The Rijksmuseum Collection as Linked Data

Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele, Jan Wielemaker

Abstract. Many museums are currently providing online access to their collections. The state of the art research in the last decade shows that it is beneficial for institutions to provide their datasets as Linked Data in order to achieve easy cross-referencing, interlinking and integration. In this paper, we present the Rijksmuseum linked dataset (accessible at http://datahub.io/dataset/rijksmuseum), along with collection and vocabulary statistics, as well as lessons learned from the process of converting the collection to Linked Data. This dataset contains over half a million objects, including detailed descriptions and high-quality images released under a public domain license.

Keywords: Linked Data, Open Data, Image Collections, Cultural Heritage, Museums

1. Introduction

Publishing cultural heritage collections as Linked Data improves reusability of the data and allows for easier integration with other data sources [1,13]. Concepts providing context for collection items are often shared among multiple cultural heritage organisations, which is an ideal basis for creating connections between collections and allowing reuse of information [9,15]. The availability of data models tailored towards publishing cultural heritage data helps to make the data available in an interoperable way [3,4]. These benefits have become apparent to the sector, resulting in an increase of attention and the development of methodologies to help institutions overcome the hurdles involved in publishing data according to the Linked Data principles [1,8,13].

The Linked Data version of the Rijksmuseum collection has some unique features. The data is a result of a joint effort between the museum, CWI and VU University Amsterdam and has evolved with input from many research projects [12,14,16]. Nowadays, employees of the museum are in control of the publishing process, creating and maintaining a conversion layer from content management system to Linked Data. The museum’s digitisation process includes the use of external datasets for adding contextual concepts (e.g. creator or technique), creating manually curated links towards external datasets [7]. The data is continuously extended: every day new objects and descriptions are added and both metadata and images are released under open licenses when possible.

This paper describes the current state of the Rijksmuseum Linked Data and provides insights into the lessons learned during its creation. In the next section, we describe the characteristics of the Rijksmuseum collection and its digitisation process. The historical development of the dataset is given in Section 3. Sections 4 and 5 provide details on the data model and the number of digital objects currently available. In Section 6 we give an overview of the links from collection objects to external data sources. In Section 7 we give an overview of how the museum data is used, before we conclude in Section 8 with discussing the current dataset.

2. The Rijksmuseum in a digital age

The Rijksmuseum Amsterdam is one of the most visited museums in the Netherlands, with a mission
to provide a representational overview of Dutch art from the Middle Ages onwards. It is well known for its Golden Age paintings, including artworks by Rembrandt and Vermeer. The collection comprises over a million objects, of which only a fraction can be on display at a given time. To open up the remaining collection the museum started digitising objects and publishing them online.

Digitising large collections is a time consuming and costly endeavour. To address the backlog of items to be digitised, the Rijksmuseum started a dedicated digitisation project, employing catalogers and a professional photographer. The catalogers register objects in the content management system and describe the objects, using structured vocabularies if available [7]. The photographer takes high-quality images which are released under a public domain license when possible, waiving the rights of the museum.

The digitised collection items are accessible through the website of the museum. Online visitors can explore the collection using categories or they can search for specific keywords. The presentation of the website focusses on high-quality images of collection objects, encouraging users to save, manipulate, and share them [6]. Developers can use an Application Programming Interface (API) to get access to information about the collection objects, sub-collections created by users, and event information1.

3. History of the Rijksmuseum Linked Data

The Linked Data version of the Rijksmuseum dataset has a long history, influenced by a number of research projects. A first Resource Description Framework (RDF) version comprising 750 top pieces was created by converting a datadump of an educational database [5]. As a next step, in an effort to integrate Dutch cultural heritage collections, the datamodel was changed to follow the VRA Core specification2, with the key advantages of allowing the use of Dublin Core constructs3 and making a distinction between the physical artwork and its digital representations. The meta-data values of objects were represented in plain text.

In a next version, contextual concepts from in-house thesauri of the Rijksmuseum were aligned with the Getty thesauri and WordNet, resulting in a dataset of 27,993 triples [12]. At the time, the Getty vocabularies were only available under license and in XML format, which resulted in the need for an internally maintained conversion to RDF. In a similar effort, Iconclass was converted and aligned using the Simple Knowledge Organization System (SKOS) to formalise its structure [15]. The experiences gained served as input for the SKOS specification.

The Rijksmuseum dataset was one of the first entries in the Europeana Thought Lab4, an initiative for showcasing experimental technologies. This entry marks the first conversion of all available Rijksmuseum collection data: 46,000 objects with images were obtained from the content management system and converted to comply with the VRA data model. The experience of modelling the complete collection and integrating it with collections from other institutions required the ability to model different (potentially conflicting) metadata records from different sources describing the same artwork. These and other obtained requirements influenced the creation of the Europeana Data Model [4].

