



Open Access to Data about Silk Heritage: A Case Study in Digital Information Sustainability

Jorge Sebastián Lozano, Ester Alba Pagán, Eliseo Martínez Roig, Mar Gaitán Salvatella, Arabella León Muñoz, Javier Sevilla Peris, Pierre Vernus, Marie Puren, Luis Rei, Dunja Mladeníč

► To cite this version:

Jorge Sebastián Lozano, Ester Alba Pagán, Eliseo Martínez Roig, Mar Gaitán Salvatella, Arabella León Muñoz, et al.. Open Access to Data about Silk Heritage: A Case Study in Digital Information Sustainability. Sustainability, 2023, 15 (19), pp.14340. 10.3390/su151914340 . hal-04314464

HAL Id: hal-04314464

<https://hal.science/hal-04314464v1>

Submitted on 20 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Open Access to Data about Silk Heritage. A Case Study in Digital Information Sustainability.

Jorge Sebastián Lozano ^{1,*}, Ester Alba Pagán ^{1,*}, Eliseo Martínez Roig ¹, Mar Gaitán Salvatella ¹, Arabella León Muñoz ¹, Javier Sevilla Peris ², Pierre Vernus ³, Marie Puren ³, Luis Rei ^{4,5}, and Dunja Mladenec ⁵

¹ Department of Art History, Universitat de València, Av. de Blasco Ibáñez, 28, 46010 València, Spain.

² Institute of Robotics and Information and Communication Technologies (IRTIC), Universitat de València, 46980 Paterna, Spain; javier.sevilla@uv.es

³ Laboratoire de Recherche Historique Rhône-Alpes (LARHRA), Université Lumière Lyon 2, 14, avenue Berthelot, F-69363 Lyon, France; Pierre.VERNUS@msh-lse.fr

⁴ Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia; luis.rei@ijs.si; dunja.mladenec@ijs.si

⁵ Jožef Stefan International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia; luis.rei@ijs.si

* Correspondence: jorge.sebastian@uv.es, esther.alba@uv.es

Received: date; Accepted: date; Published: date

Abstract: Silk textiles are a valuable yet very endangered part of world heritage. Museums entrusted with its preservation are slowly transitioning into the digital management of their collections, like most heritage institutions during the past two decades. It is not only a sensible decision vis-à-vis the inherent fragility of these objects, but also a responsible, sustainable approach in terms of efficient resource allocation. This article builds on the work done and lessons learnt within SILKNOW, a research project that aimed at enhancing the preservation and digital dissemination of silk heritage. It provides an overview of recent developments in the field of textile heritage, the tools implemented to allow the semantic access and text analysis of descriptive records associated with silk fabrics, and the spatiotemporal visualization of that information. Finally, it shows that institutional elements -namely, the creation and free dissemination of open data related to cultural heritage- are just as important as technical developments, showing why any future effort in these areas should take data sustainability into account.

Keywords: silk; cultural heritage; museums; open-access data; information sustainability.

1. From Silk Heritage to Digital, Public Humanities.

The conservation of cultural heritage requires its prior study and documentation. Through time, this highly specialized task has been done by individuals (researchers, scholars) and institutions (museums, collections, companies...). This has increased its respect, enjoyment and appreciation by the society at large.

Textiles have garnered less attention than other kinds of heritage. Not so many people pay attention to them; less institutions, in many cases of small or medium size and scarce resources, are devoted to it. It is usually a small and specialized field within heritage, even while its objects and intangible resources are connected in various ways to most members of our society. As a result of the high specialization that it requires, the efforts for its documentation show results of varying quality. An added problem is the inherent physical fragility of the textiles themselves.

In the last few decades, heritage institutions (also known as GLAMs: Galleries, Libraries, Archives, Museums) have been transitioning towards digital tools and platforms, in a process that is still ongoing. This general trend within heritage institutions and the humanities has also affected textile heritage. In some cases, large, national, well-funded museums are able to carry out large digitization projects. For the rest, the path can be much more uneven. Some parts of their collections might have been cataloged (rarely in their entirety), in-house systems are developed and then discontinued, databases become obsolete as technology evolves... But a fair amount of information in various digital forms is already available.

Often, these efforts suffer from the lack of a sustainable approach. Individuals, instead of teams, are in charge of them. Cataloguing staff may have little or poor training in digital tools, and in all cases, scarce funding and resources. The field is severely affected by an irregular, limited adoption of cataloguing standards. Terminology tends to lack normalization, a problem replicated and aggravated within each national or local language, as well as by the diversity of technical or historiographical approaches to the subject. The very technicalities that form the core of textile production (weaves, looms, patterns...) make it hard to deal with the topic, although the final products are very appealing for large sections of the population.

The SILKNOW project has aimed at providing an answer to some of these challenges, as section 3 in this article shows. This has been done thanks to digital tools and approaches, combined with scholarly expertise (from silk specialists, art historians and historians, textile engineers...). The final goal was to provide methods and best practices for heritage institutions that want to take their textile collections into the information and knowledge age. It paid particular attention to small and medium-size institutions, often lacking the technical resources and staff to venture into cutting-edge ICT and research. In this regard, the project offered paths towards the sustainability of their collections, their data, as well as to the fulfillment of their mission and institutional motivations.

2. Background and Conceptual Framework

2.1. Sustainability of Heritage Information: Toward Open Access.

Before speaking about information sustainability, it may seem worthwhile to speak about why museums or heritage institutions should be interested in this topic. Access to heritage and culture should be at the center for that conversation. It could well start two centuries ago, with the birth of large national museums across Europe, but we will not go that far.

Let us just remind now that access to culture is recognized as a human right in article 27 of the Universal Declaration of Human Rights, and that access to information about cultural objects is one of the best ways to protect and preserve them, for today's society, but also for future generations, so that both they and we can enjoy those objects, share our experiences around them, and learn about the people that created or used them.

The advent of the internet (or even more generally, the Information Age and the Network Society, following Castells' terminology [1]) has made this discussion even more poignant for cultural heritage institutions, traditionally charged with those tasks of preserving the memory of the past while making it accessible and understandable for current citizens.

The practical implementations of all these general principles vary greatly, of course, depending on several circumstances:

- Ownership: public institutions (meaning state-owned), or collections owned by private organizations or individuals.
- Funding models: be they fee-based, paid through taxes, established as non-profits but aiming at sustainability, and many different mixed approaches.
- Intellectual property rights (moral and economic ones) connected to the works, or their associated documentation have varying consequences on the display, dissemination and reuse of all that information.

- Information tools: catalogs and inventories kept only for professional and scholarly (internal) use; sometimes slowly and partially published over decades or centuries; sometimes disseminated through exhibition catalogs or research journals.
- Digital availability: varying degrees of transition to digital tools and repositories for all that information, a true wealth of data kept by heritage institutions.

One of the many aspects in this discussion has been the adoption of open-access policies within GLAMs, or CH institutions. The definition of “open” in open access is itself a contested topic, but for our purposes let us agree that “open refers to a policy or practice that allows reuse and redistribution of materials for any purpose, including commercial” [2, 3].

Examples of important museums that have made all or large parts of their holdings freely available on the web are well known: the Rijksmuseum, the Metropolitan, the Getty, museums belonging to universities such as Yale or Harvard... But is this policy only available for well-funded institutions? Incidentally, some examples indicate that a temporary closing for renovations or other reasons was an important push to take those decisions, making their collections available online while various reasons made physical access to them impossible. To our knowledge, COVID lockdowns did not have this exact kind of consequence -a big effort for the massive digitization of collections- among other museums. For good reasons, of course: lockdowns were not planned, and many other issues were more pressing for museums during that difficult time. Adaptation to changing environments and visitor behaviors, however, has now become much more evident as a good reason to invest into digitization and open access to museum collections.

This discussion is by no means recent [4]. Anyway, the important fact here is that some small and medium institutions have also adopted ambitious schemes of digitization, open access and (importantly) interoperability for the information about their holdings. For instance, OPENGLAM, an international network of heritage professionals, did some important work preparing a Declaration on Open Access to Cultural Heritage (<https://openglam.org/>).

We are addressing here particularly two of all the benefits that this approach involves. Our argument is that, once information is digitized (both images and catalog records) and incorporated into a structured data repository, sharing it across institutions is really at hand, in most cases. Secondly, instead of only expecting users to find our website among millions of other webs, an additional path to follow is to aggregate our data, our cataloguing records, into larger repositories.

There are many other reasons for shared repositories: ensuring the practice of better cataloguing and information management strategies, opportunities of increased visibility for small and medium-size institutions (that usually do not have such a big public profile among the general audience), better chances of being indexed and made visible by generic search engines, guaranteeing a permanent exercise of citizens’ rights of access to culture in spite of failing institutional abilities during any given crisis, sustainability of digital resources against data obsolescence and in order to ensure appropriate usage of the required investment.

Some challenges do appear quickly, too. Searching in a repository that contains dozens of millions of records seems daunting (but not much worse than any ordinary internet search). Many will say that more information does not equal more knowledge, invoking the all-too-present danger of infodisinformation. With data available in such huge amounts, doubts about information quality, relevance and homogeneity are perfectly understandable. Will non-expert users be able to find useful data in such a deep ocean?

As regards textiles, if we are dealing with fabrics, fashions, and in general with material objects of an inherently fragile material condition, digital preservation and access seem the only forward-looking option, as the following examples will show.

2.2. *Access to Textile and Fashion Heritage: Some Approaches.*

2.2.1. “Universal” Repositories.

Moving forward to some examples, We Wear Culture, from Google Arts & Culture, illustrates one possible approach. It is an aggregator of ad-hoc, usually highly curated content. In a way, it is the traditional answer, and a very successful one, if done properly. The innovation here is bringing together content from almost 200 different institutions and providing it in a compelling way, prioritizing extraordinary photographs and audiovisual content over consistent and extensive documentation.

In all likelihood, most non-expert users will feel more attracted to this approach. A pre-selection has been done by each institution and approved by the Google team. However, in terms of discovery of new information, or of specific pieces (beyond the usual collection highlights), it is fairly limited. Its searching abilities are very basic and irregular in results.

Sharing information across institutions in structured repositories is another approach, the one we are dealing with in the next two examples. It offers clear advantages and some prospects that are, at least, worth exploring.

- Opportunities for discovery: large databases can provide “windows” of visibility for less-known pieces, many times kept in storage, that are less likely to attract the attention of the general public, but which can be interesting for other experts or targeted audiences. This is the main benefit, and one that “only” requires having the cataloguing data in digital formats, adapting them to existing standards, and sharing them through available repositories.
- Workflow optimization: information generated primarily for institutional, internal usage can to a certain extent be repurposed for later, external reuse, instead of incurring the costs of time-consuming, one-off curated content publishing.
- Multilingualism: joint efforts are helping to overcome linguistic barriers. Thanks to automated translation and, in specialized contexts, multilingual thesauri, it is possible to gather information in different languages, and not just the language employed by the user to interrogate the system. Museum catalogs tend to be rather specialized resources that use scholarly terminology. Therefore, it will always be better to count on multilingual controlled vocabularies and not just general automated translation.
- Opportunities for automatization of some tasks. Large bodies of information (for instance, objects covering an entire period or style) are hard to grasp in their entirety, even for experts. Artificial intelligence and big data might be ready to help us in some cases, where computers can take care of repetitive and cumbersome tasks. For instance, searching for previously unknown shared features, or for unexpected patterns within large numbers of objects and records, both in visual analysis and in textual analysis. Automated annotation might be a great help for catalogers, providing suggestions based on comparison with many other instances, but always ensuring that the AI-generated content is curated and supervised by domain experts.

