Schleider [a], Raphaël Troncy [a], Mar Gaitán [b], Ester Alba [b], Jorge Sebastián [b], Dunja Mladenić [c], Avguštin Kastelic [c], M. Besher Massri [c], Arabella León [d], Marie Puren [e], Pierre Vernus [f], Dominic Clermont [g], Franz Rottensteiner [g], Maurizio Vitella [h] and Georgia Lo Cicero [h]

[a] *Data Science Department, EURECOM, Campus Sophiatech, France*
E-mails: thomas.schleider@eurecom.fr, raphael.troncy@eurecom.fr
[b] *Departamento de Historia del Arte, Universitat de València, Av. de Blasco Ibáñez, 28, 46010 València, Spain*
E-mails: silknow@uv.es, esther.alba@uv.es, jorge.sebastian@uv.es
[c] *Artificial Intelligence Laboratory, Jozef Stefan Institute, Jamova 30, 1000 Ljubljana, Slovenia*
E-mails: dunja.mladenic@ijs.si, avgustin.kastelic@ijs.si, besher.massri@ijs.si
[d] *Garin 1820 S.A, Calle Ramón de Villarroya 1, Moncada, 46113, Spain*
E-mail: aleon@garin1820.com
[e] *CNRS - LARHRA UMR5190, Lyon, France*
E-mail: marie.puren@msh-lse.fr
[f] *Université Lumière Lyon 2 - LARHRA UMR 5190, Lyon, France*
E-mail: pierre.vernus@msh-lse.fr
[g] *Institute of Photogrammetry and Remote Sensing, Leibniz Universität Hannover, Jamova 30, Nienburger Straße 1, 30167 Hannover, Germany*
E-mails: clermont@ipi.uni-hannover.de, rottensteiner@ipi.uni-hannover.de
[h] *Dipartimento Cultura e Società, Università degli Studi di Palermo, Italy*
E-mails: maurizio.vitella@unipa.it, georgia.locicero@unipa.it

Abstract. SILKNOW is a research project that aims at improving the understanding, conservation and dissemination of the European silk heritage from the 15th to the 19th century. This paper presents the SILKNOW knowledge graph (KG) that lies at the center of the application of Semantic Web technologies and computing research to the needs of museums and every other user of this knowledge. The underlying data model is based on CIDOC-CRM and data mappings which are realised and implemented with conversion tools developed for SILKNOW. The full integration pipeline consists also of our own crawling software to retrieve the original data from both public sources and project partners. We developed an API access for the KG and created the exploratory search engine ADASilk on top of it. Finally, we present how we apply automatic image and text analysis to predict missing metadata in the knowledge graph.

Keywords: multilingual thesaurus, cultural heritage, silk heritage

1. Introduction

Inventory and cataloguing, including texts and images, are indispensable requirements for the identification and conservation of cultural heritage artifacts. In the last years, many museums and libraries made great efforts to make their collections available in open access datasets. In this respect, controlled vocabularies allow to obtain better information, especially if they exist in several languages, thus, enabling the integration of information. On the other hand, information and communication technologies have been gradually been incorporated in museums. Two digital projects have showcased the holdings from textile museums and collections. Interestingly, they have been produced by the two major contenders in the online arena for cultural content: The Google Cultural Institute (with its *We Wear Culture* resource) and Europeana (through the *Europeana Fashion* portal). This attests the interest shown by digital content aggregators towards collections of textiles, dresses and fashion.

The keepers of these types of collections, however, are sometimes in a difficult position when it comes to taking advantage of new digital tools. Innovations such as 3D printing of textiles and automated image matching -to name but a few- have the potential to revolutionize the ways in which this cultural heritage is explained and made accessible. In some cases, small and medium sized textile museums are struggling for their very survival, and bold technological ventures are understandably not seen as a priority. Otherwise, digital cataloguing is very common today, but the differences among institutions are bewildering: from entire collections served online through updated web and API outlets, to custom-made Excel records stored on a local hard drive. In this context, interoperability between independent catalogs becomes very difficult, even though the necessary technologies and standards are well known for the museum community.

Facing these challenges and opportunities lies at the core of the research presented in this paper.

2. Related Work

The development of the web has led Cultural Heritage (CH) organisations to provide information on their objects to many portals or aggregators. But the fact that each collection management system or cataloguing database has potentially its own metadata format makes information integration costly and time consuming. To make data integration easier and less costly, different organisations elaborated guidelines or guides for best practices.

Thus, in 1995 the "Documentation standard" Working Group of the International Council of Museums (ICOM) published its International Guidelines for Museum Object Information[1] which describes the "Information Categories that can be used when developing records about the objects in museum collections". Moreover, standard XML schemas have been provided in order to enable institutions to use the Open Archives Initiatives Protocol for Metadata Harvesting (OAI/PMH). For instance the CDWA lite, developed in the United States by the Getty foundation, museumdat, largely built upon the former developed by the Deutscher Museumsbunde (DMB) or Spectrum, developed in the UK.

Presently, the Lightweight Information Describing Objects (LIDO)[2] schema published by the Working Group "Data Harvesting and Interchange" of the ICOM has superseded both the CDWA Lite v1.1 schema and the museumdat v1.0 schema. LIDO provides an explicit format to deliver museum's object information. It is an application of the CIDOC Conceptual Reference Model (CRM). This model, specifically developed for information integration in the field of cultural heritage, is the outcome of over 20 years of development originally by the ICOM's International Committee for Documentation (CIDOC) Documentation Standards Working Group and, presently, by the CIDOC-CRM SIG. Since December, 2006, the CIDOC-CRM is an official ISO standard, a status renewed in 2014.

Another example of technology for the archive, search and exploration is the Europeana platform, launched in 2008. It consists over 57 million objects from more than 3500 institutions in Europe [3], [4]. All Europeana datasets can by now be explored and queried through a SPARQL API. Its data is represented in the Europeana Data Model (EDM). [1]. It is defined as an "integration medium for collecting, connecting, enriching the description provided by Europeana's content providers". In fact, EDM defined a limited set of elements, some reused from other namespaces (RDF and RDFs, OAI ORE, SKOS, Dublin Core, DCAT and Creative Commons and SIOC Services Ontology Module) and some other introduced by Europeana. But even this subset is also partly based on other models: for instance, among the 11 classes introduced by Europeana, 6 are noted as equivalent to CIDOC-CRM classes [2].

As a last example, in France, the Ministry of Culture and Communication has initiated the development of a Harmonized Model for the production of cultural data. Although not an exchange model, this model is semantically compatible with the CIDOC-CRM and its extensions, the Europeana Data Model (EDM) and the LIDO schema [3].

Given these efforts to harmonize data produced by cultural heritage organizations, it is not surprising that they are resorting to Semantic Web technologies and

[1] https://icom.museum/en/ressource/international-guidelines-for-museum-object-information-the-cidoc-information-categories/

[2] https://en.wikipedia.org/wiki/LIDO

[3] https://pro.europeana.eu/page/reasons-to-share-your-data-on-europeana-collections

[4] https://www.europeana.eu

knowledge graphs. CultureSampo [5] is one example, where the main aim was to publish heterogeneous cultural content on the Semantic Web [4]. It had to deal specifically with challenges, such as converting legacy data into linked data and to make heterogeneous, but interlinked cultural heritage content interoperable on a semantic level. For example, the challenge of finding connections between two persons can be solved by using the Getty ULAN structured vocabulary of artist names and biographical information [5], [6].

