Semantic Web 0 (0) 1
IOS Press

The Rijksmuseum Collection as Linked Data

Chris Dijkshoorn [a], Wesley ter Weele [b], Lizzy Jongma [c], Lora Aroyo [a]
[a] *Department of Computer Science, The Network Institute, VU University Amsterdam, the Netherlands*
[b] *History Department, VU University Amsterdam, the Netherlands*
[c] *Rijksmuseum Amsterdam, the Netherlands*

Abstract. Many museums make their collections accessible online. To reuse concepts and ease the integration of collections it is beneficial for institution to release their datasets as Linked Data. In this paper we present the Linked Data version of the Rijksmuseum, accessible at http://sealinc.ops.few.vu.nl/rijksmuseum/. We describe and provide statistics about the collection, the links to structured vocabularies and the links to other collections. The data presented in this paper is used in multiple ways: to enable aggregation, by users to explore the collection and by scholars to use structured queries to answer complex research questions.

Keywords: Linked Data, Cultural Heritage

1. Introduction

Museums are institutions that collect and preserve cultural heritage objects. Traditionally these objects are presented to the public using exhibitions. During the last decade a trend started were museums began to provide access to their collections through the internet, by digitising collection objects and making metadata available [4,10]. In this paper we present the Linked Data version of the Rijksmuseum Amsterdam collection, which enables users to explore the collection, researchers to answer complex research questions and allows for easy integration with other collections.

Digitising large collections is a long and costly process, an endeavour which the Rijksmuseum started in 2004, by photographing, registering and annotating collection objects. The museum uses thesauri to add information to the collection objects [6], which simplifies a translation to Linked Data. The data is modelled, generated and curated by the institution itself. This illustrates the shift from the research oriented approach of the Semantic Web to a more general adoption of the Linked Data principles [1].

This paper is structured as follows. In the next section we describe the characteristics of the Rijksmuseum collection and the used digitisation process. Section 3 provides details on the data model and the number of digital objects currently available. In Section 4 we give an overview of the links from collection objects to structured vocabularies and in Section 5 we mention applications using the Linked Data. We describe the process of finding links to the short-title catalogue Netherlands in Section 6 and conclude with discussing the current dataset and its known weaknesses.

2. The Rijksmuseum collection in a digital age

The Rijksmuseum Amsterdam has a collection of over 1,000,000 objects, including masterpieces by Rembrandt and Vermeer. Besides being one of the most visited museums in the Netherlands, the museum is also a valuable source of information for scholars. On average the museum has 8,000 object on display, just a fraction of the whole collection. To open up and improve access to the collection, the museum made parts of its collection available online in 2011.

In 2007 the museum started the Print Room Online project, specifically addressing the backlog of collection objects that needed to be digitised. Normally curators are solely responsible for the registration and annotation of objects, but for this project a team of six cataloguers, a photographer, a curator and a project manager was formed. Together they focus on digitis-

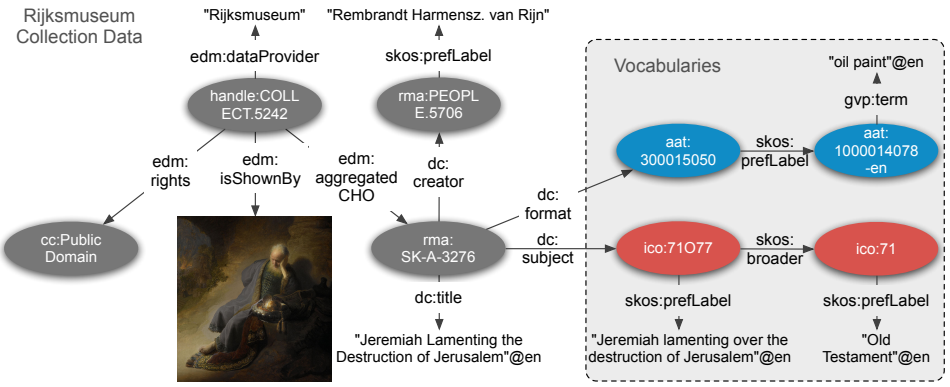


Fig. 1. Example of the painting *Jeremiah Lamenting the Destruction of Jerusalem* modelled according to the EDM data model.

ing the Rijksmuseum's 700,000 works of art on paper (prints, drawings and photographs), one of the largest collections in the world. This team registers, annotates and digitises nearly 30,000 prints a year.