Summing up, massive, shared repositories can help to go beyond the walls of individual institutions, of whatever size; but they should be particularly useful for smaller museums and collections, such as the ones scattered throughout Europe as memories of the essential roles that textile industries once played across the continent.

2.2.1.1. Europeana

Europeana is one of the largest experiments in open access to culture that the internet is making possible. It works with European archives, libraries and museums in order to share cultural heritage, providing access to millions of books, music, artworks and other content.

The important thing here is that Europeana brings together cataloguing records and digital surrogates from -literally- thousands of institutions. This might seem a simple accumulative effort, but far from it, it is a technical feat of data harmonization and interoperability. A large part of it

comes from national libraries, but museums and collections of material culture also have an important presence in the repository. It is decentralized in nature, working through national and thematic aggregators instead of a single data ingestion node. Very diverse institutions, regardless of size, share their contents and become data providers for Europeana.

It is very revealing that one of the first thematic clusters within it was devoted to fashion and costume (<https://pro.europeana.eu/project/europeana-fashion>). Europeana Fashion currently gathers around one million records of cultural objects related to fashion, from catwalk photographs to drawings from the great designers of couture brands. It was born as a research project that later became a network of fashion-related institutions and an aggregator for this kind of contents into Europeana.

2.2.1.2. Wikimedia Commons

A different model is that of Wikimedia Commons, the file repository that hosts public domain and freely licensed media content for the various projects of the Wikimedia Foundation. The Wikipedia is just one (and the most used) among them.

This approach is very different to the previous ones. Wikimedia Commons can be used to find images (or multimedia) of cultural heritage, but not as a direct provider. Instead, search engines increasingly rely on Commons as the first option in image searches about historical cultural objects. Since, most often, images uploaded to Wikimedia have little or no limitations to their reuse, they get downloaded and copied in ever-increasing numbers. Any cultural institution should ask itself what to do, in this regard: whether to fight a long, uphill battle to get their own website as the top search result for objects kept by them; or to “join the enemy” and simply make sure that the image shown in Wikimedia is one provided by them, properly referenced and linked to the owning institution.

We have outlined just two global repositories that contain data about textiles and fashion, among many other subjects, of course. This short overview simply aims at making clear that digital platforms offer various models for the dissemination of cultural heritage data – particularly, about fashion and textiles. Those platforms evolve over time, and in this regard as in others, Europeana seems the most stable option for any institution within Europe that wants to share their collection information beyond their own digital resources.

In any case, it does not have to be an either / or dichotomy. All three approaches can be useful for the same institution, and even for the same user.

2.2.2. National Databases.

A few countries have followed an approach that aims at a central, state-wide repository for material heritage, or at least of some of its varieties. These catalogues require a substantial effort to establish cataloguing guidelines, use common schemes for the records (i.e., data models) and employ controlled vocabularies. Some kind of shared software, within a decentralized system, or an online platform for a centralized one, are common features for these databases, too. These coordination efforts pay off in the longer run, enabling users to access data from dozens or hundreds of museums through a single gateway, instead of having to search for them at each individual institution. They also provide greater financial efficiency, sharing one information system and software application among many museums, instead of having each one of them make the expense to develop or acquire their own solution. On the downside, it is also worth noting that a single record scheme may not always do justice to many kinds of data about heterogeneous museum objects, such as Baroque altarpieces, folk crafts, traditional African masks, filmed records of performance art, or textiles. The advantages in easy, homogeneous access must balance the losses in information specificity and detail.

- Joconde is the classical model in this regard. This database, created and maintained by the French Ministry of Culture, as of 3 July 2023, gathers 661164 records from more than 250 museums having received the legal status of “Musée de France” [5]. Its records are also

shared through other platforms, notably, on POP, the Plateforme Ouverte du Patrimoine (<https://pop.culture.gouv.fr/>). Some 21,000 of them are related to textiles and costumes [6].

- CERES, the Red Digital de Colecciones de Museos de España, offers a similar framework. It is built on an information system named Domus, developed by the Spanish Ministry of Culture and currently used by 195 museums throughout Spain, both public and private. While the system was originally built for the internal management of the collections, sharing the catalogue records through the Ministry's centralized repository is a permanent feature of the software. This repository is then made public online through the CERES website. It also relies on a set of common controlled vocabularies and cataloguing rules. It covers large parts of Spanish heritage kept in museums, but it cannot be said to be fully comprehensive, either. Some regions have developed their own, independent systems. Even among museums contributing to CERES, the quantity and depth of their records on the platform can vary widely. Despite such shortcomings, it offers a tremendous amount of information, and serves as an outstanding example of the feasibility and advantages of centralized repositories. In its current version it offers more than 341,000 records from 118 museums (<http://ceres.mcu.es/>). It is particularly useful for small and medium institutions: among them, many specializing in textile heritage. They can benefit greatly from shared resources like Domus and CERES, as they usually lack the funding, human resources and expertise to embark on large digitization campaigns on their own.
- A partly similar approach lies at the basis of BeWeB, the census of heritage owned by Catholic dioceses and institutions from Italy (<https://beweb.chiesacattolica.it/>). While organized by a private institution, the Italian Bishops' Conference (CEI - Ufficio Nazionale per i beni culturali ecclesiastici e l'edilizia di culto), it offers a coverage even larger than the ones just mentioned. It contains records on more than 10 million objects, including archival documents and books, with historical and artistic objects exceeding 4 million [7]. Inevitably, the quality and standardization of all these records is quite a challenge, and often offers ample room for improvement. The resource itself, however, is staggering in its ambition and reach, and offers a good example for private owners of art historical heritage.

The last two instances show, on the other hand, some of the limitations of the centralized model. Even when controlled vocabularies are available, cataloguers do not always follow them consistently. In CERES, identical pieces may be catalogued as either "Textiles" or "Tejidos" (or as any of their many subtypes), which makes systematic recovery quite unpredictable, sometimes. Semantic web technologies can help to overcome these problems, but only to a certain extent. On the other hand, these repositories bring together records prepared over long periods of time (decades, sometimes), in widely different institutions, about very heterogeneous records, by catalogers with varying levels of expertise and dedication to the task. Resulting records are also dissimilar, in quality, depth and scientific validity. In any case, the main advantages for these large repositories are, again, the new opportunities they provide for discovery into the less visible parts of our heritage, for the cross-reference of objects between institutions or across disciplines, for quantitative analysis and innovative visualizations.

2.2.3. Major Museums.

In many developed countries, large museums of national -and sometimes even global, encyclopedic- reach are guardians of massive historical holdings. Some of them are owned by the State, usually originating from royal -and national, later- collections. This is the situation in many European countries. In other cases, equivalent institutions have been formed during the 20th century, through acquisitions and bequests. Some of them are more focused on fine arts than on decorative arts, but their textile holdings can be impressive, in any case. Others are specialized in decorative arts, traditional crafts or modern design, but they remain large-scale institutions with dazzling collections.

For a number of institutional reasons, these large museums tend to offer their catalogues independently, not as part of national repositories like the ones in the previous section. This gives

them more visibility and reinforces their exceptionality. Moreover, their leading position often means that they count on the human and technological resources required to have a strong online presence, including comprehensive, in-depth cataloguing information on most of their holdings. Some of them are also champions of the open-access movement among museums.

We may group in this category museums such as:

- The Victoria and Albert Museum in London was created in the mid-19th century with a focus on the applied arts and on science. Cataloguing records on part of their collections are available on their website, numbering more than 1.2 million. Their holdings of textiles and fashion are truly impressive: their online presence almost reaches 80,000 pieces, not including embroidery and fashion items. They are also accessible through an API, an uncommon but forward-looking feature for museums (<https://developers.vam.ac.uk/>). Similar institutions in other European countries are the Musée des Arts Décoratifs in Paris and the Museum für angewandte Kunst in Vienna.
- The Musée des Tissus in Lyon (<https://www.museedestissus.fr/>), until recently known as the Musée des Tissus et des Arts Décoratifs. Widely considered as the best European collection of historical silks, it is also exceptionally rich in fashion and other textiles, with 2.5 million pieces in total. Online access to such a vast collection is quite limited, however. As a private establishment, dependent on both public and private funding, it has suffered serious institutional crises in the last years that seem to have been overcome by now.
- The Metropolitan Museum of Art, in New York City, is an encyclopedic art museum, and a public/private partnership. It houses world-class collections in many kinds of fine and decorative arts. In the last decades, its Costume Institute has gained huge visibility for its temporary exhibitions and celebrity-oriented events, such as the annual Met Gala. Less known but equally important are the textile collections. It is a global leader in the field of open access to collections, providing full information and high-resolution images on some 400,000 pieces, from the total 1.5 million objects it holds (<https://www.metmuseum.org/about-the-met/policies-and-documents/open-access>). Some 40,000 catalogue records on textiles are freely available on its website.
- The Smithsonian Institution is a system of research centers, libraries and museums, part of the US federal administration. One of its 19 museums is the Cooper Hewitt Design Museum, located in New York City. While it may seem focused on modern design, it houses impressive historical collections, too, with textiles among them. It is also a champion of open access, as part of the Institution's general policy, featuring an API (<https://edan.si.edu/openaccess/apidocs/>) and an image repository shared across all of the Smithsonian's collections (<https://www.si.edu/openaccess>).

2.2.4. Other Projects:

Research projects are providing yet another model for the study and dissemination of textiles by means of online repositories. Sometimes they are linked to individual institutions, while in some other cases one of their goals is precisely to bring together resources from separate organizations, adding the challenge of interoperability of shared information. Some have been born as aids for a personal research project, later outgrowing that framework in order to incorporate data from other researchers. Others are born from a private grant or thanks to competitive funding received from national or international research agencies. The sustainability of these projects is open to question, since the information obtained and the expertise gathered during the years of their design, implementation and publication, are seldom kept in use for a long time, after the funding disappears. However, a proper institutional support, one that incorporates future maintenance and -when necessary- technical support for migration and scalability, can help to overcome this problem. It is also essential to adopt an open framework for the information, allowing present or future semantic linking between information resources.

- SilkMemory is, according to its own website <https://silkmemory.ch/>, a “web portal [that] provides access to the archive database of the Lucerne School of Art and Design with

digitised text and image sources about the silk industry of the Canton of Zurich". Born after the commercial demise of the once-thriving Swiss silk industry, it was funded by the Zurich cantonal government, and went online in 2018. It provides a thoughtful answer to a danger that is common to many European countries: the dispersal and loss of the valuable archival and material heritage generated by those industries, most of which have gone out of business during the last decades. It offers a database of fabrics, books and images kept in those archives, together with a selection of some personal or institutional stories obtained from the same archival fund.