3. SILKNOW Ontology

In this section we give examples of competency questions the SILKNOW knowledge graph is able or supposed to be able to answer. We will also describe the underlying data model, which consists of the combination of several existing ontologies and our own extensions, built on with OntoMe. Finally, we will give an overview of the controlled vocabularies to which we link the data by replacing string values with URIs (see section 4.2).

3.1. Competency Questions

In order to better solve what the SILKNOW knowledge graph should be able to answer, the domain experts established a set of competency questions that ideally, our different target audiences would like to ask. In order to do so, we followed the DOREMUS (https://www.doremus.org/) project. The domain experts wrote around 70 questions that can be grouped in questions related to material and techniques, location, time, artists or style, etc. It should be noted that even though these questions were made by cultural heritage experts, we took other stakeholders into account to cover a wide range of social interests related to the silk heritage as a whole (tangible, intangible and living heritage) [7] .

Questions related to the needs of museums, researchers, curators, etc. refer to the specificity of silk heritage description, such as: location, period (time), typologies, materials, artists, style. Additionally, to the multiple associations between these basic questions: time and location, artists and location, type of items, time, location and material. On the other hand, we can find other examples related to the probability of the expected results in both simple and advanced search processes, including the solving of complex temporal space questions: e.g., in which museums and collections around the world are Spanish textiles located?, which items have been produced in 1815?. Time-localization questions might be: which items were produced in France during the 18th century? Down to those more complex questions like: give me all available information on silver ribbons produced in Italy during the Renaissance.

However, SILKNOW is not only focused on cultural heritage experts [8], but its tools are meant to be applied by several communities and stakeholders [9]. The definition of different scenarios and target audiences was established this this goal in mind.

Scenario 1: Cultural Heritage. It is one of the essential user groups for SILKNOW. In fact, typical users work in a museum or are frequently related to this field; otherwise they are simply interested in historical silk products. Typical users might be: collectors, museum curators, museum conservators, museum visitors, staff from international organizations, museum directors.

Scenario 2: The research and educational sectors. This platform is a precious resource for educational purposes for people who wish to know about the silk history, taste and fashions that have influenced the creation of textiles, but mostly for those who need to search for technical information, and can benefit from a rich thesaurus: undergraduate students, graduate students, design students, fashion students, high school teachers, design professors, etc.

Scenario 3: Creative industries. Silk, as a material, has qualities very much appreciated throughout history. New technologies, such as 3D printing are naturally related to this precious material that continues to be a highly appreciated natural product, especially in the fashion industry. The typical involved users are: silk company CEOs, fashion designers, textile designers, photographers, 3D-printing company CEOs.

Scenario 4: Tourism. It is a scenario that increasingly involves the various geographical and social realities. Silk textiles are essential in the definition of important social identities (elites, cultural and religion symbols, among others), so the results must answer questions about their value as part of human cultural heritage. Notably, typical users could be: local guides, museum marketing professionals, regional associations, museum visitors.

Scenario 5: Media. We also wanted to involve media, specifically in the figure of the fashion journalist, as a user who directly draws on the information offered

[5]http://www.kulttuurisampo.fi

by SILKNOW to disclose it as part of a publication, a news piece, blog post, audio-visual media coverage, etc.

3.2. SILKNOW data model

Small parts of the total ontology for the SILKNOW Knowledge Graph are based on several properties of schema.org and the W3 time ontology. The majority of the classes and properties used in SILKNOW come from the current published version of CIDOC-CRM (6.2) and its extensions, the Scientific Observation Model (CRMsci) [10] and CRM Digital (CRMdig)[11]. The complete usage and implementation of these ontologies and data models can be retrieved from GitHub where it is part of the converter software. [6]

In order to aggregate numerous data sets collected from various sources, it is necessary to harmonize them by designing and implementing a unique and complete data model. To define the SILKNOW data model we first analyse the structure of records from several institutions especially the Victoria and Albert Museum, the British Museum, the Musée des Tissus in Lyon, the Garín collection at the Museu de la Seda in Moncada, the Musée des Arts Décoratifs in Paris, the Museum Baselland, and the French Joconde Database. We also used the ICOM guidelines for Museum Object, the Europeana data model, the norms and methods relative to the inventory keeping in French museums (*arrêté du 25 mai 2004*), and the French Harmonized Model for the production of cultural data. From this analysis we elaborated the data dictionary, i.e., a list of information groups or metadata interesting for the SILKNOW project. Then we selected in CIDOC-CRM the classes and properties useful to express these metadata.

We have chosen the CIDOC-CRM because it has been elaborated to express the underlying semantics of documentation on Cultural Heritage [12]. Moreover it is an international standard, recognized as an ISO norm. It has already been used in several research projects, included EU-funded projects, such as Ariadne which developed an extension of CIDOC-CRM suitable for archeological documentation [13]. SILKNOW is using version 6.2. It is an event centric data model, very flexible and extensible by nature: while it consists of a limited set of classes and properties, it is in fact a core ontology allowing the development of more specialised extensions. In other words, it is possible to add new sub-classes and sub-properties to express more specific relationships and properties, without modifying the basic structure of the model. The classes and properties selected for the SILKNOW ontology are publicly accessible and documented via OntoMe, an ontology management system, developed by the LARHRA research center [14].

[6] https://github.com/silknow/converter/tree/master/src/main/java/org/silknow/converter/ontologies

On the one hand, the bottom up approach adopted by SILKNOW spurred us to use CRMsci as a global schema for integrating metadata about scientific observations, performed by domain experts on silk-related artefacts:

Sources	Metadata examples	Mapping rules
Musée des Arts décoratifs	Laize à décor de bouquet noué par un ruban sur fond semé de quintefeuilles	**S4_Observation 08 observed** E22_Man-Made Object **S4_Observation** P3 has note E62_String
Victoria and Albert Museum	Furnishing fabric of brocaded silk satin, possibly designed by Philippe de Lasalle, France, ca. 1790	**S4_Observation 08 observed** E22_Man-Made Object **S4_Observation** P3 has note E62_String
Chiesa Madre di Caccamo	1 ordito, di fondo, organzino di seta, 2 capi, S, colore celeste	**S4_Observation 08 observed** E22_Man-Made Object **S4_Observation** P3 has note E62_String

Table 1

Mapping rules using classes and properties from CRMsci to modelize metadata examples about scientific observations

On the other hand, CRMdig was used to express the relationships between data sets and metadata records describing them:

D1_Digital Object P2 has type E55_Type (Data set)
D1_Digital Object P106 is composed of **D1_Digital Object** P2 has type E55_Type (Metadata record)

After evaluating the pertinence of the ontology by providing mapping rules between metadata examples and the SILKNOW ontology, it was observed that, so far, all fields can be represented by using existing classes and properties from the ontology.

Generally, scientific observations are expressed with free-text fields analysing the structure and the decoration of fabrics, and/or presenting the historical context of their production or use. This first mapping aimed at storing these metadata "as they are"; but the complex semantics included in data about the creative and productive process of silk textiles cannot accurately be represented with the basic CRM entities and its existing extensions. In order to address the complexity of textile data integration, it requires elaborating new CRM classes and properties.