The Europeana Data Model today has a set of core and contextual classes that can capture collection information. The data model is designed with reuse of existing classes and properties in mind. It includes elements from the Dublin Core metadata initiative and the Object Reuse and Exchange definition of the Open Archives Initiative5. Cultural heritage organisations can extend the set of classes and properties when needed, reusing elements of other data models or by defining their own.

The possibility of making the collection available on the Europeana portal led to the museum taking matters into their own hands. The museum converts the collection to Linked Data by adding a translation layer on top of the API of the collection management system. Every relevant metadata field of a collection object is mapped to a property in the Europeana Data Model that most closely resembles the values of the field. We describe the resulting data model for the Rijksmuseum collection in the next section. Mappings are recorded in Extensible Stylesheet Language Transformations (XSLT) and enable a transformation into RDF. We provide statistics of a harvest of data obtained using the resulting API in Section 5 and list uses of the data in Section 7.

1https://www.rijksmuseum.nl/en/api
2http://www.loc.gov/standards/vracore/schemas.html
3http://dublincore.org/
4http://labs.europeana.eu/apps/SearchEngineEuropeana
5http://www.openarchives.org/ore/
4. Data model and URIs

The Linked Data version of the Rijksmuseum collection is modelled according to the Europeana Data Model (EDM). The model makes a distinction between a collection item and its digital representation(s). This is achieved with three core classes: edm:ProvidedCHO for cultural heritage objects, edm:WebResource for web resources and ore:Aggregation for aggregations of resources. Figure 1 shows the metadata of a Rembrandt painting and its core and contextual classes.

An ore:Aggregation is used to connect the metadata of a cultural heritage object to web resources. Every collection item in the content management system gets an aggregation object with its persistent identifier as URI. Information can be added to the ore:Aggregation, Figure 1 for example shows that the Rijksmuseum served as data provider.

Every ore:Aggregation is connected to a resource with class edm:ProvidedCHO, representing a description of the physical cultural heritage object. Figure 1 shows four of the properties used to describe objects in the Rijksmuseum dataset: dc:creator, dc:title, dc:format and dc:subject. When possible, concepts are used to describe aspects of the artwork, such as the thesaurus term purl:PEOPLE.5706 for Rembrandt and the concept aat:300015050 for oil paint. Section 5 lists the occurrences of predicates used to describe objects in the Rijksmuseum dataset.

When a digital representation is available, the aggregation points to the URL were the image can be obtained. This URL is of type edm:WebResource and can in turn be described with metadata, adding for example information about its creator. Note that the creator of the image most often differs from the creator of the artwork. The Rijksmuseum dataset currently includes information about the date of creation and the file format of the image.

Not all subtleties of the collection data can be captured by using constructs included in the Europeana Data Model description. While the original data includes fields like ‘rejected creator’, no such property exists in Dublin Core. EDM allows for refining and extending the data model, so in the future the museum can choose to introduce its own more specific constructs or find others to reuse. This could increase the coverage of data in the collection management system included in the Linked Data version.

Persistent identifiers in the form of handles\(^6\) are used for the URIs of the ore:Aggregation. Since an aggregation connects metadata of the artwork and its digital representation, the persistent identifier is not related to the object number of the artwork. The URI of the cultural heritage object descriptions is based on the purl scheme\(^7\) and consist of five elements: purl prefix, dataset type, country code, organisation, and object number. This results in the following URI for the edm:ProvidedCHO resource of the Rembrandt in Figure 1: http://purl.org/collections/nl/rma/SK-A-3276.

When values refer to one of the thesaurus databases of

\(\text{6}http://www.handle.net/\)
\(\text{7}https://purl.org/\)
the museum, a URI is generated based on the internal reference used, linking the collection object with the thesaurus.

5. Rijksmuseum dataset statistics

As of September 2015 the Linked Data version of the Rijksmuseum collection comprises 16,403,040 triples, describing 437,607 objects, of which 144,185 have a graphical depiction. Metadata about the collection is made available using the Vocabulary of a Friend (VOAF). Ten sub-collections are maintained, including sculptures (46,684 objects), historical items (26,902 objects), paintings (5,246 objects) and Asian art (5,029 objects). The print collection has 353,564 objects and is by far the largest sub-collection, including prints, drawings and photos.

Table 1 lists the predicates used to describe collection items. A title is provided for almost all artworks, of which the majority is unique. Over half of the objects have a textual description, while curatorial comments are added using skos:note. There are over thirty thousand unique creators, with dc:contributor referring to additional persons involved in the creation process. dc:subject provides information about the subject matter, while dc:creator identifies the type of artwork (e.g. print or painting) and dc:format is used to specify materials.