- ART-CHERIE (Achieving and Retriving Creativity Through European Fashion Cultural Heritage Inspiration - <https://www.artcherie.eu/>) was a project funded by the Erasmus programme of the European Commission, lasting from December 2016 to May 2019. It brought together partners from Belgium, Greece, Italy and the United Kingdom, from a quite broad scope, including an interesting connection to the training of fashion designers. Among other outputs, it aimed at providing a Digital Database or "Catalogue and Digitisation of Museo Prato Exhibits and Collection". However, it is not openly accessible.
- María Judith Feliciano, an independent scholar specialized in Medieval Iberian textiles, is the principal investigator of the "Medieval Islamic Textiles in Iberia and the Mediterranean" research project (<https://maxvanberchem.org/en/scientific-activities/projects/art-history/16-histoire-de-l-art/160-medieval-islamic-textiles-in-iberia-and-the-mediterranean-2>). Funded by the Fondation Max van Berchem in 2016-17, it was a crossroads for a number of research projects from other scholars in the same field, like Ana Cabrera, Laura Rodríguez Peinado and Therese Martin. It reportedly aimed at producing a website and database to make available the results of the research carried out, but these have only been disseminated through articles and essays.
- IMATEX is the online database offering information about the collection of the Centre de Documentació i Museu Tèxtil de Terrassa (<http://imatex.cdmt.cat/>). Created in 1996, it was originally built as a gateway for designers searching for inspiration in CDMT's historical collection, and later transformed into a generic online information resource, open for everyone [8]. It is extremely rich in content, including costumes, accessories, designs, paraments, sample books, a library and an outstanding collection of more than 9,000 textiles. Available in Catalan, Spanish and English, initially it was made possible by the European Regional Development Fund, and by the CDMT's own budget afterwards.
- The MINGEI project aims at exploring the possibilities of representing and making accessible both tangible and intangible aspects of craft as cultural heritage (<https://www.mingei-project.eu/>). One among the crafts under study is silk weaving, led by one of the project partners, Haus der Seidenkultur in Krefeld. It is a Horizon 2020 project, led by FORTH, and it is being carried out between 2019 and 2022. The project does not directly intend to build a database, but rather a repository of innovative storytelling models, including interactive Augmented Reality and Mixed Reality. It does have a strong emphasis on developing content description tools that comply with existing semantic web standards, such as CIDOC-CRM.
- The PARVENUE project was recently funded by the German Federal Ministry for Education (<https://www.parvenue-projekt.de/>). Led by art historians from the Heinrich-Heine-Universität in Düsseldorf, it is built on a tight collaboration with the Deutsche Textilmuseum in Krefeld (<https://www.deutschestextilmuseum.de/>), one of the European capitals of silk industry. One of the areas of the project is built on a preliminary cataloguing of the collection of 30,000 fabrics and costumes in the museum, not yet available online.
- Another recent initiative is the Restaging Fashion project, based in the Lipperheidesche Kostümbibliothek – Sammlung Modebild in Berlin, and the Fachhochschule Potsdam (<https://uclab.fh-potsdam.de/projects/restaging-fashion/>). Active between 2020 and 2023,

and also funded by the German Federal Ministry for Education, it purports to build an online catalog of costumes, prints and drawings, held in different institutions, adding 3D visualizations of some of these objects.

- Finally, the authors of this article have been involved in two projects focused on the dissemination and sustainability of heritage via digital tools and platforms. One of them is SILKNOW, a Horizon 2020 project, active between 2018 and 2021, that among other things has built ADASilk, a repository of some 40,000 records about silk-related objects from different museums and collections (<https://ada.silknow.org/>). The joint team of art historians and computer scientists in Universitat de València has also been working on SeMap (<https://www.uv.es/semap/>), a project funded by Fundación BBVA and built on the data from Spanish museums made available by CERES, the web portal of the Red Digital de Colecciones de Museos de España presented in a previous section.

3. Silk Heritage Resources and Tools

3.1. *The SILKNOW thesaurus*

Smeets [9] affirms that since traditional, intergenerational ways of transmission are losing ground worldwide, additional ones are being and must be explored. Most heritage contains both tangible and intangible elements, whose proper safeguarding requires careful documentation of the link between them, i.e., terminology. While cultural heritage institutions strive to use controlled vocabularies based on their own collections [10], some efforts have been done with the aim of standardizing all available vocabularies, regardless of important differences among them, being produced by different professionals, by different nationalities or even different disciplines, etc.[11] In such a recent date as 2016, a proposal was made within an ICOM General Conference to develop a textile thesaurus based on merging and enlarging the existing vocabularies [12].

For instance, the CIETA Vocabularies have been and still are the most common terminological resource within the field of historical textiles. CIETA, the Centre International d'Étude des Textiles Anciens, based in Lyon, offers a hub for researchers, collections and institutions. Its vocabularies (<https://cieta.fr/cieta-vocabulaire/>), available in all the major European languages, have been continuously expanded and translated, with the latest versions being made fully available online only very recently, for the first time.

The SILKNOW project has built a multilingual thesaurus dedicated to the specific vocabulary of historic silk textiles, which also includes local term variants [13]. In addition to giving a chance to preserve and transmit heritage, we seek to give a useful tool to the target institutions which, at the moment, are using different terms to describe their objects [14]. A thesaurus is a controlled vocabulary, but also a hierarchical tool, one that incorporates the relationships between terms: hierarchical, equivalence, association, etc. Terms related to silk textiles, productive crafts, motifs, etc. can vary greatly from one context to another, which makes information related to them less accessible. These local and traditional terms are usually forgotten or ignored, since they are only present in archival records. However, such specialized lexicon is in use among practitioners, especially in the domains of traditional knowledge and handicrafts, and it needs to be collected for its proper preservation. Documenting and standardizing this terminology is thus a great help for professionals, students, and researchers.

The SILKNOW thesaurus also covers the already noted need for a common framework, a standard tool that could gather as many terms as possible, with all their variants or synonyms, for independent institutions (<https://skosmos.silknow.org/>). Nowadays, the collaboration among museums and collections requires tools that foster data interoperability. A multilingual thesaurus not only facilitates information exchanges across collections and institutions, but it also makes heritage accessible to non-specialist audiences by lowering language barriers [15]. Making it freely accessible, reusable and linkable to other resources is the way to go, if it is intended as a sustainable tool. In fact, the SILKNOW thesaurus has been built as an extension to the most widely used

thesaurus within the cultural heritage community, the Getty Foundation's Art and Architecture Thesaurus (or AAT - <https://www.getty.edu/research/tools/vocabularies/aat/>).

As of July 2023, the latest version of the thesaurus has 666 preferred terms and more than 600 alternative terms in the 4 languages in which it was developed. Its validation [16] was carried out following a coverage analysis which permitted the validation of textual data of online resources on the full thesaurus in all four languages it covers. We calculated the frequency of the individual thesaurus concepts that are present in the data coming from collections included in SILKNOW. Later, we compared two words and determined whether they have the same stem. Additionally, the thesaurus was evaluated among domain experts who compared data from SILKNOW ontology with the concepts included in the thesaurus. The result is an interdisciplinary and multilingual thesaurus that covers not only the most frequent concepts used in museums but also those that are used in academic papers and in the current and traditional silk industries. Hence, not only protecting tangible and intangible heritage but standardizing silk heritage language.

3.2. *The SILKNOW Ontology*

The data collected by SILKNOW is by nature heterogeneous. Each institution has its own cataloguing practices, and these practices may have evolved over time. The resulting metadata can therefore vary greatly. The inherent heterogeneity of these data results in the creation of data silos that are incompatible with each other, and therefore mutually incomprehensible [17]. Moreover, data heterogeneity is further increased by the multiplicity of languages used. This makes the discovery of these data all the more difficult, as it requires users to master various languages and very different information management systems, as well as explicit or implicit data models. To overcome the problem posed by data heterogeneity, SILKNOW uses a formal ontology, to formalise these cultural heritage data in a logical language made up of classes and relations [18]. This coherent and uniform representation of information will facilitate the discovery of information which, until then, was difficult to access due to its fragmentation in incompatible data silos.

To develop an ontology which will thus enable a better data integration, we have chosen to use the CIDOC Conceptual Reference Model or CIDOC CRM, developed to express the underlying semantics of cultural heritage documentation. Recognised by both the museum and ICT worlds, the CIDOC CRM is also an international standard, recognised as an ISO standard for version 5.0.4. It should be noted that the latest version of the CIDOC CRM is the version 7.2.2 published in October 2022 (<http://www.cidoc-crm.org/versions-of-the-cidoc-crm>); but we are currently using the version 6.2, which was the latest version published at the time this work began.

The creation of this ontology required different steps to be taken. First, we analysed and compared numerous records coming from different cultural heritage institutions - such as the Victoria and Albert Museum, the British Museum, the Musée des Tissus in Lyon, the Garin collection at the Museu de la Seda in Moncada, the Musée des Arts Décoratifs in Paris, and various French museums via the Joconde database -, and consequently respecting different data models and cataloguing standards. We also relied on the standards and documentation used by these institutions to produce metadata, notably the inventories in French museums [19], the HADOC Harmonized model for cultural data production [20], the ICOM guidelines for Museum Object information [21] and the Europeana data model [22]. We then drew up a list of the descriptive fields most commonly used by cultural heritage institutions to describe the textiles they preserve, eliminating those that are not of interest to the SILKNOW project - in particular, we have not selected the information concerning the administrative management of these artifacts. These descriptive fields were then grouped into "information groups". We have, for example, defined an "Object acquisition information group" to identify the acquisition method and date, the previous owner of the object, its current owner, and additional information about the acquisition:

Table 1. Contents of the object acquisition information group.

Object acquisition and legal status information group
--

Definition: information about the acquisition and ownership of a cultural heritage object. Several such information groups can be available for one object depending on the history of the object.

Acquisition method

The method by which an object was acquired.

Ex: gift; purchase

Acquisition time-span

The timespan or the date of acquisition of the object.

ex: Before 1998; 1950

Previous owner

The name of the person from whom, or organization from which, the object was acquired.

New owner

The name of the person who, or organization that, acquired the object.

Acquisition complement

Any additional information about the acquisition of the object.

Acquisition note

If necessary, additional comment on the acquisition of the object

These categories of information allow us to make the best use of the functional overviews provided by the CIDOC CRM official documentation (<http://www.cidoc-crm.org/functional-units>). These functional overviews divide the CRM entities and properties into different categories of information, with their graph representation, thus offering technically neutral templates of modelling applied to the metadata that describe cultural heritage artifacts. For example, in the functional overviews, we find the category "Acquisition information" which corresponds to "Object acquisition information group". This preliminary step is therefore particularly useful because it allows us to rely on the functional overviews to express the semantics of these "information groups" with the entities and properties of the CIDOC CRM. This categorisation of information has made it easier to express the underlying semantics of these descriptive fields; it enabled the selection of CRM classes and properties capable of expressing the meaning of these categories. For example, the Object Acquisition information group was thus expressed with the following classes and associated properties:

Table 2. CRM classes and properties used to express information on acquisition.

