NETWORK DIGITAL HERITAGE

A *knowledge graph* for the integration of heritage information

February 2016

spinque

Commissioned by Network Digital Heritage

Authors:
Michiel Hildebrand
Wouter Alink

Table of contents

Summary

The National Strategy for Digital Heritage contributes to the effort to provide unlimited access to cultural heritage collections. In order to achieve this, the collections must be linked in a meaningful way. In this context, the controlled vocabulary used by heritage institutions to describe their collection items plays a significant role. In this report, we give advice on how to link the controlled vocabulary in order to form a sustainable and usable alignment layer. We propose to construct a knowledge graph for heritage information to act as an alignment layer. For users the knowledge graph offers a uniform view of persons, places, events and concepts appearing in the collections. For contributors the knowledge graph is realised and managed in a distributed manner. Heritage institutions make their contribution by sharing and linking their controlled vocabularies. Sources from outside the heritage sector are also relevant.

There are four prerequisites for creating a knowledge graph for heritage information in a sustainable manner:
1. everyone can become a node and register a controlled vocabulary;
2. nodes guarantee that a given controlled vocabulary is and remains publicly available and that the source is interoperable with other sources by integrating their schemes;
3. those who manage a controlled vocabulary are responsible for its integration, together with other vocabularies, in the network by linking the terms;
4. collection managers describe their collections with terms originating from the network.

We describe the four actors that should guarantee these four prerequisites in a sustained manner: the Network for Digital Heritage, the nodes, the domain experts and the collection managers. With regard to the tasks of these four actors, we describe the requirements that must be met by their respective services in order to carry out the tasks.

Based on this report, the steering group of the 'Usability' project can prepare a resolution for the starting points, global concept and implementation of the alignment layer as a part of the national infrastructure for digital heritage.

1 Introduction

1.1 Background

With its National Strategy for Digital Heritage, the Network for Digital Heritage (NDE) strives to enhance digital facilities and services for heritage institutions. These facilities and services must contribute to improve sustainability, usability and visibility of the cultural heritage of the Netherlands. The initiative covers different domains of cultural heritage, each one being represented by a national node. The present report stems from, and is relevant to, the 'Usability'.This project aims to improve accessibility to the collections by improving thematic searches across collections in all heritage domains. In order to meet this goal, the national strategy creates an alignment layer encompassing all individual collections.

Fig. 1 shows the common situation in the heritage sector. Each institution manages a *collection* and describes the collection items with *terms* (or concepts) based on its own terminology list, thesaurus or register. In this report, we refer to this as a *controlled vocabulary*. In some cases, we also refer to it as a *source*; however, in the context of this report we mean by this 'a controlled vocabulary' and not a 'collection'. In the current situation, institutions provide access to their collections themselves, using their own vocabulary.

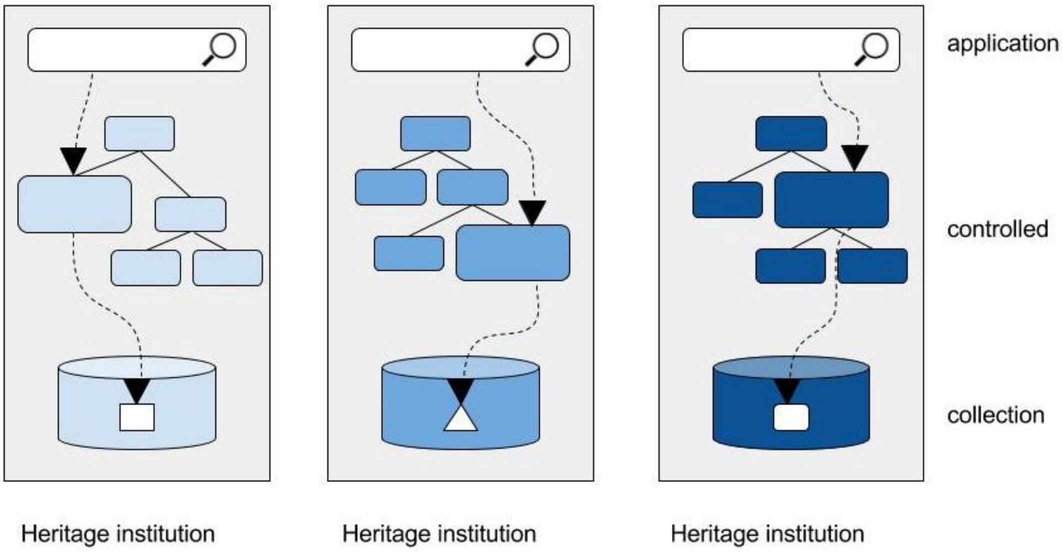

Fig 1. Common situation in the heritage sector. Each heritage institution describes collection items in its own way with its own metadata systems and terms from its own internal vocabulary. A heritage institution manages its own controlled vocabulary. Users can search the collections with the terms from the vocabulary. The collections are thus made accessible in isolation.

