Enriching a Small Artwork Collection Through Semantic Linking

Mauro Dragoni[1(✉)], Elena Cabrio[2], Sara Tonelli[1], and Serena Villata[3]

[1] FBK, Trento, Italy
{dragoni,satonelli}@fbk.eu
[2] University of Nice Sophia Antipolis, Nice, France
elena.cabrio@unice.fr
[3] CNRS, I3S Laboratory, Sophia Antipolis, France
villata@i3s.unice.fr

Abstract. Cultural heritage institutions have recently started to explore the added value of sharing their data, opening to initiatives that are using the Linked Open Data cloud to integrate and enrich metadata of their cultural heritage collections. However, each museum and each collection shows peculiarities, which make it difficult to generalize this process and offer one-size-fits-all solutions. In this paper, we report on the integration, enrichment and interlinking activities of metadata from a small collection of verbo-visual artworks in the context of the Verbo-Visual-Virtual project. We investigate how to exploit Semantic Web technologies and languages combined with natural language processing methods to transform and boost the access to documents providing cultural information, i.e., artist descriptions, collection notices, information about technique. We also discuss the open challenges raised by working with a small collection including little-known artists and information gaps, for which additional data can be hardly retrieved from the Web.

1 Introduction

In the last years, cultural heritage institutions have been involved in several initiatives in order to exploit digital means to increase their visibility. Galleries, libraries, archives and museums (GLAM) typically own rich and structured datasets developed over many years and organized by domain, which in principle could be easily connected with the databases of other institutions and then made available online to a larger audience. However, several issues need to be faced, for instance the need for a standard format for data sharing, and the lack of technical skills especially in small museums, so that data manipulation and conversion can hardly be achieved. Relevant standardization efforts such as that of Europeana[1], a framework at European level to publish and link cultural heritage metadata through a unified data model, go in this direction and contribute

This work has been partially carried out in the framework of the VVV Project supported by Fondazione Cassa di Risparmio di Trento e Rovereto.

[1] http://www.europeana.eu.

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 724–740, 2016.
DOI: 10.1007/978-3-319-34129-3_44

to raise museums' awareness on the importance of knowledge sharing among cultural heritage institutions. The advantages include driving users to new content, stimulating collaboration in the cultural heritage domain, enabling new scholarship through the availability of new digital content, and more generally increasing the relevance of cultural heritage institutions [1].

In this work, we present the process performed to map the metadata from the Verbo-Visual-Virtual Project [2] to the Linked Open Data[2] (LOD) cloud and the related data enrichment. Although the work was largely inspired by past efforts by other cultural heritage institutions [3–5], we face new challenges, partly related to the small size of the collection, with little-known artists and few information available from other online sources, and partly to the integration of Natural Language Processing (NLP) techniques to enrich the metadata. On the one hand, we show that linking metadata to DBpedia[3] contributes to improving the quality and richness of the data owned by the museum. On the other hand, that small collections with little-known artists present specific issues, e.g. the limited coverage of external resources, that need to be addressed in a semi-automatic way. We make available both the developed ontology and the RDF data set containing the enriched metadata of our collection.

The remainder of the paper is as follows. Section 2 describes the Verbo-Visual-Virtual cultural heritage collection, Sect. 3 presents the `VVV` ontology we have defined to represent our cultural heritage data, and details about the data interlinking and enrichment steps we have addressed to enrich the available information. The approach is evaluated in Sect. 4 to prove its feasibility, and the encountered difficulties are discussed. Conclusions end the paper.

2 The Verbo-Visual-Virtual Project

The Verbo-Visual-Virtual Project (VVV) started in 2013 as a joint effort between two museums in Trentino-Alto Adige and a technological partner, with the goal to create a unified virtual collection of "Archivio di Nuova Scrittura" (ANS) [6]. The collection, albeit international, is mainly centered around the artworks of Italian artists active between 1950 and 1990, and finds its origin in the collecting activity of Paolo Della Grazia, an entrepreneur with a passion for interdisciplinary forms between art and poetry. Towards the end of the nineties, Della Grazia decided to donate to a public institution the archive, which had been steadily growing and needed an appropriate site. He decided first to contact MUSEION[4], the Museum of Contemporary Art in Bolzano. However, since the museum did not have enough space available, Della Grazia decided to split the collection into two parts, one assigned to MUSEION through a long-term loan, and the other to MART[5], the Museum of Modern and Contemporary Art in Rovereto. For this reason, a collection which was originally conceived as a single archive is now

[2] http://lod-cloud.net/.
[3] http://www.dbpedia.org/.
[4] http://www.museion.it.
[5] http://www.mart.trento.it.

divided in two and hosted by two different institutions. ANS fulcrum is represented by works linked to concrete poetry, visual poetry, Fluxus, and conceptual art. The main features of the collection are its small size (around 5,000 artworks in total) and its homogeneity, i.e. all works were created in a limited time span by a relatively small number of artists.