Domain	Property	Range
E8_Acquisition	P14_carried out by	E39_Actor
E8_Acquisition	P22_transferred title to	E39_Actor
E8_Acquisition	P23 transferred title from	E39_Actor
E8_Acquisition	P24_transferred title	E22_Man-Made Object
E8_Acquisition	P7_took place	E53_Place
E8_Acquisition	P4_has time-span	E52_Time-Span

This first step is followed by the mapping process which consists in producing semantic data from the data produced by the cultural heritage institutions and stored in relational databases, giving them an equivalent semantic expression by means of the chosen formal ontology. The mapping

process was produced manually by domain experts in collaboration with computer scientists. The method adopted is the one suggested in [23], which proposes to interpret each of the descriptive fields as entity-relationship-entity (e-r-e). More precisely,

- tables and columns in the relational database are interpreted as entities;
- complete records are interpreted as entity instances;
- fieldnames are interpreted as both relationships and entities;
- and field contents are interpreted as entity instances.

The whole scheme is decomposed into e-r-e's, and each e-r-e is aligned with the CIDOC CRM [24]. In other words the mapping consists in interpreting these entities and relationships and in expressing them in CIDOC CRM semantics. In doing so, we aim to preserve as far as possible the original meaning of the data. Concretely this process produces triplets that link nodes together through properties, forming a network of human- and machine-readable data and enabling information exchange and integration.

Given the data heterogeneity, carrying out this mapping process implied precisely understanding their meaning. After studying the structure of the different catalogues from which the data were extracted, we analysed their contents to understand what information was expressed in them, but also to assess their internal consistency. Cataloguing practices within the same institution may have varied over time, and the meaning given to these descriptive fields may also change, depending on the practices adopted. As a result, the consistency of content is generally weak. Based on the categorization of the data carried out previously, we not only selected the classes and properties most likely to express the semantics of this information, but we also refined this initial selection by adding new classes and properties, and removing those that proved to be useless.

To understand the mapping process, it is necessary to mention that we have chosen to use the CRM class *E22_Man-Made Object* to model the artifact preserved and therefore described by cultural heritage institutions. Indeed, the CIDOC CRM uses this class to model "physical objects purposely created by human activity". This class is therefore at the centre of the SILKNOW ontology. In the following example from the collections of Museu de la Seda in Moncada, the field "Denominación principal" contains the title given by the heritage institution to the object kept in its collections. We can express the underlying semantics as follows: the title of the artifact is "Abundancia" in the database. This means that we can interpret this field as a title, modelled with the class *E35_Title* in CIDOC CRM. The fieldname describes the relation that exists between the object (*E22_Man-Made Object*) and its title, which implies interpreting it with the property *P102_has title*.

Table 3. Information contained in the descriptive field "Denominación principal" and its mapping in CIDOC-CRM.

Fieldname	Content	Path
Denominación principal	Abundancia	E22_Man-Made Object P102 has title E35_Title

The SILKNOW ontology, consisting of the selected classes and properties, is publicly accessible (<https://ontome.net/profile/7>) and documented via OntoMe [25], an ontology management system developed by the LARHRA research centre into which the CIDOC-CRM documentation has been imported. To model the data collected by the SILKNOW project, we have therefore used part of the classes and properties proposed by the CIDOC CRM model, but also those offered by an extension of this model, the Scientific Observation Model (CRMsci) [26] which is a formal ontology elaborated to integrate metadata about scientific observation. We have more particularly used the class *S4_Observation*, defined as "the activity of gaining scientific knowledge about particular states of physical reality gained by empirical evidence, experiments and by measurements". This class seemed to us quite appropriate to model the historical and technical analyses resulting from the observation of ancient fabrics.

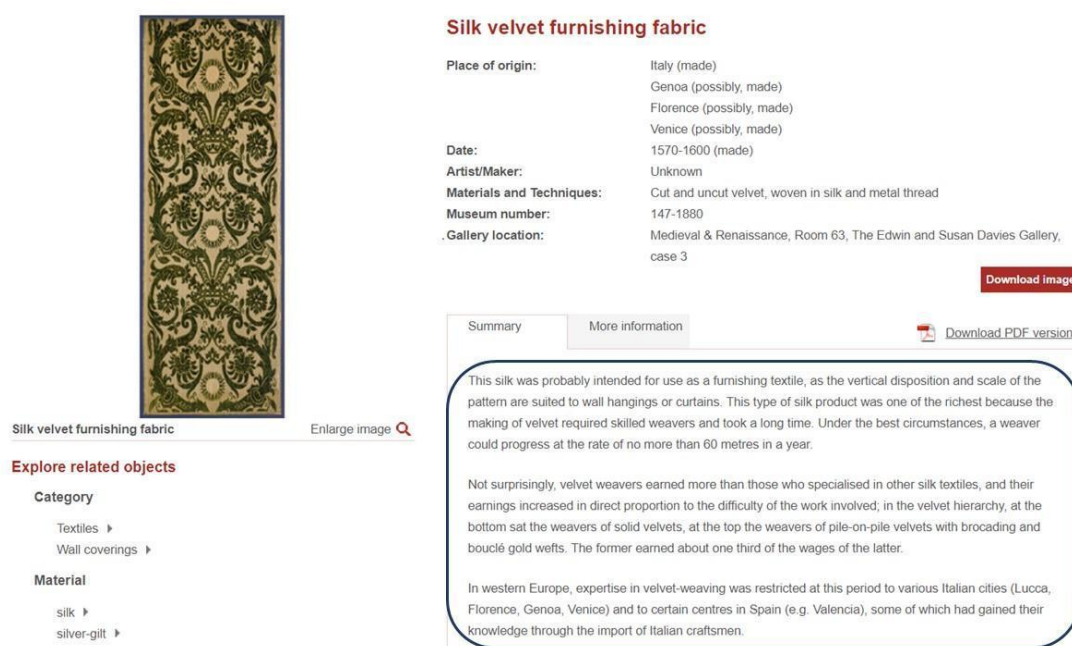


Figure 1. Example of historical and technical analyses from the Victoria and Albert Museum catalogue

The quality of the data model was assessed by providing mapping rules between cultural heritage institutions' records and the SILKNOW ontology. We observed that all fields can be represented using the existing classes and properties of the SILKNOW ontology. In practice, we selected two representative records from each dataset, and provided mapping tables and associated RDF graphs. The RDF graph below shows triplets that we have systematically created to model crucial information about the object described. On this graph, we visualise the triplets modelling the information about the production of the artefact (*E12_Production P108_has produced E22_Man Made Object*): when it was produced (*E12_Production P4_has time-span E52_Time-Span*), where it was produced (*E12_Production P8_took place on or within E53_Place*) and by whom it was produced (*E12_Production P14_carried out by E39_Actor*). As we are studying ancient fabrics using silk and specific manufacturing techniques, it is also essential to model information about the material(s) used (*E12_Production P126_employed E57_Material*) and the techniques employed (*E12_Production P32_use general technique E55_Type*) - information that is usually detailed in historical and technical analyses (*S4_Observation O8_observed E22_Man-Made Object*).

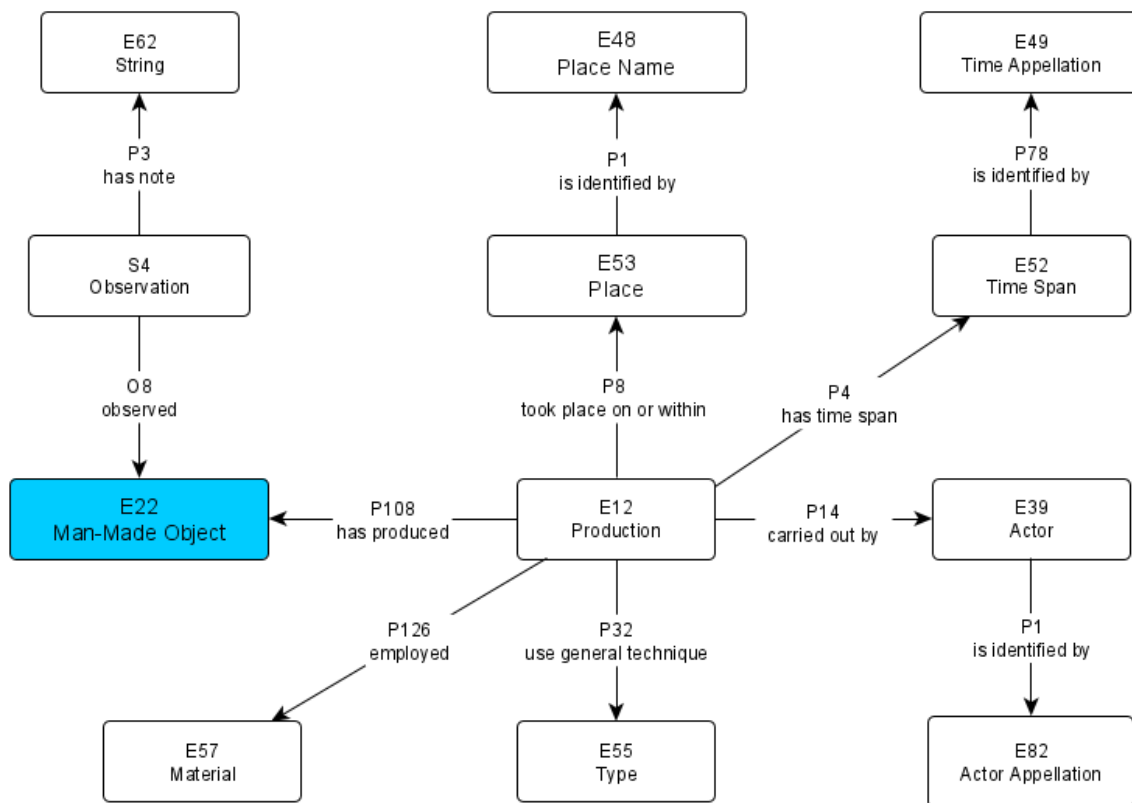


Figure 2. RDF graph : weaving process modeled with CIDOC CRM and CRMsci

SILKNOW has also developed text and image analysis methods which, from the data describing silk-related objects, infer new properties on these objects, and ultimately enrich the existing metadata. We have thus modelled the integration of the new data produced by these analyses. The modelling should make it possible to make a clear distinction between these predictions and the original data, and to provide the ADASilk users with information on the degree of reliability of this information. For this, we have chosen to use the Provenance Data Model (Prov DM) [27], recommended by the W3C. Image or text analyses are represented in the form of a *Prov:Activity* which can be qualified by a type (image analysis or text analysis). Depending on the case, this *Prov:Activity* takes an *E38_Image* (image analysis) - or or a text - *E62 String* (text analysis) - as input (*prov:used*) and produces two statements as output (*prov:WasGeneratedBy* properties). Each of these declarations has an *E54_Dimension*. The date of the analysis can be specified (*prov:AtTime*). If necessary, we can specify the analysis module with a *prov:Agent class* (of type Software Agent) and document it (*E31_Document*).