When these collections and related vocabularies are merged (aggregated), a problem occurs: Vocabularies of different collections can contain similar (or related) concepts, but these relationships are not explicitly recorded. So a system which combines collections will not be able to show which objects deal with particular subjects and which do not. Therefore an alignment layer is needed to reveal the relationships between the various terms of the vocabulary. The terms of the vocabularies are linked to each other to form a terminology network. In this way, corresponding objects are identified and the collections are indirectly integrated.

Managing a vocabulary at the level of the institution has another disadvantage: The management of a vocabulary is time intensive while this usually is not a primary task of the institution. Consequently, the coverage and richness of the vocabulary is often restricted to the necessities of the collection in question. In addition, a number of sources have been created through public initiatives, that are often actively managed, have a large coverage and rich content, for example the controlled vocabularies managed by the domain nodes within NDE. But sources managed outside the heritage sector could also be relevant, such as Geonames, DBPedia and Wikidata. All of these large public sources will play a significant role within the alignment layer.

1.2 Mission

Spinque has been commissioned through the NDE project ‘Usability’ to translate these goals into a usable service concept with particular attention to:

- the role of controlled vocabularies and related services necessary to actually implement the integration and availability of heritage data,
- the envisioned technical components and to which degree they are already available, or in development,
- the way in which the heritage sector can follow the developments regarding linked data and knowledge networks outside of the heritage domain.
- a user-based perspective through which the intended results can be communicated effectively

The focus of this report lies on the alignment layer made possible by a sustainable and user-friendly term network. To illustrate the way in which this alignment layer works, points of relevance with, for example, the management of collections and the aggregation of collection data can be described. Collection management and aggregation are itself beyond the scope of this report. The topic of aggregation of collection data is developed in the ‘Usability’ project position paper by Johan Oomen, Wilbert Helmus and Enno Meijers.

Based on the expert opinion expressed in this paper, the steering group of the NDE project ‘Usability’ can formulate a decision about the basic principles, global implementation and the realization of a term network and related services.

1.3 Approach

This advisory report has been composed through (desk)research, interviews with steering group members (or representatives thereof) and the chairman of the steering group. The steering group members represent the national nodes of the Network for Digital Heritage. We carried out interviews with:

- Dirk Houtgraaf and Joop VanderHeiden of the Cultural Heritage Agency of the Netherlands (RCE)
- Enno Meijers of the Royal Library (KB)
- Vincent Huis in ‘t Veld and Willem Melder of the Netherlands Institute for Sound and Vision (Beeld en Geluid)
- Pieter Koenders of the National Archives (NA)
- Annette Gaalman of Brabant Heritage (EB)

1.4 Structure

Based on the interviews, we describe current developments, aspirations and challenges in section 2. We translate the observations into requirements for the terminology network. In section 3, we combine the requirements with our research to form a proposal for a *knowledge graph for heritage information.* We define the roles and tasks for the management and the use of this knowledge graph, in which all stakeholders are involved: the Network for Digital Heritage, the nodes, domain experts and collection managers. In section 4, we provide an overview of the services necessary to carry out the tasks efficiently. In the concluding sections we discuss the consequences and opportunities of the terminology network for aggregation and access to collections.

Versions

The first draft of this advisory paper was presented and discussed at the steering group meeting of 20th January 2016. The first concept was made available to the steering group and to external advisor Victor de Boer on February 8th. Following this feedback round, the final version was made available on February 18th.

We would like to thank all people interviewed for their precious collaboration and enthusiasm, and the steering group for their feedback. We would also like to thank Victor de Boer (VU), Chi Shing Chang (Spinque) and Arjen de Vries (Spinque) for their contribution to the present paper.