Given the possibility offered by current digital technologies to access art collections online, VVV was launched to create a unified virtual collection of ANS works, where all information about the collection would be semantically enriched and made available through a web-based platform. The Digital Humanities group at Fondazione Bruno Kessler has therefore worked in the last two years to make all the work records consistent, possibly add information automatically retrieved from the Web, and implement a navigation platform to display and search works in a virtual exhibition. The project will end in Spring 2016, when the VVV platform will be made accessible.

2.1 Project Challenges

The VVV project brought about several challenges to address:

1. The *quality* of information available in databases describing small collections exposed in local museums is a critical aspect, especially when archives have been created incrementally and at different time points. Information could be inconsistent even for important elements such as the artwork title or author. Here, experts have to be supported in the management of such items by providing facilities to interlink information about collection items with as many external data sources as possible, to improve the qualitative description of each item. In VVV, this is particularly relevant to guarantee that the two portions of the archive are curated with the same quality.
2. Besides quality, also the *quantity* of available information is an important issue. Local authors are a good example scenario. Generally, it is not easy to find information about them on the Web, and, when it is available, it should be manually extracted for enriching the related content in the knowledge base. Addressing this challenge implies the design and implementation of information extraction approaches supporting experts in the enrichment of artworks information.
3. The exposure of information in semantic formats requires artworks to be classified through a classification schema. For this reason, the use of ontologies in the cultural heritage domain has gained a lot of attention in the last years. However, well-known classification schemas (for example the Europeana Data Model[6]) may be too generic to capture the peculiarities of minor collections, e.g., the local museum in a village in which a specific battle of World War I took place. In this case, a *conceptual modeling* activity is needed to tailor existing models to the specific needs of a collection.

[6] http://www.europeana.eu/.

4. Finally, disseminating the cultural heritage of local environments is effective only when information is published in languages currently spoken in the territory. In Trentino-Alto Adige, where the VVV use cases are located, Italian and German are the official languages, while English is widely used for tourism purposes. For this reason, it is crucial that artworks preserved in local museums are described at least in these three languages. When necessary, a *translation* task has to be performed for breaking the language barriers and yielding an effective publication of information.

2.2 Data Description

As a first step in this direction, in this paper we investigate the possibility to perform automatic enrichment of data through semantic interlinking and information extraction from the Web, based on the data split belonging to MART. We focus first on this part of the collection because the information stored at MART is more stable, while the records at MUSEION are still being updated. Nevertheless, the framework introduced in this paper (Sect. 3) is designed to support multiple sources and to fuse them through a domain ontology. MART has adopted a well-known record management system called *MuseumPlus*[7], that is used by the museum personnel to fill information about the artworks and curate them, as a knowledge base of the museum objects. Since the system is used by several people, and required information is mostly filled as free text, some inconsistencies are present in the data, especially spelling errors.

In order to perform semantic enrichment, we first export the VVV database stored at MART as raw data in CSV format. Since *MuseumPlus* offers the possibility to input a variety of information about each artwork and each artist, which however are not present for all entries of the VVV collection, we export only a subset of fields, which domain experts consider mandatory to identify a work of art, that is:

```
title,inventary_code, dimension, date, technique,
author_name, author_surname, author_born_place,
author_born_year
```

Overall the raw dataset contains 592 artworks created by 287 artists. However, prints are often collected in portfolios, which are also recorded as artworks. If we merge the works that belong to the same portfolio, we obtain 495 entries. Another issue is related to vague information in the fields: 187 works have no title ("Senza titolo"), not because of missing information but due to the artist's choice. Besides, 27 works have been created by an unknown artist.

3 The Framework

In this section, we describe the steps carried out to transform VVV raw data into the semantically enriched VVV data set. Figure 1 shows the proposed framework.

[7] http://www.zetcom.com/en/products/museumplus/.