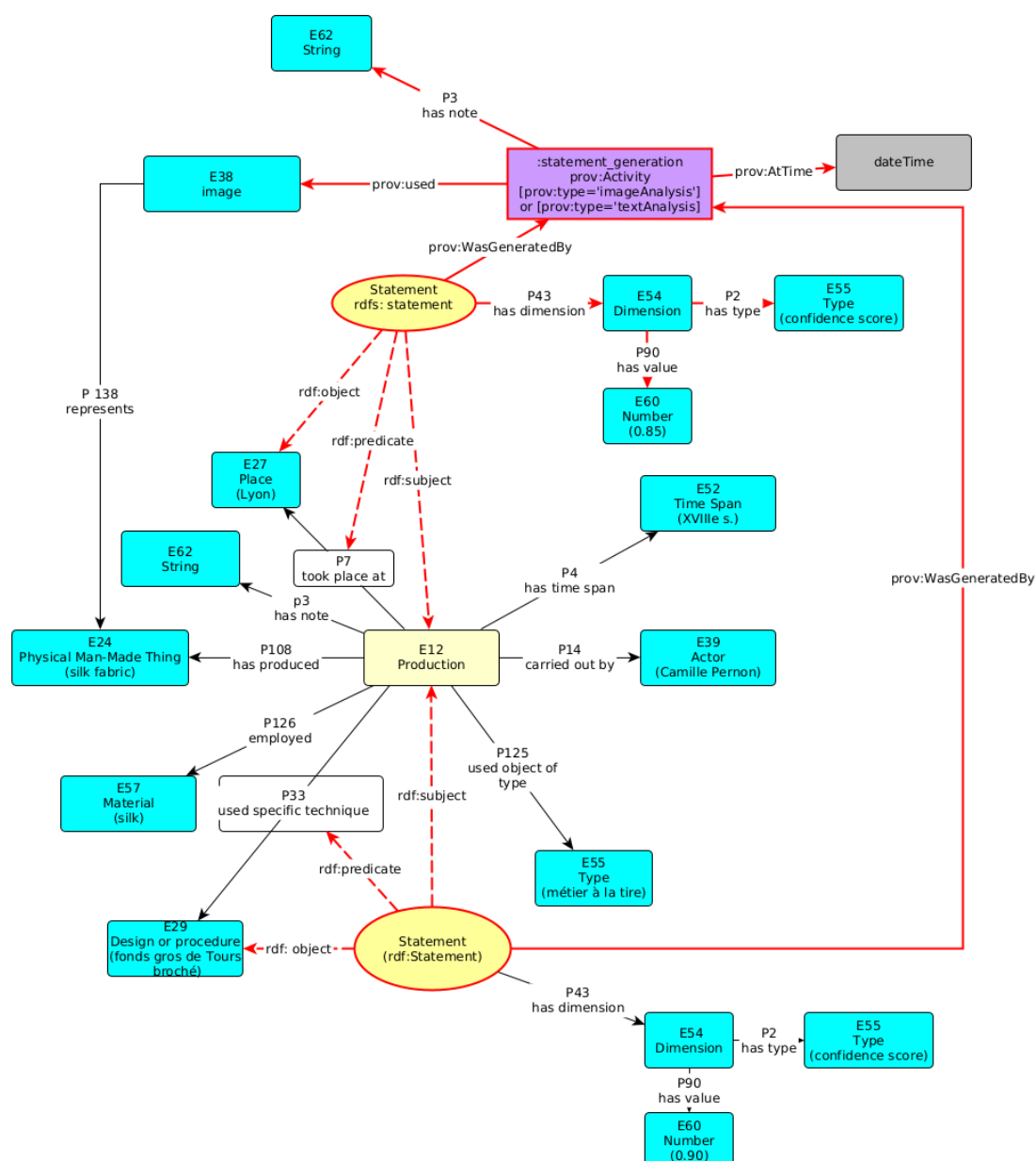


Figure 3. Integrating data from the image analysis module with the Provenance Data Model

The metadata describing the textile artifacts are also very rich; and this first mapping, aimed to store these metadata "as they were", cannot fully reflect this richness. In particular, we note the use of free text to analyse the structure and decoration of the fabrics, or to present the historical context of their production or even their use. During the first step of this work, this information has been stored as a "note" or using the CIDOC CRM and Scientific Observation Model: *S4_Observation P3_has note E62_String*. Table 2 provides an example from the collections of the Chiesa Madre di Caccamo (Sicily). Free text is used here to describe the complex construction of the patterned fabric:

Table 4. Information contained in the field “Costruzione” modeled with S4 Observation.

Fieldname	Content	Path
Costruzione	fondo in raso da 5, diffalcamento 2, faccia ordito, prodotto da tutti i fili e da tutte le trame di fondo. Opera	S4_Observation O8_observed E22_Man-Made Object

creata dal raso da 5, diffalcamento 3	S4_Observation P3_has note
faccia trama prodotto da tutti i fili e	E62_String
da tutte le trame di fondo,	S4_Observation P2_has type
unitamente a 2 trame braccate [...].	E55_Type (Costruzione)

The extraction of information from these textual data (cf. section 3.4) shows the extent to which these observations produce detailed technical analyses and new historical perspectives on these artifacts. It is thus possible to have access to information on description of patterns and weaves, weaving techniques, or styles. By choosing to model this information with a simple note, however, it is not possible to fully reflect the semantics of this information, nor to provide easy access to it. Indeed, users will not be able to formulate fine queries on this data, which nevertheless offers particularly interesting information.

Fortunately, CIDOC CRM is a very flexible and extensible model. This means that, if necessary, it is possible to create new classes and properties to express new types of information, without modifying the basic structure of the model. This allows the development of more specialised extensions - such as the Scientific Observation Model for example, or FRBRoo (Functional Requirements for Bibliographic Records) [28] for the process of creation, production and expression in literature and the performing arts. In line with these compatible models, we have therefore created a CRM extension designed to formally describe the process of creating and producing textile artifacts.

We have created 23 classes and 12 properties, accessible via Ontome (<http://ontome.dataforhistory.org/namespace/36>). We adopted a "bottom-up" approach, first of all based on the collected data. We also worked closely with domain experts and ICT experts to verify that these classes and the properties we proposed to create were useful and meaningful. For example, we created the class *T1_Weaving*, that is a subclass of *E12_Production*, to easily express how a *T7_Fabric*, that is a subclass of *E22_Man-Made Object*, was woven. We also created the class *T8_Part Weaving* to express the fact that the weaving process can include different but simultaneous actions - especially in the case of complex fabrics such as patterned fabric - using various techniques as well as several warps and wefts. We have then created classes and properties to accurately model this complex process, which often involves the use of several *T25_Weaving Technique*, and therefore various *T21_Weave*, and different *T17_Weft* and *T16_Warp*.

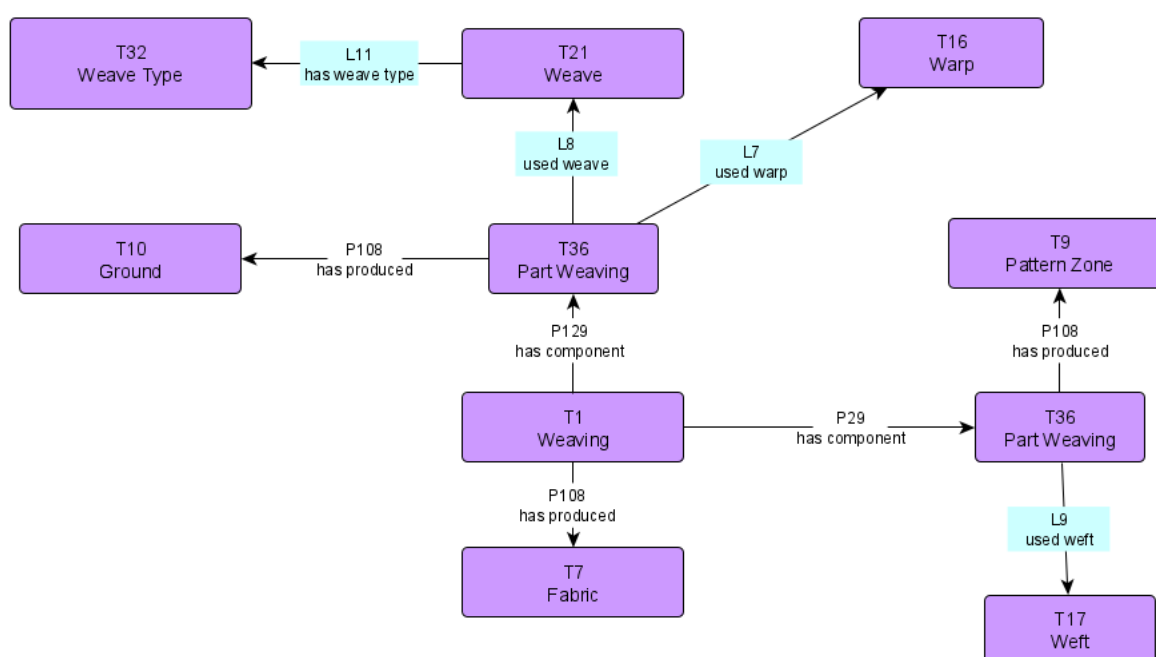


Figure 4. New classes and properties to model the weaving process.

Some of these classes also make it easier to create links between these data and the definitions provided by the SILKNOW thesaurus (cf. section 3.1). This thesaurus provides additional information that users can access directly from the data they are currently studying. Thanks to these classes, it is therefore possible to create links between the data, regardless of the language in which it is expressed, and the thesaurus, which not only enriches the user's experience, but also provides useful contextual information for a better understanding of the data itself. For example, the class *T32_Weave Type* makes it possible to create links between the types of weaves described in the technical analyses and the thesaurus.