The current workflow of the Print Room Online project is highly standardised and well documented. The catalogers register each object in the collection management system by entering its information (e.g. acquisition date, creator's name and name of the object) from the inventory books. Relevant literature regarding the particular artist is retrieved, serving as the main source of information. All books used in the annotation process are recorded as reference. Also annotations about subject matter are added, typically concerning the depicted event, people, place and objects. Upon completion of the annotation process, objects are handed over to a professional photographer for digitisation.

The correctness and high quality of the information describing collection objects is essential for the museum. The Print Room Online project leader does the majority of the quality control, e.g. checking that the right fields are filled according to the documentation, as well as the correctness and consistency of the description fields and the bibliography. The project leader spends roughly 8 hours per week on quality control, which results in checking about 150 objects a week. The curator is responsible for verifying the title and description fields.

3. Data model and dataset statistics

The Linked Data version of the Rijksmuseum collection is modelled according to the **Europeana Data Model** (EDM) [7]. This model distinguishes three core classes: cultural heritage objects, web resources and aggregations. An aggregation connects the metadata of a cultural heritage object to web resources, thereby allowing one object to be linked to one or more images. Figure 1 shows an example of one of the Rijksmuseum objects modelled according to EDM.

The EDM model is designed with the purpose of allowing integration on a collection level. To accomplish this, Europeana requires a content provider to include a predefined set of metadata fields. For each cultural heritage object the title and subject should be specified using predicates originating from Dublin Core terms[1]. The aggregation class should provide a link to the digital object, refer to the provider of the metadata and indicate copyright information regarding the digital object. Persistent identifiers in the form of handlers[2] are used for the URI's of the aggregations. The URI's of the collection objects are based on the purl scheme[3].

As of September 2014 the Linked Data version of the **Rijksmuseum collection** comprises 20,717,995 triples describing over 548,785 objects, of which 159,659 have a graphical depiction. The metadata is based on a conversion of a selection of the data in the content management system. Not all data is included, potentially sensitive data such as insurance value and the provenance of how a collection object was obtained is left out. Ten sub-collections are maintained, including sculptures (51,653 objects), historical items (46,756 objects), paintings (11,478 objects) and asian art (6,494 objects). The print collection is with 424,098 objects by far the largest sub-collection and includes prints, drawings and photo's. Metadata about the col-

[1] http://dublincore.org/documents/dcmi-terms/
[2] http://www.handle.net/
[3] https://purl.org/

lection is made available using the Vocabulary of a Friend (VOAF)[4] and the collection objects are linked to two structured vocabularies: the Art & Architecture Thesaurus and Iconclass.

4. Links to structured vocabularies

The **Art & Architecture Thesaurus**[5] (AAT) consists of concepts about arts from antiquity to the present. Concepts include art styles, materials and agents. It is maintained by the Getty foundation, which released a Linked Data version in February 2014 with 38,619 concepts. The focus of the thesaurus lies on generic concepts: instead for example describing individual artists, it includes the concept *printmakers*. New concepts originate from cataloging and documentations projects and labels of concepts are available in multiple languages.

The Rijksmuseum uses the Art & Architecture Thesaurus for the subject, type and format metadata. A small subset of the available concepts is used: 99 distinct types, 56 distinct subjects and 25 distinct formats. As can be seen in the subject frequency distribution in Figure 2a, a small number of concepts is often used. This is also the case for the type and format fields. For example the top three types are prints (302,459), stereoscopic photographs (10,952) and easel paintings (6,073). The museum refrains from assigning art styles to objects, since it is often debatable to which art style an object belongs.

The **Iconclass vocabulary**[6] contains 39,578 concepts, providing 'a systematic overview of subjects, themes and motifs in Western art'. A Linked Data version was released in 2012. Concepts are identified with codes and SKOS relations are used to create an hierarchy between them. Labels of concepts are available in English, German, French, Finnish and Italian.

An example of a code used in Iconclass is 7, which refers to the *Bible* and is connected to the concept 71O7, *the book of Jeremiah*, using skos:narrower predicates. Context dependent modifiers can be added to the codes, for 71C131(+3) the code 71C131 indicates *the sacrifice of Isaac*, while the modifier (+3) indicates that one or more angels are depicted on the artwork. These modifiers are not included in the database dump of the Linked Data version in order to maintain a reasonable file size, although they can be dereferenced.