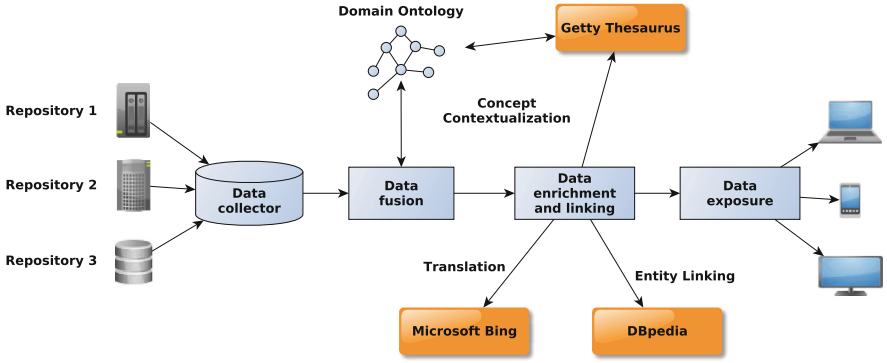

Fig. 1. The platform developed in the context of the VVV project.

Data can be collected from diverse repositories connected to the platform. As mentioned earlier among the challenges of the project, raw data are affected by problems like incomplete records and missing information about authors. Interoperability is promoted by the fusion and annotation of input data with semantic information modeled through an ontology, described in Sect. 3.1. After the merging and annotation process of all collected information, each record is enriched in two different ways:

1. Recognized entities are linked with information available in the Linked Open Data (LOD) cloud. More precisely, in the current version of the system only DBpedia[8] is exploited for interlinking (Sect. 3.2).
2. Natural Language Processing methods are used for extracting information from web pages containing relevant details for the record that needs to be enriched. For example, if for an author only the name is available, extraction patterns are applied for collecting information such as the birth place, the birth date, and so on (Sect. 3.3).

Finally, the last component of the pipeline consists in exposing the created knowledge base, by using LOD formats, to make it available to third-party services. Data are exposed through a RESTful interface providing requested information in Turtle format[9].

Such a pipeline, specifically designed and implemented for the VVV project, is generally applicable to any process whose goal is to convert raw cultural heritage data into an enriched version. Each task of the pipeline is run when a new raw dataset is imported. Given that the final purpose of the process is to expose an enriched dataset, and the fact that collected data are automatically translated and linked by the different components of the pipeline, we provide a tool that implements facilities enabling the manual verification and refinement of all information. More precisely, the described pipeline has been implemented

[8] http://wiki.dbpedia.org/.
[9] http://www.w3.org/TR/turtle/.

in a collaborative knowledge management tool called MoKi [7,8]. MoKi[10] is a collaborative MediaWiki-based[11] tool for modeling ontological and procedural knowledge in an integrated manner[12]. MoKi is grounded on three main pillars, which we briefly illustrate: *(i)* each basic entity of the ontology, i.e., concepts, object and datatype properties, and individuals, is associated with a wiki page; *(ii)* each wiki page describes an entity by means of both unstructured, e.g., free text, images, and structured, e.g., OWL axioms, content; and *(iii)* a multi-mode access to the page contents is provided to support easy usage by users with different skills and competencies. In the VVV project, the tool has been customized for supporting the manual refinement activity performed by experts after the automatic execution of the entire pipeline[13].

3.1 The VVV Ontology

In the cultural heritage domain, a number of ontologies has been proposed to represent the semantics of cultural heritage data. The most known ontologies proposed in this context are CIDOC-CRM[14] and the Europeana Data Model (EDM)[15].

CIDOC Conceptual Reference Model (CIDOC-CRM) claims to be a "formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" [9]. It is developed by International Counsel of Museums (ICOM), and is accepted as an ISO Standard. CIDOC-CRM is one of the most recommended models for cultural heritage. It combines knowledge about artworks together with all the events concurring to its creation. CIDOC-CRM is defined with the aim to facilitate the identification and sharing of knowledge about cultural heritage data, and the interoperability among the different sources of cultural heritage data and their own data representation models. This ontology mainly defines general concepts, allowing for ontology extensions with the concepts and properties needed by each source.

EDM is a data model that has been defined by a set of European museums in the Europeana project. The aim of Europeana is to build a computational library including the cultural heritages from various European museums. It allows to access to the collections of galleries, museums and libraries of all types (including images and audiovisual resources). The British Library in London, the Rijksmuseum in Amsterdam and the Louvre Museum in Paris are among the 1500 institutions that have participated in the construction of this cultural library. The EDM data model is constructed based on the CIDOC-CRM model, i.e., it inherits some concepts and properties of CIDOC-CRM. This reuse of properties and

[10] http://moki.fbk.eu.
[11] http://www.mediawiki.org.
[12] Though MoKi allows to model both ontological and procedural knowledge, here we will limit our description only to the features for building ontologies.
[13] A read-only version, but with all functionalities available, of the MoKi instance described in this paper is available at https://dkmtools.fbk.eu/moki/3_5/vvv/.
[14] http://www.cidoc-crm.org.
[15] http://pro.europeana.eu/edm-documentation.