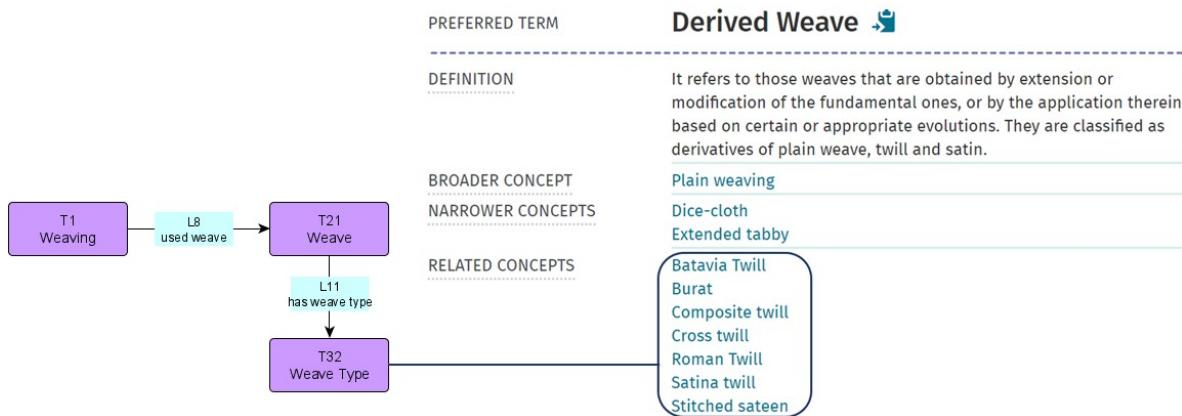


Figure 5. Examples of type of weaves (modelled as *T32_Weave Type*) defined in the thesaurus

3.3. Towards automated annotation through AI: text analysis.

A large number of culturally significant historical artefacts have been digitized and made available online. This means that experts in cultural heritage, and often the general public, now have the ability to search for and access information about artifacts instantly even when these are stored in distant parts of the world. However, in order to make the digitized information truly sustainable in the long run, certain challenges to exploring and accessing it remain, two of which we will address now. The first challenge is, obviously, language. For example, the European Union, which includes most of Europe, a continent with a very closely linked cultural history, has 24 official languages. The second challenge is the lack of standardized representation of knowledge across the different archives or catalogues. That is, the lack of a common ontology. Important data, such as the production technique or material used to create an object, is either not specified categorically, or when specified, it does not necessarily use the same term as in a different archive. This difference in terminology has a negative impact on the ability of an expert to find and understand related artefacts across archives. It also makes it harder for automated methods to provide useful features for the exploration of large groups of artifacts such as suggesting similar artifacts and providing visualizations of groups of artifacts along their properties. Most catalogues, for each digitized artifact, often have only a title, a short text description of the artefact, an image, and less often, an incomplete set of arbitrarily defined categorical descriptions in a non-standard terminology, although not all of these are necessarily present for different artifacts. Here, we use the available text, in whatever language it is written in, both title and short description, to infer categorical properties of the underlying object, through the use of supervised text classification. The properties we infer are intended to be useful for cultural heritage experts. These properties can be aligned with a specific ontology (such as our extension of CIDOC-CRM) or thesaurus which can be cross-lingual (such as our silk heritage thesaurus).

3.3.1. Data Description

A supervised text classification approach requires a labeled dataset. Our dataset was obtained by crawling online catalogues relevant to silk cultural heritage. These included the following: Victoria & Albert Museum, London (VAM); Boston Museum of Fine Arts (MFA); and the Red Digital de Colecciones de Museos de España (CERES) – a catalogue of multiple museums. Only pages containing information regarding silk fabric artefacts were retrieved. From these, the title, text description of the artefact, and any categorical fields present were extracted. These categorical fields were then normalized, that is, converted to a standard representation in English, defined by domain experts. The categories, their possible values, and the total number of samples for each can be seen in the following table.

Table 5. Summary of the data: the categories we attempt to infer, the list of their possible values given our data, and the total number of samples for each variable.

	Technique	Material Used	Production Place (country)	Production Date (century)
Values	brocading, embroidering, knitting, lace, printing, sewing, velvet, weaving	cotton, leather, linen, metal_thread, wool, printed, other	Africa, AT, AZ, BE, UK, CN, FR, DE, GR, IR, IT, JP, NL, PT, RU, ES, SY, TR, US, South Asia	10, 14, 15, 16, 17, 18, 19, 20
Number of Samples	3783	4058	8116	7765

3.3.2. Methodology

Our methodology to infer properties of a silk fabric from its short text description is based on a supervised text classification approach using a machine learning algorithm. This entails several steps which we will describe presently. We start by converting text into a normalized form and segmenting it into tokens which mostly correspond to words. Individual words are then converted, via a lookup table called an embedding layer, into (word) vectors also known as word embeddings. These vectors are learned, typically through co-occurrence, in a way which captures both semantic and syntactic properties of words. We use multilingual aligned word vectors, where vectors that represent words in one language are aligned with vectors that represent the same words in other languages. This means that to our learning algorithm, the same word in different languages will look similar (e.g., the English word “silk” will look similar to its Spanish translation, “seda”). In particular, we use the pre-trained multilingual aligned embeddings described in [29]. Finally, these vectors are fed into a classifier, a Convolutional Neural Network which outputs a predicted class value (1 class out of N possible predefined choices) via a softmax layer.

The architecture of our Convolutional Neural Network, shown in Figure 6, follows from previous work in applying CNNs to text [30, 31]. The word embeddings are concatenated and a predefined number of convolutional filters (feature maps) with different fixed window sizes (kernel sizes) are applied to each possible window of words to extract “features”. These are then passed through a non-linearity and a max-pooling operation. The idea is to capture the most important feature for each feature map. Pooling over time (1d max-pool) deals with variable text lengths - we used a fixed maximum of 300 word-tokens, determined from analysis of the data. After the pooling, the different features for each window are concatenated together, regularized by a dropout layer and put through a final fully connected output layer with a softmax activation to give a distribution of probabilities over the classes. The general intuition behind the algorithm is that each window of size $h = 2, 3, 4$ learns to extract something similar to word n -gram features where $n=h$. In this work, the sequence of operations consisting of Convolution, Activation, and Max Pool form a convolutional block. A single convolutional block handles a single window size. We use 3 blocks in parallel, corresponding to the window sizes $h = 2, 3, 4$. We use the Gaussian Error Linear Unit (GELU) [32] as our activation function and the Alpha Dropout [33] variant of dropout. The activation function, dropout variant, dropout probability, convolutional kernel sizes (window sizes), and the number of filters were all treated as hyper-parameters and selected through hyper-parameter tuning on a subset of the data. In all experiments, the network was trained for 300 epochs, using mini-batch stochastic gradient descent, with a batch size of 64, and an initial learning rate of 0.005.

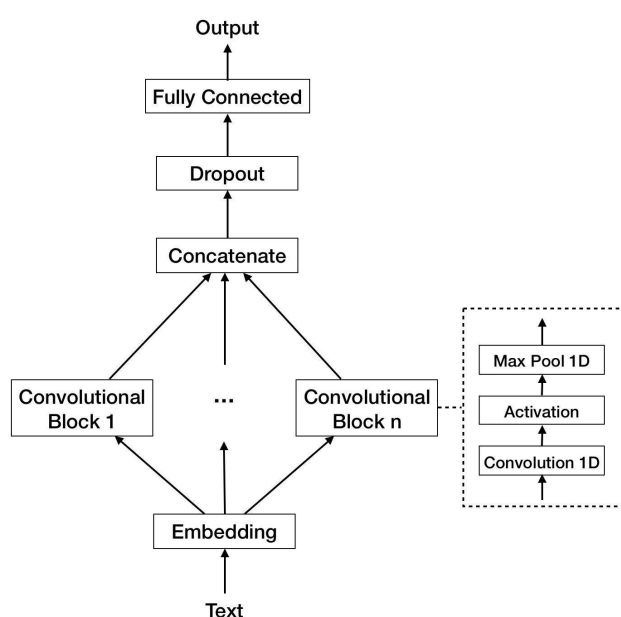


Figure 6. Convolutional Neural Network Architecture for text classification.

3.3.3. Experiments and Results

Our experiments focus on answering specific research questions. Our research questions are posed in terms of specific application scenarios:

1. Given labelled data (digitized artifacts) from one catalogue (e.g., a museum), can we infer those labels (properties) in non-labelled data in **the same catalogue**? The practical applications of this include the ability to infer properties in a catalogue from a subset of that catalogue's data which was semi-automatically or manually labelled, filling missing data, and semi-automatic conversion to a different ontology.
2. Given labelled data from one catalogue, can we infer the labels of non-labelled data **in a different catalogue**? Practical applications include aligning the ontologies of two or more different catalogues, and, if one can be labelled with a standard ontology then that effort can be leveraged to provide those categorical labels to other catalogues.
3. Given labelled data from one catalogue, can we infer the labels of non-labelled data **in a different language catalogue**? Applications are the same as in the previous case, but cross-lingual.

For each of these questions, we created a corresponding experimental evaluation scenario. For the first scenario, we use the data we collected from VAM and split it into a separate train and test sets (Scenario 1). This artificial split of the data was performed using random stratified splitting, a technique which randomly selects the examples in each set while preserving the distribution of the labels from the original set. 80% of the examples are used as training and 20% as test examples. In the second scenario, we used the VAM catalogue as the training set and the MFA collection as the test set (Scenario 2). In the third scenario, we again use VAM, which is in English, for training but we use CERES, in Spanish, as the test set (Scenario 3). Note that we use VAM as a training set in all scenarios purposefully to enable a better comparison between the results.

Our results given in Table 6 clearly show that it is possible to infer properties of silk fabrics from a short description of them, even across catalogues and across languages. The best results are obtained when labelled data from the same catalogue is used (Scenario 1). The biggest challenge faced by a text classification algorithm when dealing with short descriptions, in the context of digitized artifacts, is how different these texts are across catalogues, in both form (syntax and length), content (semantics), and in the objects they describe (e.g., museums are not random collections of objects but rather curated, often thematically and locally). We can clearly see the difference in results with regards to “production place”, and “production date” between Scenario 1 and 2. MFA descriptions rarely contain any words relevant to these two properties, while VAM often explicitly mentions regions, cities, and even countries with regards to one and dates with regards to the other. Thus, a text classifier trained on VAM descriptions is ill-prepared to handle MFA descriptions. MFA descriptions, though much shorter than VAM descriptions, usually do include techniques and materials, making the results in Scenario 2 for these properties much closer to Scenario 1. In the cross-lingual Scenario 3 we can see a further performance drop attributable, in part, to the difference in language. Descriptions in CERES, while focused primarily on depictions, often explicitly mention locations (e.g., cities) which helps explain the better-than-expected results for “production place”. City names are easier to align in pretrained embeddings than very domain-specific techniques and materials since these embeddings are primarily trained on Wikipedia and aligned through the use of dictionaries that are not domain specific. With regards to dates, when explicit, VAM tends to express them using Arabic numerals and ranges (e.g., “1740-1800”) while dates in CERES tend to use roman numerals (e.g., “siglo XIX”) which is responsible for part of the difference in accuracy.

Table 6. Evaluation results (accuracy) for the different scenarios.

	Technique	Material Used	Production Place	Production Date (century)
Scenario 1 (within museum)	97.6%	91.4%	97.4%	88.6%
Scenario 2 (across museums)	88.3%	77.7%	24.22%	48.2%
Scenario 3 (across museums and languages)	54.9%	59.8%	86.4%	20.7%

3.3.4. Text analysis: conclusions

We have shown that it is possible to infer, from a short text description of a silk fabric, properties relevant in the cultural heritage domain. We have also shown that this is possible in a cross-catalogue and cross-lingual setting. Applications of this development include, but are not limited to, machine-aided improvement of categorical digitized data within an archive, changing ontologies of categorical properties in a catalogue to align them with a different ontology or thesaurus, and helping a centralized resource (e.g., an open knowledge catalogue that includes data from multiple museums) homogenize digitized artifacts across its sources.

In this work we have provided a methodology for creating and evaluating a text classifier that can handle this challenge. The source code of the classifier is available online under an open-source license [34]. Several interesting avenues of future work remain open, especially along domain adaptation. For example, it would be interesting to have pretrained embeddings that are more tuned to the cultural heritage domain. The challenge to this lies in obtaining enough text to perform such adaptation. A second avenue would be the adaptation of the multilingual alignment to include the use of such data as well as domain-specific dictionaries and thesaurus.