Table 1

<table>
<thead>
<tr>
<th>predicate</th>
<th>distinct artworks</th>
<th>distinct objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc:contributor</td>
<td>96,608</td>
<td>9,614</td>
</tr>
<tr>
<td>dc:coverage</td>
<td>184,365</td>
<td>216</td>
</tr>
<tr>
<td>dcterms:created</td>
<td>433,036</td>
<td>26,313</td>
</tr>
<tr>
<td>dc:creator</td>
<td>434,986</td>
<td>31,691</td>
</tr>
<tr>
<td>dc:description</td>
<td>266,754</td>
<td>210,622</td>
</tr>
<tr>
<td>dcterms:extent</td>
<td>361,628</td>
<td>79,147</td>
</tr>
<tr>
<td>dc:format</td>
<td>402,942</td>
<td>609</td>
</tr>
<tr>
<td>dcterms:hasPart</td>
<td>2,365</td>
<td>65,848</td>
</tr>
<tr>
<td>dc:identifier</td>
<td>437,607</td>
<td>437,607</td>
</tr>
<tr>
<td>dcterms:isPartOf</td>
<td>67,207</td>
<td>2,764</td>
</tr>
<tr>
<td>dcterms:isReferencedBy</td>
<td>124,073</td>
<td>85,960</td>
</tr>
<tr>
<td>dc:language</td>
<td>437,607</td>
<td>&quot;nl&quot;</td>
</tr>
<tr>
<td>skos:note</td>
<td>58,973</td>
<td>28,017</td>
</tr>
<tr>
<td>dcterms:provenance</td>
<td>140,893</td>
<td>7,164</td>
</tr>
<tr>
<td>dc:publisher</td>
<td>437,607</td>
<td>&quot;Rijksmuseum&quot;</td>
</tr>
<tr>
<td>dc:rights</td>
<td>437,607</td>
<td>18</td>
</tr>
<tr>
<td>dcterms:spatial</td>
<td>260,926</td>
<td>2,883</td>
</tr>
<tr>
<td>dc:subject</td>
<td>263,312</td>
<td>62,157</td>
</tr>
<tr>
<td>dc:title</td>
<td>435,522</td>
<td>343,058</td>
</tr>
<tr>
<td>dc:type</td>
<td>437,529</td>
<td>3,779</td>
</tr>
<tr>
<td>edm:type</td>
<td>437,607</td>
<td>&quot;IMAGE&quot;</td>
</tr>
</tbody>
</table>

Overview of the predicates in the Rijksmuseum collection that describe collection items

code of the country of the institution, in this case "nl". The edm:type is "IMAGE".

6. Contextual concepts and links to external datasets

Institutions often maintain their own vocabularies containing their perspective on contextual objects. When the contextual objects of collection items are replaced with objects from standardised vocabularies such as the Getty vocabularies, these nuances in perspectives are in danger of disappearing. So while collection objects and contextual concepts in the thesauri of the Rijksmuseum are linked to an increasing number of available datasets maintained by other institutions, the Rijksmuseum chooses to also maintain its own.

Five contextual classes are defined in the data model for relating collection items to contextual information: edm:Agent, edm:Place, edm:TimeSpan, and skos:Concept. These classes correspond to the types of thesaurus records in the databases of the Rijksmuseum: the person database maps to the agent class and the general thesaurus database contains informa-

---

8http://datahub.io/dataset/rijksmuseum
9http://purl.org/vocommons/voaf
tion about places, historical events, and materials. Hierarchies in these thesauri, such as broader and narrower relations, are modelled using SKOS. In the remainder of this section we describe how external datasets are used to extend the thesauri and annotate the collection data of the Rijksmuseum.

The Art & Architecture Thesaurus \(10\) (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 of for example describing individual artists, it includes the concept "printmakers". New concepts originate from cataloging and documentation projects and labels of concepts are available in multiple languages.

The museum uses the Art & Architecture Thesaurus for the \textit{dc:subject}, \textit{dc:type} and \textit{dc:format} metadata fields. A small subset of the available concepts is used: 317 distinct formats, 126 distinct types and 54 distinct subjects. 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 (235,410), stereoscopic photographs (10,651) and easel paintings (4,639). The museum refrains from assigning art styles to objects, since it is often debatable to which art style an object belongs.

The Iconclass vocabulary \(11\) contains 39,578 concepts, providing ‘a systematic overview of subjects, themes and motifs in Western art’. An official 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 \textit{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.