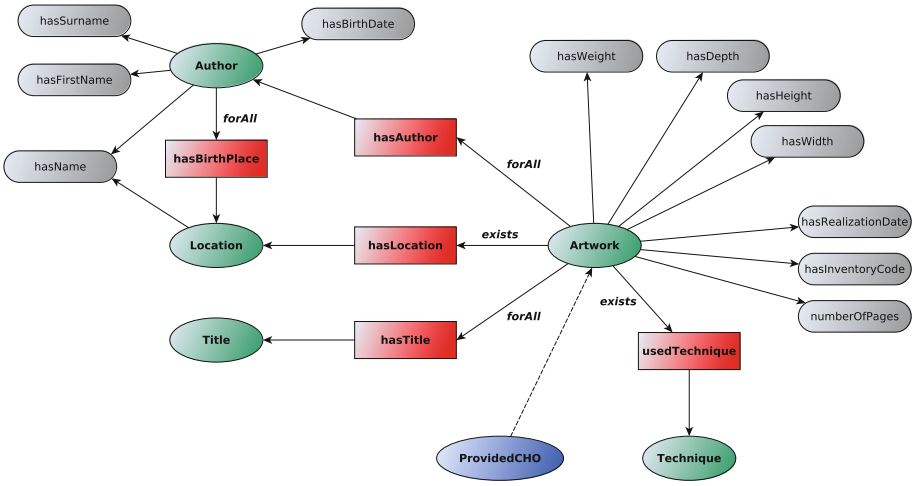

Fig. 2. The ontology of the VVV project.

concepts from CIDOC-CRM supports interoperability with other data sources represented through the CIDOC-CRM ontology.

In analyzing these two models with the purpose of selecting an existing ontology (if any) that suits the requirements of the VVV project, we choose to take EDM as the basis of our ontology[16], so as to reuse EDM general concepts and properties and further extend it with the additional information we need to represent the VVV data semantics in a compliant way. More precisely, our data contain a set of mandatory fields reported in Sect. 2. Moreover, we need to address the challenge of managing several missing data about the items in local museums. For these reasons, we design a new ontology, called `VVV`, reflecting a specific model that fits well our project needs. As illustrated in Fig. 2, our ontology extends the EDM ontology with new concepts and properties.

The `VVV` ontology reuses two concepts from the EDM ontology. Such concepts have been redefined in terms of label for accommodating the schema readability by the domain experts. The `Author` concept is used for representing artists, and it is aligned with the concept "Agent" contained in the EDM, while the concept `Location` is used for representing either the places where an artwork is exposed or the birth place of an author. This concept is aligned with the concept "Place" contained in the EDM.

In addition, the `VVV` ontology introduces the following concepts: *(i)* `Artwork`, defined as a child of the concept `ProvidedCHO` modeled in the EDM ontology, is used for instantiating artworks in the knowledge base; *(ii)* `Technique` represents the style adopted for realizing a certain art work; *(iii)* `Title` defines the caption of an artwork.

[16] CIDOC-CRM model is not appropriate to represent the granularity of the information we are interested in.

The definition of the concept `Title`, instead of adopting a simpler annotation property, is required since such information may be missing in the knowledge base. Indeed, our intention is to have a conceptual modeling of the missing title of an artwork, and not only mark the absence of an annotation. This fact is modeled through the use of the `hasTitle` object property. The same situation occurs, always for the concept `Artwork`, when data about its author are missing. Thus, the concept `Author` is linked to the concept `Artwork` through the `hasAuthor` object property. Both object properties have been modeled using the universal quantifier.

Moreover, among the object properties, also the `usedTechnique`, `hasLocation`, and `hasBirthPlace` object properties are defined. Such properties allow to model the artistic technique adopted for realizing a certain artwork, the location is which the artwork is exhibit, and the birth place of an author, respectively.