3.4. Spatiotemporal visualization of maps.

The amount of openly accessible data continues to grow sharply. The proliferation of information published through institutional websites, journals and social networks generates huge amounts of data. The visualisation and analysis of this information has become an emerging field with extensive scientific activity [35]. For this reason, many old data visualization techniques have been redesigned while new ones have been developed, too [36].

An important case in point is the visualisation of spatiotemporal data [37]. Within it, the visualisation of cultural heritage information brings in additional complexity, due to the frequent uncertainty regarding both time and space of historical events [38].

The STMaps tool [39], designed and developed in the SILKNOW project, aims at visualising and analysing spatiotemporal data stored and represented in a knowledge graph. It allows the interactive visualisation of the data and the relationships between them, as well as their evolution over time. Using advanced visualisation techniques, it allows us to find unusual patterns and behaviours.

3.4.1. Implementation.

The Visualization Ontology (VISO) is used to configure the visualization aspect of STMaps tools [40]. VISO is a generic approach, mostly related to two-dimensional space visualization. It was extended in order to manage virtual reality concepts and the data visualization techniques used in STMaps.

The tool is implemented in Unity, a technology that allows to develop a cross-platform application with state-of-the-art graphics. It can be used by embedding a WebGL plugin into an HTML web page. The WebGL plugin technology is executable in most of the operative systems. Previously to the render process, the access to the domain knowledge graph must be defined, in addition to how to visualize the data and which data to visualize. STMaps has been released into a Github

repository [41], where the tool can be downloaded. Figure 7 shows a schema of a system embedding STMaps.

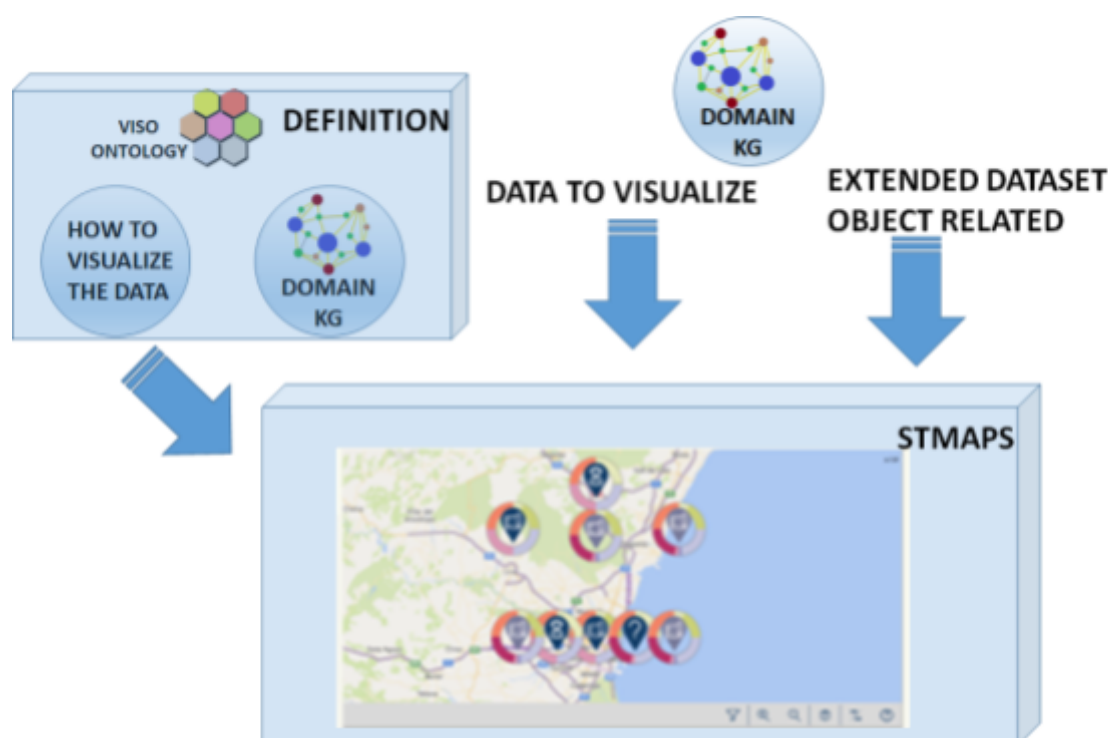


Figure 7. STMaps system integration schema.

3.4.2. Functionality

STMaps gets the adequate map images to represent the objects of the domain data from their built-in spatial coordinates. It also creates a quad tree-based representation, which splits the map into clusters and groups the data according to the zoom level. Thus, depending on the active user zoom level, the cluster points are depicted by a representative icon, or the single points are directly displayed. By zooming in, the clusters are divided into subclusters, until clusters are replaced by independent points. The cluster zoom levels and the icon aspect are determined in the configuration file. This clusterization is essential to keep the map readable without losing information.

The tools support the uncertainty related with space and time, a frequent problem in cultural heritage data. To deal with it, STMaps represents these objects with uncertain data by using special icons and also displaying data about alternative instances.

STMaps offers two possibilities to visualize the existing relationships between the displayed objects on the map. The first one is a classic, basic style, just connecting the related objects by coloured lines. By displaying a window with extended information related to a data point, the user may select a set of object properties. With this selection, the tool depicts the relationships from this object to the other objects with the same value in the selected properties.

The second way to display the relationships between the objects on the map is the outer ring, having segments filled with different colors. The size of these segments is proportional to the percentage of points with the same value for this given property of the object. For example, if the red segment covers the 25% of the ring it means that the 25% of the points has the same value in the property represented by the red color. This option is an easy, graphic way to detect objects with no relationships or a high number of relationships.

Both ways to show the relationships within STMaps are displayed in Figure 8.

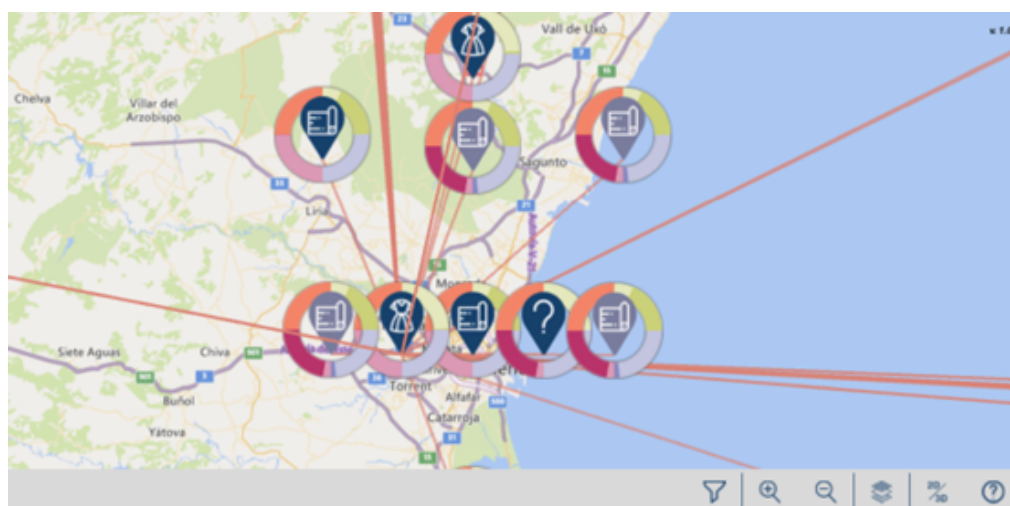


Figure 8. STMaps screenshot, showing the two possibilities to represent relationships between dataset objects.

STMaps also offers two ways to display the changes in data across time. The first is a classic timeline, showing a time interval with a time slider. The user can drag this slider in order to see the data status in a specific moment of time, according to the time resolution defined in the configuration file.

The second option is a time layer. With this functionality, the user may define a number of layers to visualize (from 2 to 4). Then, a time interval is associated with each layer and the application represents in each layer all the data related to its corresponding time interval. The layers and the data are displayed in a 3D environment. A user interface allows to adjust the desired layer for a better visualization. This second option incorporates simultaneity, as the user can simultaneously visualize different time steps data on screen. Figure 9 shows a screenshot of STMaps with the time layer functionality activated.

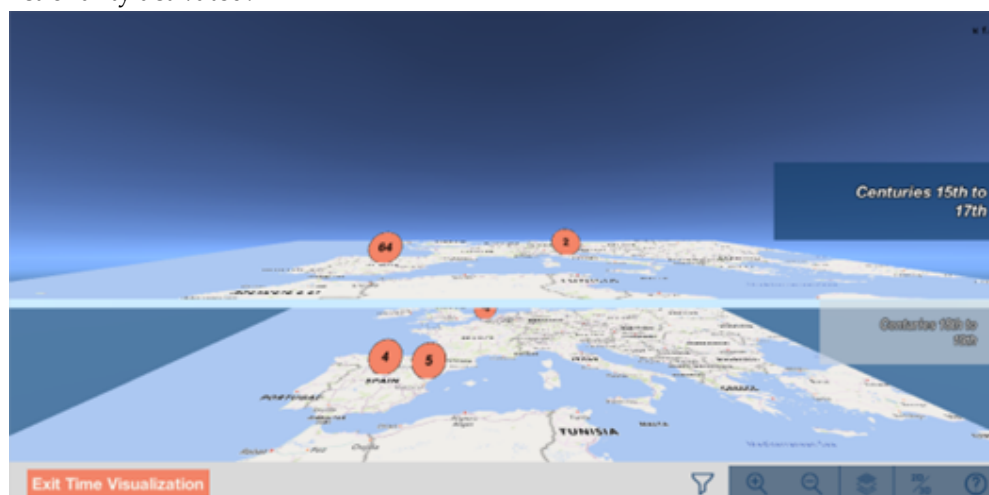


Figure 9. STMaps screenshot where the time layer visualization mode is active.

4. Heritage institutions need to focus on information sustainability and open access.

In conclusion, this case study shows the possibilities and demands that information sustainability places on heritage institutions, based on a collaborative project such as SILKNOW. Among the possibilities, interoperability does not only offer users the opportunity to discover data on

repositories shared across institutions, instead of having to knock on every museum's door (or website) to find the data. It also allows the application of algorithms to the massive amounts of data that these repositories contain (or can arrive to contain), facilitating cross-lingual access to the information, searching for unexpected patterns and matches among previously unrelated pieces, or suggesting automated annotations for poorly cataloged objects, based on the information within similar objects' records. Data visualization is a new field that enhances our understanding of those massive quantities of information, for instance thanks to spatiotemporal maps. Crucially, it ensures the long-term sustainability of heritage information, since the efforts to standardize it also entail improving its quality and ensuring its permanent availability.