There is yet no CRM extension for dealing with the production of textile artefacts, something similar to FRBRoo, for the creation, production and expression process in literature and the performing arts. A CRM extension is currently in development for this purpose, and a complete overview of these new classes and properties is publicly available via Ontome [15].

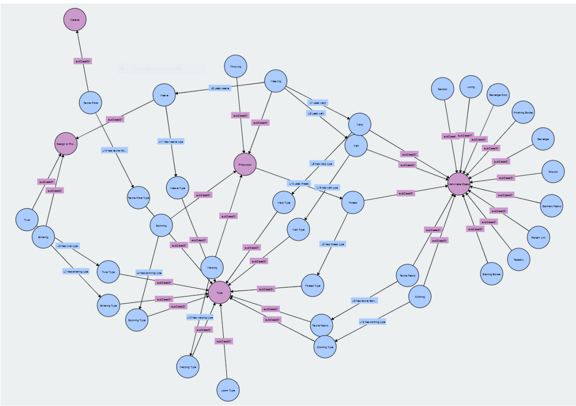

Fig. 1. CRM extension - RDF graph

Source	Metadata example	Mapping rules
Chiesa Madre di Caccamo	1 ordito, di fondo, organzino di seta, 2 capi, S, colore celeste	**T16_Warp L6 has warp type T30_Warp Type** (ordito di fondo) **T16_Warp** P57 has number of parts E60_Number (1) **T16_Warp** P45 consists of E57_Material (seta) **T15_Thread L3 has thread type T28_Thread Type** (organzino di seta) **T15_Thread** P56 bears feature E26_Physical Feature (color) P2 has type E55_Type (colore celeste) **T15_Thread** P56 bears feature E26_Physical Feature (twist) P2 has type E55_Type (S) **T15_Thread** P43 has dimension E54_Dimension E16_Measurement P40 observed dimension E54_Dimension E54_Dimension P91 has unit E58_Measurement unit (numero di capi) E54_Dimension P90 has value E60_Number (2) E54_Dimension P2 has type E55_Type (capi)

Table 2

A CRM extension to modelize the creation and production process of silk textile

3.3. SILKNOW Thesaurus

A museum can be understood as a huge data base where cultural objects are stored. In order to properly identify these objects, the documentation area emerges as a specific and important area in the museum. Documenting a cultural object means to register and to catalog it. Doing it properly is the precondition to ensure the physical persistence of objects as the registration of a cultural asset assumes its importance as cultural heritage that requires conservation and protection. Indeed, the basic element for conservation is to classify objects, understanding it as symbolic organization of meanings: "cultural artifacts constitute the network that sustains their institutions, they are symbols that are defined as the locally objectified sites of meaning." [16]. In other words, the conservation of cultural heritage begins with its registration and identification, tasks that are carried out through inventories and catalogs, which are the traditional tools for the study, analysis and especially protection of heritage [17].

As said before, in order to describe a cultural asset, proper terminology stands out as one fundamental pillar [18]. Information professionals, curators, conservators and general audience will be the end-users of these tools. Indeed, controlled vocabularies are essentials to provide access to museum collections not only to inside users (registrars, curatorial departments, conservators, education department), but also to external users who wish to know more about a subject without knowing the specific term of its search [19]. A thesaurus is defined in general, as a controlled vocabulary that has a semantic network of unique concepts [20] that enhances information retrieval, as it is based in queries based in categorized deductions [21]. It also links the object with the user as it allows to use a language that facilitates the research of a cultural asset and its related information. Moreover, the vast amount of metadata associated to it allows not only to document and describe the object, but also to find likenesses or differences between similar cultural assets, and to associate them, allowing users to find new connections [22]

Although some institutions and public administrations are striving to use standard vocabularies, most museums have generated their own methods of classification. The terminology used in the description varies widely according to different cataloging schools, fashions and curators in charge of this task. At the same time, museums around the world develop their own controlled vocabularies, that they see more fitting in order to describe their collections [23]. It is the case of The Textile Museum Thesaurus from the Textile Museum in Washington, or the Museon Arlaten. We can also mention the Domus system of Spain, hosted by the Documentary Standardization of Museums [24], or French databases such as Joconde and Gallica.

On the other hand, some standardization efforts have been carried out, such as the UNESCO thesaurus or the Getty Art & Architecture Thesaurus (ATT). Also, we can cite other generic thesauri, applicable to all types of cultural, movable or immovable property: CDWA, Object ID, ULAN, TGN, Iconclass, etc. Although they are useful for their own institutions, the result is a multitude of vocabularies that are not shared, complicating interoperability.

However, the cultural heritage domain and the silk heritage in particular are characterized by large, rich and heterogeneous datasets [19]. In this sense, the silk heritage vocabulary can change according to who (careers: weavers vs historians / disciplines: art historians vs. anthropologists) and where (Europe or North America) the term is being used [25]. This has resulted in the use of different terminologies in specialized organizations when describing their collections which makes comparisons among the same type of objects, techniques, designs quite complicated, not only in different languages but also in the same language.

On the other hand, cultural heritage data is being transformed into public Linked Data, especially in large-scale aggregators such as Europeana [23]. Plus, the Semantic Web technologies lead to a new approach in managing Cultural Heritage data interoperability [26]. Responding to these challenges, the SILKNOW thesaurus emerges as a thesaurus that aims to improve silk heritage knowledge by building an open-access thesaurus based on SKOS model. This thesaurus is multilingual and standardizes terminology providing conservators, researchers and other users an important tool, that allows systematic and coherent cataloging of museum collections, in order to avoid the lack of common criteria when dealing with these kinds of records.

3.3.1. Development method involving experts

Silk heritage experts were involved in order to develop the SILKNOW thesaurus. These experts included art historians, historians, weavers, engineers and philologists. Multidisciplinarity was essential in order to select terms, trace their evolution, historical and current use use, and how some terms evolved in time and space (e.g. local variations). As the SILKNOW thesaurus is symmetrical, all terms needed to be translated, textile specialists used specialized sources, which in some cases provided translations in other languages (such as the Castany Saladrigas dictionary, 1949). In other cases, direct translations were needed, a scope note was added when necessary or the source language was used as loan. Nevertheless, every translation was made following ISO directions for a thesaurus [27].

In order to compile the thesaurus, inductive and deductive methods were undertaken [28]. Around 80% of terms originated from inductive work; i.e., they were included in the thesaurus as soon as they were found in the literature. Specialized sources were used, such as specialized textile dictionaries, historical sources, glossaries, and other thesauri. The other 20% was deductive due to museum records and previous knowledge from the researchers. An extensive research was undertaken, not only taking into account specialized vocabularies, but also using historical sources and selecting the most representative and accurate ones.

Next, terms and concepts were controlled and described by adding scope notes, qualifiers and synonyms. A Preferred Term (PT) was used to refer a unique concept, whenever polysemy arose, qualifiers were added. In order to make clearer what those concepts meant, scope notes were added following specialized literature. Finally, these definitions were reviewed by international experts.

The next logical step was to categorize those terms. The SILKNOW thesaurus is based on the Getty AAT structure, as it is one of the most well-known thesauri in the cultural heritage field. Three relationships were established:

1) Hierarchical: when the relationship between terms is broader and narrower. Parents were also placed according the AAT structure when possible. As the silk heritage terminology is extensive and not easy to classify, compilers had to add new guide terms and subfacets in order to make it as accurate as possible.