2 Requirements for the terminology network

The interviews mainly focused on (i) the current developments with regard to the integration within the domain nodes, (ii) the demands of the domain nodes and the heritage sector in general for the terminology network, and (iii) the intended manner in which the terminology network should be implemented to enable integrated access.

2.1 Current developments: integration in the cultural heritage sector

The cultural heritage sector undergoes thorough changes in terms of integration. Collections are being prepared for shared infrastructure in various heritage domains. Within a shared infrastructure, institutions can exchange information easily and provide access to information from various collections. Examples are the National Library Catalogue (NBC), i.e. the shared infrastructure for libraries, and the Archives Portal Europe (APE), i.e. the central aggregator for archival holdings.

With regard to controlled vocabularies, extensive sources are already being re-used by associated institutions in various heritage domains. Some examples thereof are the Common Thesaurus of Audio-Visual Archives (GTAA) used by the public broadcasters, the Heritage Thesaurus used for immoveable heritage, and AAT-NL and RKD Artist used by art and architecture institutions. Other examples are the thesauri of the Common Automated Cataloguing System Centre (GGC) for libraries and the registration of stakeholders with the National Archives.

Fig. 2 provides a schematic overview of integration within cultural heritage domains. Compared to the initial situation shown in Fig. 1, institutions of the heritage sector use the same vocabularies. Thanks to a common vocabulary, items can be found from different collections.

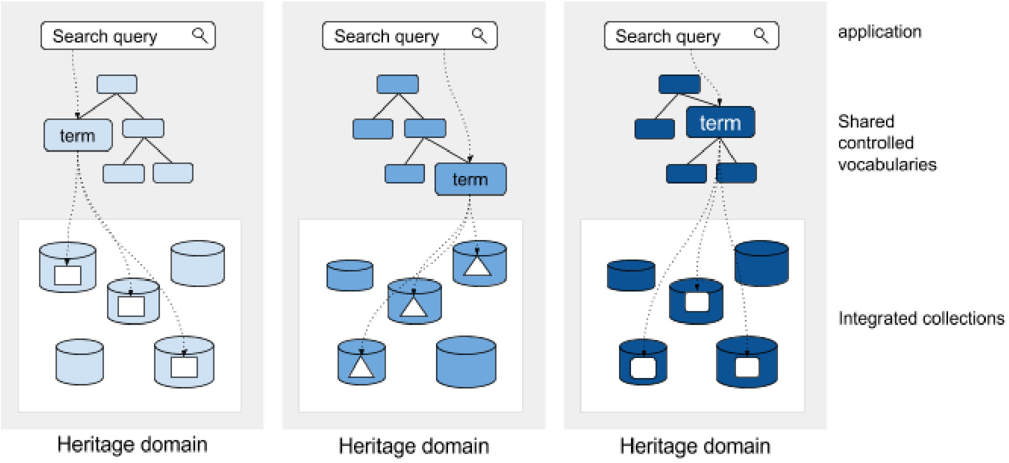

Fig. 2. Integration within the cultural heritage domains. Currently, collections are centrally organised in a particular heritage domain. Within a given domain, a common controlled vocabulary is used, so users can search across different collections from a given heritage domain.

The aim of the national strategy is not only to achieve integration within each heritage domain (*intra*-domain) but also between the different heritage domains (*inter*-domain). This way, collections can be made available at national level as a single body of information, so that, for example, all the works created by a particular person can be found together.

However, simply putting together collections is not sufficient in order to achieve uniform access to them. Because heritage institutions describe their collection items in various ways, our search queries cannot be unequivocally mapped to the underlying data from several collections.

There are two sorts of differences: (i) the *metadata scheme* (example: “maker” vs “author”) and (ii) the *terms used in de metadata sections* (example: “Johannes Vermeer” vs. “J. Vermeer”). In order to find all works produced by a particular person in the different collections, we must know which metadata schemes are used and how the person in question is described or represented in the metadata section in question.

Both within a particular heritage domain and between heritage domains the metadata schemes and terminology lists of institutions vary. Within a given domain, these issues can often be resolved by using one common metadata scheme and one (or very few) controlled vocabularies. This is the direction in which the heritage domains develops today, namely converging towards a homogenous system.But this solution is much harder to implement throughout different heritage domains.