The museum uses the Iconclass vocabulary to describe subject matter. Out of the 39,578 concepts in the vocabulary, 11,850 are used once or more to add information to an object. Of the 437,607 collection objects, 189,041 have one or more Iconclass annotations. As Figure 2b shows, many of the concepts are often used, while on average a code is used 29.3 times.

The Short-Title catalogue Netherlands (STCN) is ‘the retrospective national bibliography of the Netherlands in the period 1540-1800’ \(12\), maintained by the National Library of the Netherlands. A Linked Data version is available, containing records of 196,396 publications. This dataset contains many books that are the source of objects in the print collection of the

\(10\)http://www.getty.edu/research/tools/vocabularies/aat/

\(11\)http://www.iconclass.nl/

\(12\)http://www.kb.nl/expertise/voor-bibliotheken/short-title-catalogue-netherlands (accessed on 04-07-2014)
Rijksmuseum and linking the two collections provides valuable contextual information.

The catalogers of the Rijksmuseum add references to the National Library by adding textual descriptions of the books in a notes field. To create links these descriptions are scanned for objects from the STCN that match the title, publication date and publisher. This matching process resulted 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.

7. Data usage

In this section we give an overview of how the museum data has been used in various research projects and provide statistics about the Rijksmuseum API. Most projects that contributed to the process of data development had demonstrators illustrating the power of Linked Data. The MultimediaE-Culture project showcased a semantic search system, which won the 1st price in the 2006 International Semantic Web Conference Challenge [12]. It clustered search results based on the graph path leading from matching literal to artwork. The dataset was extended from 750 artworks to the entire Rijksmuseum collection in a search prototype of the Europeana Thought Lab, showing advanced search functionality to be included in the portal at a later stage.

Other ways of accessing data were introduced in subsequent years. The CHIP demonstrator recommended artworks based on graph patterns [16]. The STITCH project took a different approach with facets based on Iconclass concepts, allowing the user to browse the collection based on different topics [15]. The Agora demo provided access to the collection with an emphasis on the events related to objects [14]. The Accurator crowdsourcing tool of the SEALINCMedia project uses graph patterns to recommend people artworks to which they can contribute information, gathering more accurate subject matter descriptions [2].

The Rijksmuseum maintains an API for application developers, optionally returning data formatted according to the Europeana Data Model. 587 people have registered for access to the API as of August 2015. The API is used by Europeana to harvest collection data, making all Rijksmuseums structured data available through the Europeana portal. Europeana logs the page views of this portal and during a period of 20 weeks (starting from the 1st of May 2015) Rijksmuseum collection objects got 42,156 page views of which 34,206 were unique.

8. Discussion

For decades, Linked Data has been a promise for data publication and integration in the cultural heritage sector. Despite widespread interest and apparent advantages, only a limited number of institutions have managed to make their collection available as Linked Data. After a period of development influenced by many research projects, the Rijksmuseum is one of them. Furthermore, the museum is in control of the entire publication process of its own collection as Linked Data.

The majority of the Rijksmuseum collection items are part of the public domain since their intellectual property rights have expired. Although general understanding is that digitised representations of public domain works should again be released under the same license terms, many institutions are hesitant to do so, in fear of losing a possible revenue stream. The Rijksmuseum did release their high-quality images in the public domain in 2013, arguing that the increase in attention and exposure would result in a higher number of visitors [11]. In turn it allowed the museum to gain more control over the digital representations that had appeared online, replacing many inferior versions by its high-quality images.

The quality and correctness of metadata is of paramount importance to museums [13]. The Rijksmuseum has an extensive quality control process in place to ensure the correctness of metadata. By adding a direct conversion layer to the collection management system it ensures that the same level of quality is translated to the Linked Data version. All the criteria for five star Linked Data as defined in [10] are met. There is a description of the data online, the data is available in RDF, there are many links to structured vocabularies and metadata about the collection is made available. Furthermore, concepts in the Linked Data version of the Smithsonian American Art Museum are linked to the thesauri of the Rijksmuseum [13].

Data aggregators such as Europeana enticed many institutions to provide digital versions of their collection, often relying on external expertise for the conver-
sion process. This led to an increase of available collections, although providing access to data through aggregators has the major drawback that it creates a gap between the institution and its data [1]. We believe it is therefore still desirable that institutions publish their own data, if the required expertise is available. They thereby remain in control of choosing the most suitable data model, URI naming schemes, links to other datasets, and update processes.

The data of the Rijksmuseum is subject to constant change: newly digitised objects are added on a daily basis and employees extend and refine information in a broader context, allowing others to benefit from the progress made through easy reuse and the possibility to add new perspectives to the data.

Acknowledgements This publication was supported by the Dutch national program COMMIT/.

References