The demands imposed by this process are also evident from the previous pages. It is important to have a sound knowledge of the existing information environment, in order to align with standards or platforms, and not just reinvent the wheel in every new effort. Terminology standardization (and its correct application) is a key issue, one that is, however, usually forgotten in favor of other, more glamorous tasks. Multilingual thesauri are cornerstones for any cataloguing effort that aims at producing information that can still be found and properly understood in the long term. Museum records are an almost untapped resource in our era, increasingly hungry for good-quality data. Nonetheless, bringing them into the semantic web is a complex task, one that involves their mapping to standards such as CIDOC-CRM, while also extending those standards, and paying attention to detail. Turning the promises of artificial intelligence into realities useful for the sector of cultural heritage requires collaboration with computer scientists, and a fair understanding of the opportunities and limitations of algorithmic tools.

An institutional commitment towards open access is both a previous requirement and a result, in enabling all these possibilities. Without that commitment, these endeavors lack data, the basic fuel that they need to develop. It is true that a more positive attitude towards open access is now frequent among museums, compared to previous years. However, many challenges remain in this area, since not everything depends on the mere will of decision makers. Technical and organizational challenges are still important, especially for small and medium-size institutions, and cannot be overlooked. However, as long as researchers and developers have open access to data, more and more studies will confirm that advantages surpass concerns by far, in this kind of collaboration.

Funding

Research leading to these results has taken place within the research project "SILKNOW. Silk heritage in the Knowledge Society: from punched cards to big data, deep learning and visual/tangible simulations", funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 769504.

Article authors listed by institution name.

References

1. Castells, M., *The rise of the network society*. 2nd ed. 2000, Oxford Blackwell Publishers.
2. *Words Mean Things (A Glossary)*, in *Open GLAM*, by Andrea Wallace. Available online: <https://openglam.pubpub.org/pub/the-glossary/release/1> (accessed on 3 July 2023) <https://doi.org/10.21428/74d826b1.51566976>
3. *Clarifying "Open"*, in *Open GLAM*, by Andrea Wallace. Available online: <https://openglam.pubpub.org/pub/clarifying-open/release/1> (accessed on 3 July 2023).

4. Pekel, J.; Nilsson, K. *Making impact on a small budget. How the LSH museet shared their collection with the world.* Europeana Pro 2015. Available online: https://pro.europeana.eu/files/Europeana_Professional/Publications/Making%20Impact%20on%20a%20Small%20Budget%20-%20LSH%20Case%20Study.pdf (accessed on 3 July 2023).
5. *Nouvelles notices versées sur Joconde.* Available online: <https://www.culture.gouv.fr/Thematiques/Musees/Les-musees-en-France/Les-collections-des-musees-de-France/Joconde-catalogue-collectif-des-collections-des-musees-de-France/Nouvelles-notices-versees-sur-Joconde> (accessed on 3 July 2023).
6. *Lettre d'information publiée par le bureau de la diffusion numérique des collections.* 25.11.2020. Available online: <https://www.culture.gouv.fr/Thematiques/Musees/Actualites/Nouvelle-lettre-d-information-du-bureau-de-la-diffusion-numerique-des-collections> (accessed on 3 July 2023).
7. D'Agnelli, F.M.; Rizzo, M.T. *Raccontare il patrimonio religioso: identità ed etica nella restituzione sul portale Beweb*, in *Nessuno poteva aprire il libro... Miscellanea di studi e testimonianze per i settant'anni di fr. Silvano Danieli – OSM*, M. Guerrini, Editor. 2019.
8. *Renovamos IMATEX.* Available online: https://cdmt.cat/es/renovem-imatex-10_05_2021/ (accessed on 3 July 2023).
9. Smeets, R. Language as a Vehicle of the Intangible Cultural Heritage. *Museum International* **2004** 56(1-2), 156-165.
10. Schreiber, G.; Amin, A.; Aroyo, L.; van Assem, M.; de Boer, V.; Hardman, L.; Wielinga, B. Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *Journal of Web Semantics* **2008** 6(4), 243-249. <https://doi.org/10.1016/j.websem.2008.08.001>
11. Anderson, C.A. Talking about textiles: the making of the Textile Museum Thesaurus. In *Textile Narratives & Conversions: Proceedings of the 10th Biennial Symposium of the Textile Society of America*, Toronto, Ontario, 2006 October 11-14.
12. Haffner, D. *A Textile Thesaurus – Merging and Enlarging the Existing Vocabularies*, presentation at ICOM General Conference 2016, Milan. Available online: <https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2018/12/haffner-textile-thesaurus.pdf> (accessed on 3 July 2023).
13. León Muñoz, A.; Gaitán Salvatella, M.; Sebastián Lozano, J.; Alba Pagán, E.; Insa, Isabel. SILKNOW. Designing a Thesaurus about Historical Silk for Small and Medium-Sized Textile Museums. In *Science and Digital Technology for Cultural Heritage-Interdisciplinary Approach to Diagnosis, Vulnerability, Risk Assessment and Graphic Information Models: Proceedings of the 4th International Congress Science and Technology for the Conservation of Cultural Heritage (TechnoHeritage 2019)*, Sevilla, Spain, March 26-30, 2019; CRC Press, 2019, p. 187.
14. Owens, L.A.; Cochrane, P.A. Thesaurus Evaluation. *Cataloging & Classification Quarterly* **2004** 37(3-4), 87-102. http://dx.doi.org/10.1300/J104v37n03_07
15. Isaac, A.; Zinn, C.; Matthezing, H.; Van de Meij, H.; Schlobach, S.; Wang, S. The Value of Usage Scenarios for Thesaurus Alignment in Cultural Heritage Context. In *Proceedings of the ISWC 2007 Workshop in Cultural Heritage on the Semantic Web*. <http://hdl.handle.net/11858/00-001M-0000-0013-19E2-A>.
16. Alba Pagán, E.; Gaitán Salvatella, M.; León Muñoz, A.; Mladenici, D.; Brank, J. Weaving words for textile museums: the development of the linked SILKNOW thesaurus. *Heritage Science* **2022** 10(1), 1-14. <https://doi.org/10.1186/s40494-022-00681-x>
17. Halevy, A. Why Your Data Won't Mix: New tools and techniques can help ease the pain of reconciling schemas. *Queue* **2005** 3(8), 50-58. <https://doi.org/10.1145/1103822.1103836>
18. Guarino, N., Understanding, building and using ontologies. *International Journal of Human-Computer Studies* **1997** 46(2), 293-310. <https://doi.org/10.1006/ijhc.1996.0091>
19. Arrêté du 25 mai 2004 fixant les normes techniques relatives à la tenue de l'inventaire, du registre des biens déposés dans un musée de France et au récolement, *Journal officiel "Lois et Décrets"* **2004** 0135.
20. Briatte, K., HADOC Modèle harmonisé pour la production des données culturelles. 2012: Ministère de la Culture et de la Communication. Available online: https://www.culture.gouv.fr/Media/Documentation/Harmonisation-des-donnees-culturelles/Files/MCC-HADOC-REF-modele_harmonise_donnees_culturelles.pdf (accessed on 3 July 2023).
21. *International Guidelines for Museum Object Information: The CIDOC Information Categories.* International Committee for Documentation of the International Council of Museums, 1995.
22. *Europeana Data Model.* Available online: <https://pro.europeana.eu/page/edm-documentation> (accessed on 3 July 2023).
23. Kondylakis, H.; Doerr, M.; Plexousakis, D. *Mapping Language for Information Integration.* 2006. Available online: http://www.cidoc-crm.org/sites/default/files/Mapping_TR385_December06.pdf (accessed on 3 July 2023).

24. Doerr, M. *Mapping a Data Structure to the CIDOC Conceptual Reference Model*. 2002, Heraklion: ICS-FORTH. Available online: https://www.cidoc-crm.org/sites/default/files/Mapping_7_4_2003_0.ppt (accessed on 3 July 2023).
25. Beretta, F. OntoME, Ontology management environment, in *2nd Data for History workshop*. 2018: Lyon.
26. *Definition of the CRMsci. An Extension of CIDOC-CRM to support scientific observation, Version 1.2.8*. 2020, CIDOC CRM-SIG. Available online: <https://cidoc-crm.org/crmsci/ModelVersion/version-1.2.8> (accessed on 3 July 2023).
27. *The PROV Data Model, W3C Recommendation*. 30 April 2013. Available online: <https://www.w3.org/TR/prov-dm/> (accessed on 3 July 2023).
28. *FRBR object-oriented definition and mapping from FRBRER, FRAD and FRSAD. Version 3.0*. 2017, International Working Group on FRBR and CIDOC CRM Harmonisation. Available online: <https://www.cidoc-crm.org/frbroo/ModelVersion/frbroo-v-3.0> (accessed on 3 July 2023).
29. Joulin, A. *Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion*. arXiv:1804.07745 [cs], 2018. <https://doi.org/10.48550/arXiv.1804.07745>
30. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. P. *Natural Language Processing (Almost) from Scratch*. *J. Mach. Learn. Res.* 2011, 12, 2493–2537. <https://doi.org/10.5555/1953048.2078186>.
31. Kim, Y. *Convolutional Neural Networks for Sentence Classification*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, ACL, 2014; pp 1746–1751. <https://doi.org/10.3115/v1/d14-1181>.
32. Hendrycks, D.; Gimpel, K. *Gaussian Error Linear Units (GELUs)*. arXiv:1606.08415 [cs], 2020. <https://doi.org/10.48550/arXiv.1606.08415>
33. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*.
34. SILKNOW Text Classification Code Link. SILKNOW Consortium. (2021). Available online: <https://github.com/silknow/text-classification> (accessed on 4 July 2023).
35. Qin, X.; Luo, Y.; Tang, N. Making data visualization more efficient and effective: a survey. *The VLDB Journal* 2020 29, 93–117. <https://doi.org/10.1007/s00778-019-00588-3>
36. Wang, J.; Hazarika, S.; Li, C.; Shen, H. Visualization and Visual Analysis of Ensemble Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 2019 25 (9) 2853–2872, 1 Sept. 2019, <https://doi.org/10.1109/TVCG.2018.2853721>
37. Alam, M.M.; Torgo, L.; Bifet, A. A Survey on Spatio-temporal Data Analytics Systems. *ACM Comput. Surv.* 2022 54, 10s, 219 (January). <https://doi.org/10.1145/3507904>
38. Windhager, F.; Filipov, V.; Salisu, S.; Mayr, E. Visualizing uncertainty in cultural heritage collections. In *EuroRVV '18: Proceedings of the EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization June 2018, 7–11*. <http://dx.doi.org/10.2312/eurorv3.20181142>
39. Sevilla, J.; Casanova-Salas, P.; Casas-Yrurzum, S.; Portalés, C. Multi-Purpose Ontology-Based Visualization of Spatio-Temporal Data: A Case Study on Silk Heritage. *Applied Sciences* 2021 11(4) 1636. <https://doi.org/10.3390/app11041636>
40. Polowinski, J.; Voigt, M. VISO: a shared, formal knowledge base as a foundation for semi-automatic infovis systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. Association for Computing Machinery, New York, 2013, 1791–1796. <https://doi.org/10.1145/2468356.2468677>
41. STMAPS Github Repository Link. SILKNOW Consortium. (2021). Available online: <https://github.com/silknow/spatio-temporal-map> (accessed on 3 July 2023).