2) Equivalence: This relationship concerns when different names refer to the same concept as they are synonyms or quasi-synonyms. E.g. bobillo → bocillo. Either noun is accepted to designate this type of lace, however bobillo acts as the Preferred Term.

3) Associative relationships: when different terms are conceptually closely related, but not hierarchically. E.g. acanalado → otoman. Both terms refer to a type of tabby, however they are not the exact same concept.

Finally, as the SILKNOW thesaurus was initially thought to standardize museums records, experts tried to make it as wide as possible in order to expand silk heritage knowledge. Looms, equipment, iconography, colours, botanical elements were added. This will help researchers to connect these data not only in museum's collections, but also in other research areas. In living heritage, for example, it is possible to see how some of these motifs are used in other contexts. By using this thesaurus, researchers, museum professionals, students and cultural heritage specialists will improve museum information and international research thanks to a free and easily accessible tool.

3.3.2. Thesaurus coverage

SILKNOW thesaurus was validated on textual data of the selected museums in several natural languages. The frequency of individual thesaurus concepts that are present in the specific museum was calculated. Spanish, English and French translations of the thesaurus were each compared to resources in the corresponding language. The program for the calculation of coverage was written in Python. Pre-processing was done using the Natural Language Toolkit library (NLTK) [29] which contains the Snowball Stemmer. It was used on all the terms and their synonyms from the thesaurus, as well as all the words from online resources.

Table 3 gives the results showing that 76% of the terms from the Spanish thesaurus are present in the Spanish museums, followed by 87% for the English thesaurus and 90% for the French thesaurus. In more detail, the two Spanish datasets CERES and IMATEX contain 361 and 326 terms from the Spanish thesaurus respectively, 308 of them occur in both museums. Both museums contain 379 terms from the Spanish SILKNOW thesaurus.

Spanish thesaurus	concepts	
CERES	361	72 %
IMATEX	326	65 %
Spanish museums	379	76 %
English thesaurus	**concepts**	
Victoria and Albert Museum	262	82 %
Rhode Island School of Design	210	66 %
Metropolitan Museum	205	64 %
IMATEX	182	57 %
English museums	279	87 %
French thesaurus	**concepts**	
Musée des Tissus de Lyon	255	89 %
Musée des Arts Décoratifs	201	70 %
Joconde	158	55 %
French museums	259	90 %

Table 3

Coverage of the thesaurus concepts in the museums. Showing results for thesaurus in each language separately over the museums for that language.

For each online resource (a dataset from a database or museum information system) a feature vector representing all its phrases was computed using QMiner platform [30]. The result was a set of n-grams with the maximum size of three words and a corresponding number of occurrences. From here a subset was generated where all the concepts that can be found in the thesaurus were removed from the feature vector.

Online resource phrases that are not included in SILKNOW thesaurus and the frequency of thesaurus concepts that are present in specific online resources were both visualized using word clouds. A Python library named wordcloud enabled to draw these in different patterns and colors resembling a flag of the origin country for each online resource. Word clouds are designed in a way where more relevant phrases in a represented resource are drawn in a bigger font than those that are less relevant.

The following Metropolitan Museum word cloud displays how frequently each of the thesaurus terms occur in textual data of the museum. In Fig. 2, we can see that the most frequent phrases are silk, spun silk, hard silk, course silk and bourette silk.

Metropolitan Museum word cloud displaying museum phrases that are not included in SILKNOW thesaurus. In Fig. 3, we can see that most of the phrases are ordinary words (eg., canvas, border, cotton, design, fragment).

Fig. 2. The most frequent English SILKNOW thesaurus phrases that occur in the Metropolitan Museum dataset.

Fig. 4. The most frequent English phrases that occur in the Victoria and Albert Museum dataset and are not included in SILKNOW thesaurus.

Similar observations can be made for the other museums that we have considered (see Fig. 4, showing the most frequent phrases in Victoria and Albert Museum that are not included in SILKNOW thesaurus). This in a way confirms that the SILKNOW thesaurus has a good coverage of domain-specific, silk related terminology.

Fig. 3. The most frequent phrases that occur in the Metropolitan Museum dataset that are not included in English SILKNOW thesaurus.

This word cloud displays Victoria and Albert Museum phrases that are not included in the SILKNOW thesaurus. Fig. 4 shows that the most common phrases are again colors, materials, geographical locations. We can also see that the most common motifs in silk related objects from this resource are flowers, which is evident from words like flowers, leaves and floral.

The CERES word cloud is displaying in Fig. 5 how frequently each of the thesaurus terms occur in textual data of this resource, that groups data from many Spanish national museums. Here most frequent phrases are seda, seda cruda, seda ocal, seda salvaje. Some of these words directly translate to most frequent thesaurus phrases from the Metropolitan Museum Fig. 2. For example seda translates to silk and seda cruda to hard silk. This shows us that a group of similar thesaurus concepts remain as most frequent phrases across many online resources in different natural languages.

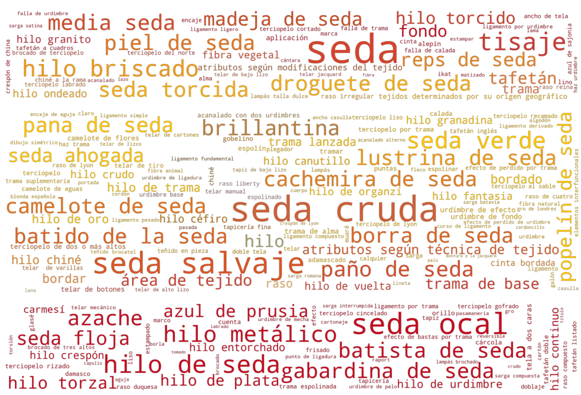

Fig. 5. The most frequent Spansih SILKNOW thesaurus phrases that also occur in CERES resources.

The Musée des Tissus de Lyon word cloud in Fig. 6 is displaying museum phrases that are not included in SILKNOW thesaurus. Similarly like with Victoria and Albert museum and Metropolitan Museum we find common ordinary words (eg., tissus, ligne, fleurs, maison) and phrases not related to silk terminology (eg., exposition, Paris, portrait).

The presented validation of the thesaurus has shown that it includes most of the silk related vocabulary that is used in the considered resources. The phrases which occur in the resources and are not included in the thesaurus are mostly common ordinary words or words not related to silk terminology.

4. Building the SILKNOW Knowledge Graph

The SILKNOW knowledge graph consists of publicly available datasets from museum collections that

Fig. 6. The most frequent French phrases that occur in Musée des Tissus de Lyon and are not included in the SILKNOW thesaurus.

contain silk items or from collections that are part of the project. As of now, it consists also of a crawler, a RDF converter and API transformer tools. All converted data is accessible online through a SPARQL endpoint based on Virtuoso [7]. This data is available through a faceted browser, too [8]. It also consists of a graphical user interface for the thesaurus [9], to which many data values in the graph are linked to, and to a general description of our ontology (see 3. Ontology). Finally, there is our exploratory search engine, ADASilk (Advanced Data Analysis for Silk heritage) [10], coined after Ada Lovelace, often considered to be one of the first computer programmers, and SILKNOW.