The museum uses the Iconclass vocabulary to describe subject matter. Out of the 39,578 concepts in the vocabulary 11,945 are used once or more to add information to an object. Of the 548,785 collection objects 172,076 have one or more Iconclass annotations. As Figure 2b shows, many of the concepts are often used, on average a code is used 25,7 times. Since every thesaurus is limited by its scope, sometimes Iconclass will not cover all depicted subject matter.

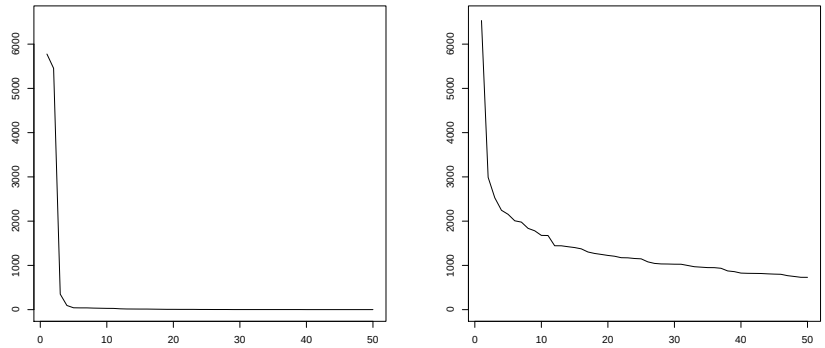
In the collection dataset 15,517 distinct subject matter annotations refer to Iconclass but are not a concepts in its Linked Data version. On the one hand this is caused by the lack of expanded notations in the loaded Linked Data version, on the other by the limited scope of the concepts in the vocabulary. Rijksmuseum catalogers are allowed to create new codes when the vocabulary does not suffice. As can be observed from the list of annotations not present in the Linked Data version, many geographic references are added. Examples are 61E(AMSTERDAM) and 61E(ITALY) which appear 1519 and 281 times respectively, but also concepts such as specific animal species are used (e.g. 25F37(SEA-GULL) and 25F62(GURNARD)).

To get access to vocabularies that allow more specific annotations of animal species, the Rijksmuseum has started a collaboration with the Naturalis biodiversity center[7], which proves to be a rich source of information regarding **taxonomies of animal species**. By using a mapping approach based on scientific names, we integrate multiple heterogeneous files containing information on animal species into a Linked Data version[8]. The resulting data is currently used in a crowdsourcing initiative, set up to annotate prints depicting birds [5].

5. Applications

The initial reason for the Rijksmuseum to start working with Linked Data was to enable **multi-lingual access** to its collection. After a ten-year long renovation the museum reopened in 2013, which got a lot of media attention. This attracted a lot of visitors from all over the world, creating an imminent problem of website visitors that were not Dutch spoken. The website already provided access to many collection items

[4] http://purl.org/vocommons/voaf
[5] http://www.getty.edu/research/tools/vocabularies/aat/
[6] http://www.iconclass.nl/
[7] http://www.naturalis.nl/
[8] http://github.com/rasvaan/naturalis



(a) AAT concept frequency distribution (b) Iconclass concept frequency distribution

Fig. 2. Frequency distributions of the top 50 subject concepts of AAT and Iconclass that the collection objects are linked to.

with well curated metadata. However, most of these descriptions were in Dutch. To enable multi-lingual access to the collection the museum started looking into using Linked Data, trying to benefit from the labels in structured vocabularies available in many different languages.

Creating a Linked Data version of the collection also enabled the ingestion of the collection by **Europeana**. Europeana is an organisation which provides a portal[9] to digital representations of collection objects of over 200 cultural institutions from 15 countries in Europe. Additionally an aggregated Linked Data set of the collections is available [8]. Having the Rijksmuseum collection accessible through Europeana increased its online exposure, although the point of access has to cater for many collections, making it impossible to tailor it to the specific characteristics of the collection.

To further illustrate the potential of Linked Data for object retrieval we created the **cluster search** system[10]. This system runs on top of a triple store with the Rijksmuseum collection loaded, alongside a number of additional structured vocabularies. Based on a keyword query users are presented with clusters of artworks. The links from collection objects to structured vocabularies create the potential of finding otherwise hidden relations between artworks. The system generates the clusters using a graph search algorithm [11]: the algorithm matches the keywords with literals and propagates back using the graph structure of the collection data and structured vocabularies, until it finds artworks. The artworks are clustered together using an abstraction of the path used to find them. The cluster search system provides an ideal basis for users to explore the collection.