Finally, the ontological model is completed with the definition of the following annotation properties:

- `hasWeight`, artwork weight;
- `hasDepth`, artwork depth;
- `hasWidth`, artwork width;
- `hasHeight`, artwork height;
- `hasRealizationDate`, the date when an artwork has been realized;
- `hasInventoryCode`, the inventory code of an artwork in the exposition where it is preserved;
- `numberOfPages`, in case of literary artworks, the number of pages;
- `hasName`, the full name of an author or of a location;
- `hasSurname`, the surname of an author;
- `hasFirstName`, the first name of an author;
- `hasBirthDate`, the birth date of an author.

The described ontology is automatically populated while importing raw data from the external databases. Information is initially available only in the Italian language. One of the challenges of the project is to provide the ontology in two additional languages, i.e. English and German. Therefore, the tool has been equipped with a component connecting MoKi with the Microsoft Bing Translation service[17]. Through the interface, the experts are able to correct the proposed translations with most appropriate ones if needed.

3.2 Data Interlinking

In this section, we describe how we address the interlinking of our dataset with external knowledge bases in order to retrieve further information to contextualize our data. The entire interlinking activity is performed in two steps:

[17] http://www.bing.com/translator.

1. when a new dataset is imported into the pipeline, an automatic procedure is performed to retrieve the candidate links from external resources;
2. considering that all data provided through the platform should contain high quality information, a further manual refinement is carried out by experts to validate the linked information.

In the VVV project, two interlinking activities have to be performed: *(i)* to contextualize the technique used for creating an artwork with respect to a domain-specific thesaurus, and *(ii)* to interlink the entities, like authors and locations, with conceptual nodes defined on external knowledge bases such as DBpedia.

Linking with the Getty Thesaurus. As already mentioned in Sect. 2.1, it is very important to translate the concept labels as a preliminary activity to the interlinking step. Since in our case all information in the database is available in Italian only, all records are first automatically translated into English through Microsoft Bing Translation service. After that, the *Getty* thesaurus is used to map modeled concepts with information on the techniques contained in our dataset. Such thesaurus is currently one of the most widely used linguistic resources in cultural heritage projects, with high-quality, manually curated domain information.

The mapping operation consists in querying the Getty web service[18] with the English label describing the used technique. Each query may return more than one result. Thus, we decided to show to the experts a selection of the top five candidate concepts, based on the confidence score provided by the web service, that can be used for defining a new mapping. Such a strategy has been already applied in [10] demonstrating its suitability for this kind of scenario. Once results are shown to experts, they are able to select which concept, from the Getty thesaurus, to map with the ones defined in our platform. In this way, an imported dataset can be semantically connected with external knowledge bases in a semi-automatic fashion.

Linking with DBpedia. A similar approach has been adopted for linking information about authors and locations with DBpedia. This activity has been performed by exploiting the lookup service connected with DBpedia[19]. Such a service works with a REST interface receiving a query containing, for instance, a named entity label, and by returning a ordered rank of candidate DBpedia nodes that can be linked with the source label. Similarly to the linking with the Getty thesaurus, also in this case we provided five candidate suggestions to the domain experts, and let them choose the final alignment.

[18] http://vocab.getty.edu/.
[19] https://github.com/dbpedia/lookup.

3.3 Data Enrichment

Given that some of the artists of the VVV collection are little-known, they may be missing in DBpedia (and therefore, it is not possible to connect them to such a resource). For this reason, in order to provide additional information also for these entities to enrich our knowledge base, we resort to a semi-automatic process combining manual selection of texts and Natural Language Processing techniques. The methodology comprises the following steps:

1. *Manual search of online information about VVV artists:* The entire enrichment process cannot be performed completely in an automatic way due to the high risk of retrieving wrong information from the Web. Thus, we first search the Web for extracting textual snippets or paragraphs describing artists (in particular their biography or career), from sources that may contain relevant and correct information about artists (for instance, art collections websites). Texts are collected independently by the language used to write them. Indeed, most of the documents about the artists contained in the VVV collection, are expressed in the language of the artist birth country. For example, the biography of the artist "Massimo Pompeo", can be found at http://www.zam.it/biografia_Massimo_Pompeo only in Italian.
2. *Automatic content translation:* For the reasons expressed above, each text written in a language other than English is automatically translated into English by using the Microsoft Bing Translation service.
3. *Part-of-speech tagging:* Information contained in the pages retrieved online is typically unstructured and presented in natural language. For enabling the detection of potential significant relations, text is preprocessed and tokens are tagged with the TreeTagger [11] library on the English translation of the text.
4. *Relation extraction:* We apply a standard method for the automatic acquisition of relations based on hand-crafted rules [12]. In particular, we define a set of basic patterns (Table 1) expressed as regular expressions, and retrieve in the text the matching strings. These are assigned to the extracted relations with the corresponding properties.