4.1. Dataset crawling

With our crawler [11] we are right now able to download datasets from 17 sources either via API or manual website crawling. All of the data is made publicly available by the respective museums or collections. We receive one more dataset directly from the Garin and UNIPA collections as they are part of SILKNOW.

The final output of the crawler is a unified JSON format: each JSON file contains two properties with single values, the ID of the record and the source URL. The latter can either be a link to the crawled website or directly to a machine-readable format like JSON via API. After that each crawled JSON file contains two arrays, one called “fields” with sets of different properties which depend on the original data. The other one is an array with all the images together with their respective URLs.

Inside the "fields" array the substructure is as follows: every field has exactly one label and then either one value or an array of values.


```
▼ 2:
    label:    "Date:"
    value:    "ca. 1785–1800"
▼ 3:
    label:    "Geography:"
    value:    "London or Birmingham"
▼ 4:
    label:    "Culture:"
    value:    "British, Birmingham or London"
▼ 5:
    label:    "Medium:"
    value:    "Steel, wood, leather, silk"
```


Fig. 7. Structure of the unified JSON format after crawling and downloading of the datasets. Example taken from the MET museum (21.180a–c).

In case of UNIPA the original format from the collection, which was Excel sheets, is converted to this common JSON format with the crawler. In case of Garin, all the integration takes place in the converter (see 4.2) and not in the crawler.

4.2. Converting and Interlinking

In all but the aforementioned case of Garin this common JSON format is then taken as the foundation for the converter software [12] in order to output Terse RDF Triple Language (Turtle) / TTL files that can finally be uploaded to a Triplestore based on the Virtuoso Universal Server.

The RDF conversion is based on a manual mapping for each dataset where fields with labels like "Técnica" and their values are mapped to properties like "P32_used_general_technique". As described in "3. Ontology", the SILKNOW ontology is based on CIDOC-CRM and the extensions CRMsci and CRMdig. Further extensions are still possible and currently under development. Two of the most central classes in our knowledge graphs are E22, which is used to represent "Man-made objects", and E12, which is a

[7] http://data.silknow.org/sparql
[8] http://data.silknow.org/fct
[9] http://skosmos.silknow.org/thesaurus/
[10] https://ada.silknow.org/
[11] https://github.com/silknow/crawler
[12] https://github.com/silknow/converter

Dataset / Museum	Downloaded files	
	Records	Images
IMATEX (CDMT Terrassa)	6802 (*3)	9201
Joconde	376	375
MAD (Musée des Arts Décoratifs)	763	763
MET (Metropolitan Museum)	8317	14208
MFA Boston (Museum of Fine Arts)	3297	3790
MTMAD (Musée des Tissus de Lyon)	663	2958
RISD (Rhode Island School of Design)	3338	4363
UNIPA	29 + 409	81 + 409
VAM (Victoria and Albert Museum)	7747	22773
CERES-MCU (Museos estatales del MEC)	1296	2872
GARIN	3125	6556
Mobilier	1296	1976
Smithsonian	147	147
Versailles	73	144
Venetian	1180	1182
Paris Musées	265	369
Gallica	230	458
Europeana	196	199
Musée d'Art et d'Industrie de Saint-Etienne	1195	1553

Table 4

Imatex has 3 records per object, one for each language. The UNIPA dataset has currently two different formats before conversion.

class for the "Production" of an object and properties like the production date and the material used.

Some strings need to be parsed with Regular Expressions (regex), for example the Dimensions field, to extract the exact width and height with its respective unit correctly. For instance, the regex pattern
`(\\d+(?:\\.\\d+)?) x (\\d+(?:\\.\\d+)?) cm`
in the MET converter makes sure to extract numbers before and after an x if the value ends with "cm". Furthermore, it makes sure to detect both integers as well as decimals. In some other cases we have one field in the JSON called e.g. "Auteur/exécutant" (Joconde) and it includes two different types of information: an actor and the role of the actor. In case of Joconde we can split it relatively easily as the order of them is always the same. The role becomes the property "ecrm:P2_has_type" of the class ecrm:E7_Activity, whereas for the actor its own class "ecrm:E39_Actor" gets created, which is also connected to the former by the property "ecrm:P14_carried_out_by".

For production dates we also developed a complex parsing and interpretation system to properly represent all dates and to make it possible to search for objects by their date. Originally many string literals were in different formats or some time periods were named differently in the different museums and languages. We can now interpret both single years, year ranges, centuries and most periods in all languages of our datasets. Every unique year or year range gets a unique URI, e.g. <http://data.silknow.org/timespan/1843> for the year 1843 that is linked with every occurrence of that year all across the data. In addition to that, we use the property "P86 falls within" to link every year with its corresponding century on Getty AAT.

Before some fields get mapped to classes and properties in RDF, their string values are getting checked if they are matching with some values in controlled vocabularies: places with Geonames, materials, techniques and motifs with the Silknow Thesaurus and the Getty Art & Architecture Thesaurus (AAT). In case of a match, the original string of the field of the dataset gets replaced with a URI of the concept in one of these vocabularies. For example: the technique "Embroidery" has the link "http://data.silknow.org/vocabulary/87" and a string that can be identified as either "embroidery", a synonym or translation like "Bordado" (Spanish) would trigger this linking.

4.3. Linked Data Publishing and API Access

To integrate SPARQL queries and their output into web development can be a challenge, even when the output format is JSON: It contains unnecessary metadata, each value has a datatype and is part of a bigger array with its own name and the attributes "type" and "value" or identical bindings that for example only differ in the language tag are not automatically merged and displayed multiple times. Mapping the results to another structure can be difficult, especially if avoiding to hard-code queries into the application's code.

With a combination of grlc [13] and SPARQL transformer [31] we were able to create an easy API access

[13]http://grlc.io/

for the SILKNOW knowledge graph, which makes it possible for web developers to avoid these problems [14]. With its graphical interface the knowledge graph can also be searched for any strings of the type time, location, material or technique and the output is displayed in a simpler JSON format.

4.4. ADASilk: An Exploratory Search Engine

Based on the aforementioned RESTful API stack we have developed the exploratory search engine ADASilk [15]. It is a user-friendly web interface to easily search and discover museum objects without the need to of any technical pre-knowledge. It is still work-in-progress, but it already offers an advanced search and many filters for all the different properties in the Knowledge Graph, for example production date or technique. Furthermore, it is integrated with the SILKNOW Virtual Loom, which makes it possible to create and modify 3D models of of the images of the silk textiles and their patterns. Finally, it has a spatio-temporal map view of the objects to further explore the geographic distribution throughout the ages.

5. Enriching and Validating the SILKNOW Knowledge Graph

5.1. Predicting missing metadata using image analysis

The information obtained from publicly available collections as described in section 4 is typically incomplete. One of the goals of SILKNOW is to tap digital images as an additional source of information. Given a sufficient amount of samples for which annotations are available, we use machine learning methods to learn how to predict the missing information from these images. Of course, this can only work for variables in the knowledge graph that are reflected in the visual appearance of a fabric. For a proof of concept, a method for the simultaneous prediction of three variables of the data model was developed: *production place*, *production timespan* and *technique or procedure*.