The sheer size makes the Rijksmuseum collection an interesting **research artefact** for humanities scholars. To illustrate this we look more closely at answering one of the many possible research questions: *"How did the popularity of bible themes change between 1650 and 1750"*. The popularity of bible themes can be derived from the prints included in printed bibles. When we talk about bible themes we do not refer to a whole bible book, like Genesis, but to *"the building of the ark"* or *"the birth of Jezus Christ"*. Not all of these themes could be depicted in a printed bible, publishers had to make a selection based on the demand of their customers. Bibles are therefore a great way to observe the popularity of themes.

Objects in the Rijksmuseum print collection can be standalone items, but can also be part of books like a diary, photo album or bible. To answer the research question stated above we need to determine which prints were originally used to illustrate bibles and when these bibles were published. Because the biblical themes of the prints are classified using Iconclass, associating them to bibles would make it possible to see the differences in popularity of a theme as time changes. This information is not included in the Rijksmuseum collection, however in the next section we describe the approach used to link to another collection, thereby gaining access to the required information.

[9] http://www.europeana.eu/
[10] http://sealinc.ops.few.vu.nl/clustersearch/

6. Links on a collection level: the short-title catalogue Netherlands

The **Short-Title catalogue Netherlands** (STCN) is 'the retrospective national bibliography of the Netherlands in the period 1540-1800'[11], maintained by the National Library of the Netherlands. A Linked Data version is available[12], containing records of over 139,817 publications. Besides having the potential of containing many books that were the source of objects in the print collection, the catalogue most likely contains the information of Dutch printed bibles in the period 1650-1750, needed to answer the research question introduced in Section 5. To generate links between prints and books either a reference from the print to book or the other way around is needed.

During the annotation process of a collection object Rijksmuseum catalogers add references to related works. In case a print was originally part of a book, they refer to the records of the National Library. Unfortunately, they do not use unambiguous identifiers, but refer to some characteristics of the book such as the title, publication date and publisher. Although the Rijksmuseum has guidelines for referring to books, the references are remarkably unstructured, requiring interpretation in order to be able to match.

We require three matches between the Rijksmuseum collection objects and the STCN books: author, year of publication and the title of the book. The STCN includes this information in separate fields, which makes querying for it straightforward. Only the title field offers some problems, mentioning sometimes also the author or publisher. After normalising this information we check the unstructured references in the Rijksmuseum collection for the presence of all three aspects. This matching process results in 3598 links from the Rijksmuseum collection to 501 publications in the STCN catalogue. The links are encoded as dc:hasPart relations from the STCN vocabulary to the Rijksmuseum collection. As can be seen below, making use of these links researchers can now query for all related themes to a specific bible.

```
SELECT DISTINCT ?theme WHERE {
 stcn:102584524 dcterms:hasPart ?print .
 ?print dc:subject ?ic .
 ?ic skos:prefLabel ?theme .
}
```

[11] http://www.kb.nl/expertise/voor-bibliotheken/short-title-catalogue-netherlands (accessed on 04-07-2014)

[12] http://github.com/wouterbeek/STCN

7. Discussion

Publishing the Rijksmuseum collection data has resulted in a number of new ways of interacting with the collection. Besides being accessible using the Europeana portal, users are now able to explore the collection using new approaches such as the cluster search system, which leverages relations found in the linked structured vocabularies. Researchers are able to use queries over multiple collection, thereby being able to answer more complex research questions.

Publishing the data is an ever evolving process. The current version is shaped by numerous discussions with aggregators and other users of the data. The choice for the EDM model was a pragmatic one, it enables Europeana to ingest the data. In the future also other models can be considered, such as the CIDOC-CRM model[13]. While many aspects of the data may evolve, the identifiers for the individual artworks will be persistent, enabling others to link to the collection objects.

All the criteria for five star Linked Data as defined in [9] are met. There is a description of the data online[14], the data is available in RDF, there are many links to structured vocabularies and metadata about the collection is made available. Furthermore, books in the Linked Data version of the short-title catalogue Netherlands link to the prints in Rijksmuseum collection.

As identified in Section 4, the currently used structured vocabularies do not cover every aspect that the Rijksmuseum catalogers wish to describe. The description of the collection could be improved by incorporating structured vocabularies with a different scope. The museum is open for doing this, but often runs into problems with availability and licences of Linked Data sets. For example, geographic information could be described using Geonames[15], but this vocabulary is not downloadable for free. Luckily more and more datasets become available under public licenses, the recent release of the Getty Thesaurus of Geographic Names[16] is an illustration of that and this vocabulary will be linked to in the future.