MoKi supports the enrichment process with facilities (i) for inserting new rules to be exploited for extracting significant relations from text, and (ii) for selecting, among the suggested relations, the ones that can be used for enriching the ontology. Figure 3 shows the interface used by the domain expert for analyzing the list of the extracted relations and for choosing which ones can be used for enriching the ontology. In the upper part of the picture, we may see how the expert is able to invoke the relation extraction component. In the middle part, we may see how suggestions are displayed to the experts; while, in the bottom part, how the selected relations are shown in the entity mask.

Table 1. Hand written rules for relation extraction.

Example patterns	Relation	Property	Examples of matched sentence
born in DATE; (LOC, DATE; born on DATE)	birthdate	`rdaa:dateOfBirth`	He was born in Brno in 1946
died in LOC	place of death	`dbpedia:deathPlace`	He died in Paris
author of .*	author of	`vvv:authorOf`	He is author of numerous children's and other radio plays
influenced by .*	influenced by	`dbpedia:influencedBy`	The first works were influenced by the metaphysical painting

Fig. 3. The enrichment mask implemented in MoKi.

4 Evaluation

In this section, we present the evaluation of the platform proposed in the VVV-project. Such an evaluation has been conducted from two perspectives:

– Quantitative evaluation (Sect. 4.1): we report on the numeric results about the effectiveness of the data interlinking and the data enrichment algorithms. These values allow to measure how much support is provided to experts by the automatic modules in the pipeline.
– Lessons learned (Sect. 4.2): we discuss the lessons learned from the experience in applying Semantic Web technologies to the use case of data enrichment and integration of small cultural heritage collections. We include some suggestions provided by two art curators, expert of the VVV collection, during a demo session.

4.1 Quantitative Evaluation

The quantitative evaluation step includes the measure of the effectiveness of the DBpedia lookup service for covering and linking the modeled knowledge base, the accuracy of the enrichment information provided by the NLP component in charge of analyzing external textual resources, and the precision of the candidate alignments with the Getty thesaurus suggested by the system.

The raw data from the VVV collection contain works from 287 artists and for only 139 of them a DBpedia page can be linked. The interlinking algorithm we implemented has been executed automatically and we manually verified the correctness of the created links. For none of the 148 artists with no DBpedia page, links have been created. For the 139 artists that have a DBpedia page, 26 artists have been linked to the wrong page (i.e. the links pointed to entities that are homonyms of the considered artist). Thus, 113 VVV entities were correctly linked to DBpedia by obtaining an accuracy of 0.812.

Concerning data enrichment, we applied the proposed approach to the 148 artists with no DBpedia page and for 93 of them we were able to find online a description of the author biography and artistic career. Hand-written rules have been applied for extracting the following relations: “place of birth”, “date of birth”, “date of death”, “author of”, “influenced by”, “description”. As mentioned in Sect. 3.3, the set of rules can be changed dynamically by adding further rules or by refining the existing ones. To evaluate the accuracy of the relations suggested by the extraction algorithm, the texts have been manually annotated to build a gold standard. Results obtained for relation extraction of the six above mentioned properties are: *Precision: 0.93, Recall: 0.89, and F-measure: 0.90*. Finally, concerning the alignment between our ontology and the Getty thesaurus, the candidate mappings service implemented in our platform obtained an accuracy of 0.976.

From these results, we can infer a couple of practical lessons useful for possible future inclusion of further collections. First, when dealing with little-known artists, a methodology based on metadata enrichment by linking artists’ names to DBpedia is not satisfactory. Artists’ contextual information are needed for improving the disambiguation capability of the linking approach. Secondly, the adoption of a semi-automatic enrichment approach based on (i) a manual retrieval of textual snippets from the Web and (ii) the analysis of such snippets

for suggesting further information about artists, is recommended to ensure a good balance between curation time and quality.