For the last few years, Deep Learning and Convolutional Neural Networks (CNNs) have emerged as powerful tools for the classification of images [32]. Consequently, they are also used for the prediction of meta-data of individual records in the SILKNOW Knowledge Graph from digital images. In this context, the prediction of one of the variables mentioned in the previous paragraph is considered to be a *task*. As we want to predict the metadata values of multiple variables simultaneously, we apply multi task learning (MTL) [33]. Our proposed network architecture takes RGB images of the size 224x224 pixels as its input. The first part of our network's architecture is a pre-trained ResNet-152 [34], which is used as a generic feature extractor. This pre-trained part of the network is followed by a fully connected layer that is shared across all tasks, i.e. it learns a shared feature representation that is independent from the specific task. In this way, we want to consider the fact that the tasks are inherently related. The network then splits up into three task-specific branches. Each task-specific branch consists of one fully connected layer and a final softmax layer which is responsible for predicting the class label of its corresponding task. The network architecture is depicted in figure 8.

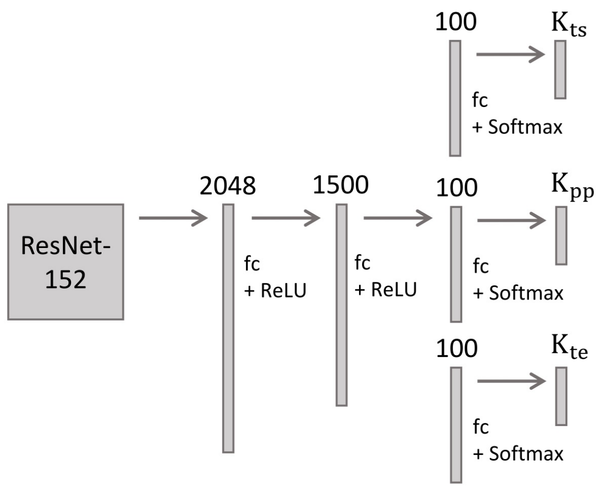

Fig. 8. Network architecture for multi-task learning. fc: fully connected layers. ReLU: rectified linear unit. K_{ts},K_{pp} and K_{te} are the numbers of classes for production timespan, production place, technique, respectively.

For training, we need training samples consisting of images with known annotations. In principle, each training image is classified by the current state of the classifier. The result is compared to the known class label of the training sample, and a loss that measures how well the prediction fits to that known class label is determined. Training itself consists of adapting the parameters of the CNN such that the loss function be-

[14] http://grlc.io/api/silknow/api

[15] https://ada.silknow.org/

comes minimal. The loss function used in our methodology is based on the softmax cross-entropy function [35], which we have modified in order to consider multiple tasks per sample:

$$E(\mathbf{w}) = -\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K_m} t_{nmk} \cdot ln\left(y_k\left(x_n, \mathbf{w}\right)\right), \quad (1)$$

where n is the index of a sample, m is the index of the task, k is the index of one of the classes of the m^{th} task. Furthermore, t_{nmk} is an indicator variable that equals 1 if the n^{th} sample belongs to the k^{th} class of the m^{th} task and zero otherwise, and y_k is the network's belief that a sample belongs to class k that is the result of the final softmax layer. During training, the loss function is minimised using stochastic minibatch gradient descent with Adapive Moments [36]. As each task of a training sample has a known class label, the sum of t_{nmk} over all k has to be 1 for all $m \in M$. In our application, the training samples are extracted from the Knowledge Graph. However, because of the nature of the Knowledge Graph, there are training samples for which this constraint does not hold, i.e. samples for which the class labels of some tasks are unknown (called *incomplete samples*). In our experiments we compared two versions of the CNN training. In the first version, we trained the CNN using only complete samples (MTL-C), i.e., samples for which the true labels for all tasks are known, while in the second case, we also considered incomplete samples (MTL-I).

The samples used for the training were extracted from a preliminary version of the Knowledge Graph (cf. section 4.1) only considering the IMATEX collection. The class labels that were considered for each task were obtained by mapping the values of each label from the knowledge graph to a meaningful class structure. In this way we generated a set of complete samples and a set of samples with incomplete annotations. Note that the set of incomplete samples includes the set of complete samples. The number of samples obtained for each class are shown in table 5.

For the experimental evaluation we trained our proposed network architecture with complete samples (MTL-C) as well as with incomplete samples (MTL-I). The evaluation results are presented in table 6. The table shows the overall accuracy, i.e., the percentage of correct decisions, for all three tasks. These numbers were determined in an experimental protocol involving five-fold cross-validation.

	Class name	Complete samples	Incomplete samples
TS	2nd half 19th c.	1022	1160
	1st half 20th c.	1611	2258
	2nd half 20th c.	488	1201
PL	Spain	394	2671
	Catalonia	2727	4322
	Italy	-	551
	Non-western	-	880
TE	drawing	1386	3854
	embroidery	336	359
	jacquard	1160	1276
	weaving	239	307
	damask	-	579
	velvet	-	500

Table 5

Overview of the class distributions for all tasks. TS: *production timespan*. PL: *production place*. TE: *technique or procedure*.

The results show that the class labels of previously unseen images could be predicted with an overall accuracy of over 92% if the CNN was trained only with complete samples. This indicates that our MTL-C CNN can be used to reliably predict the class labels for the considered tasks and, thus, enriching the SILKNOW Knowledge Graph. When also considering incomplete samples, the overall accuracy drops by 6% on average. On the one hand, these results indicate that the joint layer is dominated by incomplete samples, which leads to a loss of generality for the feature representation, resulting in a worse overall accuracy. On the other hand, it has to be noted that the requirement to have only complete samples places restrictions on the training data. For instance, table 5 shows that some class labels only occurred in incomplete samples; these classes cannot be differentiated from others if only complete samples can be considered for training. In future work, we want to improve the training with incomplete samples by only considering them for the task-specific branches. This way, we expect the joint layer to keep its generality, thus increasing the overall accuracy.

Task	MTL-C	MTL-I
production timespan	**92.3**	85.4
production place	**95.4**	86.0
technique	**92.9**	91.3
Average	**93.5**	87.6

Table 6

Overall accuracies [%] for MTL-C and MTL-I.

5.2. *Extracting structured metadata from textual descriptions*

The textual descriptors of the museums' records were used to find additional information about the records using information extraction techniques. At first, a handful of useful classes provided by domain experts has been selected for extraction from the textual fields, those classes include:

- **Technique used for the background of the textile**: Tabby(Tafetán), Satin(Raso), Twill(Sarga).
- **Technique used for the pictorial part**: Damask, Damassé(Adamascado), Espolinado(Brocading).

Figure 9 shows the distribution of the classes in the museums and the mentions in the categorical and textual fields of some of those classes.

Fig. 9. Frequencies of mentions of some of the target classes in the textual and categorical fields in museums.

Based on the target classes and the categorical and textual descriptors, a dataset for each target class has been formed to train and test a binary classification model for each of the classes. These models can then be used to extract the relevant techniques used in the textiles of the silk artifacts from its textual description.

5.2.1. *Dataset preparation*

Out of the museums data, a dataset for each of the target classes has been prepared. Initially, the specific categorical and textual fields in each records that contains the target classes has been inspected. Most of the matches were in categorical fields like 'technique', 'category', 'material', 'medium' and textual fields like 'description', 'medium (as textual field)', and 'details'. However, it has been notices that a lot of categorical fields contains mentions of more than one of the target classes at once, like having the same category field containing 'Satin' and 'Twill' or 'Damask' and 'Brocading'.