The catalogers annotating the prints have a limited amount of time and therefore only the most important aspects are being added in the collection management

[13] http://www.cidoc-crm.org/

[14] http://sealinc.ops.few.vu.nl/rijksmuseum

[15] http://www.geonames.org

[16] http://www.getty.edu/research/tools/vocabularies/tgn/

system. The Rijksmuseum will look into crowdsourcing mechanisms such as used in the Steve.museum initiative [3], to collect more information regarding for example the subject matter of prints. The same could for instance be done for art styles. While the museum itself is reluctant to add this information to the collection management system, it could be very beneficial for the end-users.

The quality and correctness of metadata is of paramount importance to museums [10]. Internally the Rijksmuseum has an extensive quality control process to ensure the correctness of the annotations. When eventually data is added to the collection by using enrichment from other structured vocabularies, these vocabularies have first to be judged on their quality. The same principle goes for adding information obtained from crowdsourcing processes. Since this will most likely result in data of heterogenous quality, the museum can look into automatic trust assessment techniques [2].

The Rijksmuseum currently selects, curates and generates RDF data. However, the data itself is hosted at a university server. This is problematic since every day around a thousand records are updated, making the data on the university server quickly outdated. Regularly updating this server with newly available data will partially solve this, although in the future we want to look into hosting the data at the institution, making the institution solely responsible for its own data.

Acknowledgements This publication was supported by the Dutch national program COMMIT/. We are grateful to all our colleagues in the SEALINCMedia and INVENiT projects for the discussions on this subject.

References

[1] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[2] Davide Ceolin, Archana Nottamkandath, and Wan Fokkink. Efficient semi-automated assessment of annotations trustworthiness. *Journal of Trust Management*, 1(1):3, 2014. ISSN 2196-064X. .

[3] Susan Chun, Rich Cherry, Doug Hiwiller, Jennifer Trant, and Bruce Wyman. Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums. In Jennifer Trant and David Bearman, editors, *Museums and the Web 2006: Proceedings*, 2006. URL `http://www.archimuse.com/mw2006/papers/wyman/wyman.html`.

[4] Victor de Boer, Jan Wielemaker, Judith van Gent, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012)*, volume 7295 of *Lecture Notes in Computer Science*, pages 733–747. Springer Berlin / Heidelberg, Heraklion, Greece, 2012. ISBN 978-3-642-30283-1. URL `http://dx.doi.org/10.1007/978-3-642-30284-8\_56`.

[5] Chris Dijkshoorn, Mieke H. R. Leyssen, Archana Nottamkandath, Jasper Oosterman, Myriam Traub, Lora Aroyo, Alessandro Bozzon, Wan Fokkink, Geert-Jan Houben, Henrike Hovelmann, Lizzy Jongma, Jacco van Ossenbruggen, Guus Schreiber, and Jan Wielemaker. Personalized nichesourcing: Acquisition of qualitative annotations from niche communities. In *PATCH 2013: Personal Access to Cultural Heritage (UMAP Workshops)*, 2013. URL `http://ceur-ws.org/Vol-997/patch2013_paper_13.pdf`.

[6] Michiel Hildebrand, Jacco Van Ossenbruggen, Lynda Hardman, and Geertje Jacobs. Supporting subject matter annotation using heterogeneous thesauri: A user study in Web data reuse. *International Journal of Human-Computer Studies*, 67 (10):887–902, 2009. .

[7] Antoine Isaac. Europeana data model primer, July 2013. URL `http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5`.

[8] Antoine Isaac and Bernhard Haslhofer. Europeana linked open data – data.europeana.eu. *Semantic Web Journal*, 4(3):291–297, January 2013. .

[9] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman II. Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

[10] Pedro Szekely, CraigA. Knoblock, Fengyu Yang, Xuming Zhu, EleanorE. Fink, Rachel Allen, and Georgina Goodlander. Connecting the smithsonian american art museum to the linked data cloud. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 593–607. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38287-1. .

[11] Jan Wielemaker, Michiel Hildebrand, Jacco van Ossenbruggen, and Guus Schreiber. Thesaurus-Based Search in Large Heterogeneous Collections. In Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin, and Krishnaprasad Thirunarayan, editors, *Proceedings of the 7th International Semantic Web Conference (ISWC2008)*, volume 5318 of *Lecture Notes in Computer Science*, pages 695–708, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-88563-4. .