4.2 Lessons Learned and Future Work

Here, we sum up our experience in using Semantic Web technologies for supporting the management and the enrichment of a little-known collection in the context of the VVV project. As already mentioned in this paper, the main issue when working with little-known collections is the quality of the raw data. Different from large, established collections, entries are often recorded once and rarely cross-checked, causing a lack of neatness and everlasting mistakes. In these cases, linking is a useful strategy both for performing a quality check and for spotting possible inconsistencies in the data. For example, in our collection, we found that the date of birth of four artists available in the VVV repository did not match with the ones published on DBpedia. We asked a domain expert for a final adjudication, and she further checked the dates in a third database, the Virtual International Authority (VIAF) File[20]. We found that: *(i)* in two cases DBpedia contained the correct information, *(ii)* in one case the VVV repository was right, and *(iii)* in another case VIAF did not match with any of the two other sources. This example shows that the quality of information stored by museums is not always more accurate than the one available, for example, in the Linked Open Data cloud. Thus, the challenge of providing accurate information to users is still an open challenge due to the necessity of double-checking all exposed data.

The problem of inconsistency in raw data affects also other parts of the collection. For instance, in some cases, the same field contains incompatible information. An example is the "author" field, which sometimes contains the author of the artwork, while in other cases it contains the editor of the artwork. These problems are caused by the inaccurate design of the data management system used for storing the raw data. However, in our data fusion activity, such issues are easily spotted and corrected automatically.

Other useful hints for current improvement and future developments were provided by the art curators during a demo session, in which the MoKi interface with the VVV records was presented. The curators were impressed by the power of such a knowledge management system and provided a generally positive feedback, praising in particular the time they could spare to complete records through automatic linking. The possibility to manually correct translations or to choose the best piece of information among a set of options was deemed to be very useful, since art experts would never trust a completely automatic enrichment process. A possible improvement, which they consider groundbreaking, would be the possibility to enrich information starting from an image, which should be linked automatically to related images or to additional information through image processing techniques. This would be, however, computationally very demanding and we will not be able to deliver it in the framework of the

[20] http://viaf.org.

VVV project. Another improvement suggested by the domain experts is to connect MoKi with the platform to navigate the collection, so that all the changes and corrections performed by them would be automatically shown to the visitors of the platform in real time, without the need to update periodically the database underlying the navigation platform. This suggestion is technically less demanding on the short run and we consider implementing it before the end of the project.

The only negative feedback provided by the curators was related to the usability of the discussion facility embedded in MoKi. Such an interface has been perceived as less intuitive with respect to the others included facilities. Curators suggested possible improvements that will be addressed in the next version of the tool.

Concerning future work, the effort for improving the results obtained in the VVV project will be focused on two directions. On the tool side, the workflow for managing data modeled in the knowledge base will be extended with an approval mechanism allowing (i) the management of a more fine-grained set of roles assigned to the users for enabling them to modify only parts of knowledge bases, and (ii) the possibility by the curators manager to decide which changes can be carried out on the knowledge base and which cannot. On the evaluation side, it is already planned to extend the evaluation to further collections coming from other local museums, in order to validate the interoperability level and the effort needed for migrating the entire platform to a different environment.

5 Related Work

Given the growing interest in enriching cultural heritage data using Semantic Web languages and techniques, a number of works have been proposed to address this issue. Hyvonen [13] presents an overview on why, when, and how Linked (Open) Data and Semantic Web technologies can be employed in practice in publishing cultural heritage collections and other content on the Web.

An example is provided by Szekely et al. [3], where the authors define a specific ontology which extends the EDM ontology. This extension is motivated by the lack of properties needed to represent specific data of the Smitsonian museum. The data is then linked to knowledge bases such as DBpedia, the Getty vocabulary "Ulan" (Union List of Artist Names), and the list of artists of the museum Rijksmuseum.

de Boer and colleagues [4,5] present the Amsterdam Museum Linked Open Data project, where the data of the museum has been transformed into RDF data in an automated way. The authors used the EDM ontology for defining the semantics of the data coupled with vocabularies such as the Dublin Core vocabulary.

On the one side, we share with these works the idea of starting from raw cultural heritage data available in different formats, e.g., tables, texts, CSV, and then transforming this data into semantic data. We all rely on the EDM ontology to start and we extend it depending on the purpose of our data translation task.

We also adopt other vocabularies like Dublin Core and FOAF to represent additional features. Moreover, a data interlinking step with DBpedia is addressed in these works as well. On the other side, several differences arise. More precisely, we do not only translate the raw data into RDF data and interlink it with DBpedia instances, but we address also an enrichment step where we extracted from the textual resources available on the Web structured information to be integrated with our starting data, e.g., information about the artists. Such an enrichment step has been achieved using Natural Language Processing techniques.

Other approaches deal with the definition of more specific ontologies and annotation tools. Benjamins and colleagues [14] present an ontology of Humanities then exploited into a semi-automatic tool for the annotation of cultural heritage data to ease the knowledge acquisition task.