To resolve the conflict mentions of different classes in the category info of the same record, each textual field in each museum record has been determined to be positive, negative, or discarded example. A positive example is a textual field where only the target class(or one of its translations) exists in the categorical fields of its record. A textual field is negative if it includes the target class (or one of its translations) but the categorical fields of its record doesn't. A field is considered discarded otherwise. The discarded cases include the cases when the target class exists along with other classes, which gives no clear sign whether the example is a positive or negative.

An initial dataset has been formed based on the above criteria for all the target variables. To provide more negative examples for each class's dataset, the positive examples of the opposite classes have been added as negatives. For example, for the 'Damask' class, the positive examples of the datasets of the 'Damassé' and 'Brocading' classes have been added as negative examples of the Damask dataset if they were there before. Most of those items were classified as discarded due to the lack of mentions of 'Damask' in both the textual and the categorical fields.

Finally, we obtained the final datasets for each of the target classes. Out of those datasets, the top three datasets in terms of the number of positive examples have been selected for experimenting. The details of each of the chosen datasets is detailed in the next section.

5.2.2. *Case Studies*

Satin(Raso)

The first dataset to experiment with was the one with the class 'Satin'. The dataset was built from the English museums with additional negative examples brought from the positive examples of the datasets of classes 'Twill' and 'Tabby'. The final dataset has 4288 items, with 3430 training and 858 testing examples.

For feature extraction; TF-IDF, word2vec, and fasttext has been used. For classifiers; linear SVM, random forests, and gradient boosting decision trees have been selected. The table 7 shows the results of training using the stated embeddings and classifiers.

Embedding	Classifier	Acc.	P.	R.	F1
TF-IDF	Linear SVM	90.2	92	93.6	92.7
	Random Forest	86.6	93.5	85.9	89.6
	Gradient Boosting Decision Trees	87.8	91.1	90.6	90.8
Word2Vec	Linear SVM	83.7	90	85	87.5
	Random Forest	78.7	87	80.1	83.4
	Gradient Boosting Decision Trees	84.7	90.8	85.9	88.3
Fast Text	Linear SVM	84.6	90.2	86.4	88.3
	Random Forest	80.8	89.4	80.8	84.9
	Gradient Boosting Decision Trees	83.6	90.6	84.1	87.3

Table 7

Model Statistics of different embeddings and classifier against the 'Satin' dataset.

From the table, TFIDF with Linear SVM seems to be the best method with good accuracy compared to the baseline model of the majority class which was 66.9%.

As an alternative method of classification, the formal definition of the word 'Satin' has been used to form a feature vector by averaging the embedding of its words using the Fast Text word embedding. The embedding of the examples of the dataset has been formed in a similar manner. The example has been determined to be positive or negative based on the cosine similarity between its embedding and the definition embedding if it exceeds a certain threshold. However, after experimenting with different thresholds and comparing it with the actual labels of the examples, it turned out that the results were worse than that of a base model.

Damask(Damasco)

For testing how the model would perform on multilingual data, a dataset was prepared for the class 'Damask' in both English and Spanish museums' data. The dataset was built with the negative examples added from classes 'Brocading(Espolinado)' and 'Damassé(Adamascado)'. The dataset has 776 train and 195 test examples with a baseline majority class of 55.9%. Table 8 shows the results against TF-IDF as feature extraction method and the previously mentioned classifiers.

Embedding	Classifier	Acc.	P.	R.	F1
TF-IDF	Linear SVM	89.7	87.5	89.5	88.5
	Random Forest	85.6	87.2	79.1	82.9
	Gradient Boosting Decision Trees	80.5	81.6	72.1	76.5

Table 8

Model statistics of the 'Damask' dataset. TF-IDF embeddings have been used with with SVM, Random Forest, and Gradient Boosting Decision Trees Classifiers.

Brocading(Espolinado)

A similar experiment to the one done on 'Damask' has been repeated for the 'Espolinado(Brocading)' class. The resulting dataset has 683 train, 171 test, and a baseline of 50.9%. Table 9 shows the models statistics.

Embedding	Classifier	Acc.	P.	R.	F1
TF-IDF	Linear SVM	93.6	96.2	90.5	93.3
	Random Forest	92.4	96.1	88.1	91.9
	Gradient Boosting Decision Trees	92.4	93.8	90.5	92.1

Table 9

Model statistics of the 'Brocading' dataset with the same settings as in 'Damask' experiment.

6. Conclusion and Future Work

With this paper we presented you much of the finished and ongoing work of the multidisciplinary project SILKNOW and its central Knowledge Graph. For the development of a Knowledge Graph, we design an ontology and a data model, based on existing ones, especially CIDOC-CRM. We developed our own Thesaurus designed by domain experts, which contains concepts of the domain of silk textiles in 4 languages. We established and implemented a data collection, extraction, conversion and upload workflow. The SILKNOW Knowledge Graph and its already enriched data are now accessible both through an API and the exploratory search engine ADASilk. Finally, we presented results from our metadata prediction experiments.

From the ontology to the data enrichment many parts are however still under development. We are for instance working on introducing new classes and properties, specifically for the domain of silk textiles, which could be used by other projects afterwards. We still experiment on prediction of metadata through image analysis and natural language processing based on existing data. Especially with regards to our exploratory search engine ADASilk we also want to work on similarity and relatedness measures inside the knowledge graph. These are some of the areas that we will still refine in our future work.

The presented results are not only focused on the cultural heritage sector, but also for other stakeholders such as designers, artisans, educators, and especially to the community that is the holder of this living heritage that is silk. In this context, SILKNOW is committed to contribute to the four principles of the 10 European Initiatives that define what the European Year of Cultural Heritage 2018 stood for: Engagement, Sustainability, Protection and Innovation.

One of the main objectives of this project is to safeguard the intangible heritage and creativity associated to the European silk history. In order to do so, we are aligned with the principles of the UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions 2005 and in particular, principle no. 7, 'Equitable access'. In fact, this Convention follows the recommendation of another UNESCO Convention, for the Safeguarding of the Intangible Cultural Heritage (October 2003), fostering the equitable access to a rich and diversified range of cultural expressions from all over the world. Access of world cultures to means of expression and dissemination constitute important elements for enhancing cultural diversity and encouraging mutual understanding. Our tools are aimed at the development of best practices for the management, conservation and dissemination of silk heritage at local, national and European levels.

On the other hand, the silk heritage can be used as a boost for innovation and sustainable development. In this regard, we followed the Green Paper "Unlocking the potential of cultural and creative industries of the EU" which was the basis for the Creative Europe Programme, in support of the European cultural and creative sector. SILKNOW understands creative industries such as those traditional industries, such as Garin, as catalysts for social development.

7. Acknowledgement

This work was supported by the Slovenian Research Agency and the SILKNOW European Union Horizon 2020 project under grant agreement H2020- CULT-COOP-769504-1.

References

[1] Bernhard Haslhofer and Antoine Isaac, data.europeana.eu - The Europeana Linked Open Data Pilot (2011). http://dcevents.dublincore.org/IntConf/dc-2011/paper/view/55.

[2] E. Consortium, Definition of the Europeana Data Model v5.2.8, 2017. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf.