Other related work has been proposed concerning multilingual access to cultural heritage information. Dannells and colleagues [15] address the problem of multilingual access to cultural heritage by adopting Semantic Web languages. They process museum data extracted from two distinct sources and making this data accessible in natural language, thanks to a grammar-based system designed to generate coherent texts from Semantic Web ontologies in 15 languages. Here, natural language processing techniques have been applied, coupled with Semantic Web technologies, similarly to our approach, even if the final goal is different from our one.

Besides the cultural heritage domain, annotations has been adopted also in the digital libraries field. In [16] the authors discuss a framework about the use of semantic annotations for enriching the content of digital archives.

Finally, the LOD cloud contains some examples of semantic data from museums. Among others, the British Museum[21] has published its data collection using the CIDOC-CRM ontology to allow data manipulation and reuse, and it provides a SPARQL access point to query the available data. The main difference with respect to this initiative is in the features of the two collections, i.e., a small and not well documented collection in the case of VVV, and a well documented and huge collection in the case of the British Museum. For these reasons, the interlinking of our data with other information sources in the LOD cloud is less efficace than in the case of the British Museum, as discussed in Sect. 4.2.

6 Conclusions

In this work, we described the process to link the metadata of a collection of verbo-visual art to DBpedia, and to enrich such a collection with additional information retrieved from the Web. We showed that the workflow proposed in the past for other important collections was applicable to our case only to a limited extent. In particular, the fact that the collection and the involved artists are in many cases little known affects the coverage of the interlinking and the amount of additional information retrieved from the Web. Besides, other issues

[21] http://collection.britishmuseum.org.

related to the management and the curation of the raw data, as well as those related to specific characteristics of VVV sample data, have to be addressed.

In the future, we plan to extend the enrichment process to the whole collection, and to make the data available. Besides, we will enrich the data also with information about the artwork content, since around 50 % of verbo-visual works contain some texts that can be transcribed and analyzed automatically with NLP techniques.

References

1. Oomen, J., Baltussen, L.B., van Erp, M.: Sharing cultural heritage the linked open data way: Why you should sign up. In: Proceedings of Museums and the Web (2012)
2. Marchetti, A., Tonelli, S., Sprugnoli, R.: The verbo-visual virtual platform for digitizing and navigating cultural heritage collections. In: Proceedings of the 2nd Annual Conference on Collaborative Research Practices and Shared Infrastructures for Humanities Computing (AIUCD-2013) (2013)
3. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the smithsonian american art museum to the linked data cloud. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 593–607. Springer, Heidelberg (2013)
4. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The Amsterdam museum case study. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 733–747. Springer, Heidelberg (2012)
5. de Boer, V., Wielemaker, J., van Gent, J., Oosterbroek, M., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Amsterdam museum linked open data. Semant. Web **4**(3), 237–243 (2013)
6. Ferrari, D.: Archivio di Nuova Scrittura Paolo della Grazia. Storia di una Collezione/Geschichte einer Sammlung. Silvana Editoriale, Milan, Italy (2012)
7. Dragoni, M., Bosca, A., Casu, M., Rexha, A.: Modeling, managing, exposing, and linking ontologies with a wiki-based tool. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), ELRA, pp. 1668–1675 (2014)
8. Ghidini, C., Rospocher, M., Serafini, L.: Modeling in a wiki with moki: Reference architecture, implementation, and usages. Int. J. Adv. Life Sci. **4**, 111–124 (2012)
9. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (eds.): Definition of the CIDOC Conceptual Reference Model. ICOM/CIDOC CRM Special Interest Group (2009)
10. Dragoni, M.: Multilingual ontology mapping in practice: a support system for domain experts. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9367, pp. 169–185. Springer, Heidelberg (2015). doi:10.1007/978-3-319-25010-6_10
11. Schmid, H.: Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop (1995)
12. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, COLING-1992, pp. 539–545. ACL (1992)

13. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, Palo Alto (2012)
14. Benjamins, V.R., Contreras, J., Blázquez, M., Dodero, J.M., García, A., Navas, E., Hernandez, F., Wert, C.: Cultural heritage and the semantic web. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 433–444. Springer, Heidelberg (2004)
15. Dannells, D., Ranta, A., Enache, R., Damova, M., Mateva, M.: Multilingual access to cultural heritage content on the semantic web. In: 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 107–115 (2013)
16. Agosti, M., Ferro, N.: Annotations: enriching a digital library. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 88–100. Springer, Heidelberg (2003)