[3] Briatte, Katell, Modèle harmonisé pour la production des données culturelles, 2013. https://www.culture.gouv.fr/Media/Documentation/Harmonisation-des-donnees-culturelles/Files/MCC-HADOC-REF-modele_harmonise_donnees_culturelles.pdf.

[4] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio and K. Nyberg, CultureSampo: A National Publication System of Cultural Heritage on the Semantic Web 2.0, in: *The Semantic Web: Research and Applications*, L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou and E. Simperl, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 851–856. ISBN 978-3-642-02121-3.

[5] Eetu Mäkelä, Eero Hyvönen, and Tuukka Ruotsalo, How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo, *Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal* **3** (2012), 85–109. doi:10.3233/SW-2012-0049.

[6] Heim, Philipp and Lohmann, Steffen and Stegemann, Timo, Interactive Relationship Discovery via the Semantic Web, in: *The Semantic Web: Research and Applications*, L. Aroyo, G. Antoniou, E. Hyvonen, A. ten Teije, H. Stuckenschmidt, L. Cabral and T. Tudorache, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 303–317. ISBN 978-3-642-13486-9.

[7] M. Herzfeld, *The Body Impolitic: Artisans and Artifice in the Global Hierarchy of the Value*, University of Chicago Press, 2003.

[8] L. Zagato and M. Vecco, *Le culture dell'Europa, l'Europa della cultura*, 1st edn, Franco Angeli, Milano, 2011.

[9] P. Orefice, *Investigación Acción Participativa*, 1st edn, FUP/USACH, Firenze; Santiago de Chile, 2013.

[10] Doerr, Martin, and Kritsotaki, Athina, and Rousakis, Yannis, and Hiebel, Gerald, and Theodoridou, Maria, and others, Version 1.2 | CRMsci, 2014. http://www.cidoc-crm.org/crmsci/ModelVersion/version-1.2-1.

[11] Doerr, Martin, and Stead, Stephen, and Theodoridou, Maria, and others, Version 3.2 | CRMdig, 2014. http://www.cidoc-crm.org/crmdig/ModelVersion/version-3.2.

[12] C.E. Ore and M. Doerr, Version 6.2.1 | CIDOC CRM, 2015. http://www.cidoc-crm.org/Version/version-6.2.1.

[13] Doerr, Martin and Theodoridou, Maria, D14.1: Extended CRM – ARIADNE Infrastructure, 2016. http://legacy.ariadne-infrastructure.eu/resources-2/deliverables/d14-1-extended-crm/.

[14] LARHRA, OntoME SILKNOW project, 2019. http://ontologies.dataforhistory.org/project/15.
[15] Puren, Marie and Vernus, Pierre, SILKNOW CRM extension.
[16] M. Lecron-Foster, Symbolisms: the foundation of culture, in: *Companion Encyclopedia of Anthropology. Culture and Social Life*, T. Ingold, ed., Routledge, London, 1997, pp. 366–394.
[17] E. Alba, Catálogo e inventario como instrumentos para la gestión del patrimonio cultural, *Educación y entorno territorial de la Universitat de València* (2014), 67–93.
[18] M. Barroso-Ruiz, La normalización terminológica en los museos. El tesauro, *Revista General de Información y Documentación* **2**(4) (1994), 121.
[19] M. Baca, Fear of Authority? Authority Control and Thesaurus Building for Art and Material Culture Information, *Cataloguing & Classification Quarterly* **38** (2004), 143–151.
[20] P. Harpring and M. Baca, *Introduction to Controlled Vocabularies: terminology for art, architecture and other cultural works*, 1st edn, The Getty Research Institute, Los Angeles, 2015.
[21] F. García-Marco, Normas y estándares para la elaboración de tesauros de patrimonio cultural, in: *El lenguaje sobre el patrimonio: estándares documentales para la descripción y gestión de colecciones*, C.y.D. Secretaría General Técnica. Centro de Publicaciones Ministerio de Educación, ed., Ministerio de Educación, Cultura y Deporte, Madrid, 2016, pp. 29–46.
[22] N. Rodríguez-Ortega, Construcción y uso de terminologías, categorías de descripción y estructuras semanticas vinculadas al patrimonio en la sociedad global de datos, *El lenguaje sobre el patrimonio. Estándares documentales para la descripción y gestión de colecciones* (2016), 115–130.
[23] G.Schreiber, A. Amin, M.V. Assem, V.D. Voer, L. Hardman, M. Hildebrand, B. Omelayenko, J.V. Ossembruggen, A. Todai, J. Wielemaker and B. Vielinga, Semantic annotation and search of cultural-heritage collections: The multimediaN E-Culture Demonstrator, *Journal of Web Semantics* **6**(4) (2008), 243–249.
[24] A. Carretero, *Normalización documental de museos: elementos para una aplicación informática de gestión informática*, 1st edn, Ministerio de Educación y Cultura, Dirección General de Bellas Artes y Bienes Culturales, Madrid, 1998.
[25] C. Gunzburger, *The Textile Museum Thesaurus*, 1st edn, Textile Museum, Washington, 2005.
[26] E. Hyvönen, *Publishing and using cultural heritage linked data on the semantic web. Synthesis lectures on the semantic web: theory and technology*, 1st edn, Morgan & Cleypool publishers, Espoo, 2012.
[27] A.E. de Normalización y Certificación, *UNE 50124: Documentación: Directrices para la creación y desarrollo de tesauros multilingües*, 1st edn, AENOR, Madrid, 1997.
[28] M. Nielsen, Thesaurus construction: key issues and selected readings, *Cataloguing and Classification quarterly* (2004), 57–64.
[29] S. Bird, E. Loper and E. Klein, *Natural Language Processing with Python*, O'Reilly Media Inc., 2009. https://www.nltk.org.
[30] B. Fortuna and J. Rupnik and J. Brank and C. Fortuna and V. Jovanoski and M. Karlovcec and B. Kazic, K. Kenda, G. Leban, A. Muhic, B. Novak, J. Novljan, M. Papler, L. Rei, B. Sovdat, L. Stopar, M. Grobelnik and D. Mladenic, QMiner: Data Analytics Platform for Processing Streams of Structured and Unstructured Data, in: *Software Engineering for Machine Learning Workshop at Neural Information Processing Systems*, Montreal, Canada, 2014. https://sites.google.com/site/software4ml/accepted-papers.
[31] P. Lisena, A. Meroño-Peñuela, T. Kuhn and R. Troncy, Easy Web API Development with SPARQL Transformer, in: *18th International Semantic Web Conference (ISWC), In-Use Track*, Auckland, New Zealand, 2019.
[32] A. Krizhevsky, I. Sutskever and G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 25 (NIPS'12)* **1** (2012), 1097–1105.
[33] M. Dorozynski and D. Clermont, and F. Rottensteiner, Multi-Task Deep Learning with incomplete training samples for the image-based prediction of variables describing silk fabrics, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. (IV-2/W6)* (2019), 47–54.
[34] K. He, X. Zhang, S. Ren and J. Sun, Identity mappings in deep residual networks, *European Conference on Computer Vision* (2016), 630–645.
[35] C.M. Bishop, *Pattern Recognition and Machine Learning*, 1*st* edn, Springer, New York (NY), USA, 2006.
[36] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization, *3rd International Conference on Learning Representations (ICLR 2015)* (2015).
[37] E. Wilson, Active vibration analysis of thin-walled beams, PhD thesis, University of Virginia, 1991.