The various domains are very heterogeneous and their daily practices and use of the collections differs. In order to integrate collections from various heritage domains, a solution is needed in which the various approaches and daily practice of working can co-exist. Furthermore, this opens the door to the possibility of also integrating the rich knowledge sources from outside the heritage sector.

- We consider the naturally developing integration of collections and infrastructure in the heritage domains as a positive trend and expect it to contribute to integration at the national level. It is indeed important to realise that developments at the national level also have an impact on the practice within the various cultural heritage domains. A sustainable solution requires the commitment of the domain nodes, but it is just as important that institutions are ready to work in a spirit geared towards integration.

2.2 Terminology network: distributed management, unified access

The alternative to an integration via one homogenous system is a flexible form of integration in which, when applicable, similarities and relationships are being recorded. In this paper, we explain how this approach can be applied to the integration of heritage information.
Inspired by initiatives in the domain of *Linked Data*, we call this procedure of establishing connections “*linking*”. Integration by linking can be applied to both metadata schemes and controled vocabularies such as terminology lists and thesauri.

Linking metadata schemes

Metadata schemes can be linked within the framework of a specific domain or as an umbrella scheme (Dublin Core, Europeana Data Model, schema.org). Schemes can also be linked directly to each other. For the Digital Collection for the Netherlands[1] and Europeana the integration of metada-

[1] http://digitalecollectie.nl/

ta schemes plays an important role. For example, Europeana maps the schemes of individual collections to the Europeana Data Model[2].

In this paper, we do not delve further into the integration of metadata schemes *for collections* but focus on the integration of controlled vocabulary. As we will see later on, scheme integration indeed also plays a role in the integration of *vocabularies*.

Linking controlled vocabularies

In the case of controlled vocabularies, linking consists in establishing connections between terms from different sources. The links between these terms thus form a terminology network that covers several sources. Managing each individual vocabulary body remains the mission of the initial suppliers. Creating and managing this terminology network thus requires a decentralized and distributed initiative. For external users, the network must provide uniform access to the collections. Linking vocabulary plays an important role in the technical implementation of this goal. At least as important is that the terminology network is presented as a unity to the external users. The distributed character of the network and its management are, at least initially, not relevant for professionals, researchers and the general public using it. Most search queries pertain to the "who", "what", "where" and when" of heritage information and rarely focus on a specific controlled vocabulary.

The most important requirements for setting up and managing a usable terminology network are the following principles:

- Heritage institutions use common public sources to describe their collection items. Institutions must ask themselves to which extent their internal controlled vocabulary is relevant for the public domain. Internal vocabularies disappear if they have no added value with regard to publicly available sources. If it is evidently relevant, the institution transforms the internal vocabulary into a new publicly available source or a part of an already existing public source.
- Everyone can contribute to the terminology network by providing a controlled vocabulary as publicly available source. The supplier is responsible for the quality of the vocabulary, its continuous availability and its integration with other sources. Within the Network for Digital Heritage, there are five domain nodes which can naturally offer public sources. Other heritage institutions can also supply their controlled vocabularies to the terminology network and make them publicly available and integrate them in other related sources of the network.
- Associations, foundations, and private individuals who do not have the ambition to play a role as supplier of controlled vocabulary, can bring in their specialist knowledge directly in the management of an existing supplier, or through a public environment such as Wikidata. The domain expert continues to contribute to the management of the terms with regard to content, but is under the responsibility of a new supplier.
- Public sources in the terminology network are linked. The source manager is responsible for establishing the links to other sources. The sources of the terminology network are also linked to other sources outside the cultural heritage sector where possible.
- The terminology network contains the information necessary to describe the "who", "what", "where" and "when" of collection items. A Simple Knowledge Organisation System (SKOS) is used for representing concepts (what). Other vocabularies are necessary to represent additional description features such as persons (who), places (where) and events (when).

[2] http://pro.europeana.eu/page/edm-documentation

2.3 Access to the terminology network for annotations

According to the vision of the nodes, internal non-public vocabularies will eventually be completely absorbed in open public vocabularies. The collection manager will then describe items with concepts from the terminology network. The requirements for supporting such an annotation process are as follows:

- A user can search several sources in the terminology network. Linked terms are shown as a single result. But the user has access to information about the sources in which the searched terms occur.
- A collection manager can annotate items directly in his or her collection management system, by adding concepts from the terminology network. The sources in which the collection manager will search can differ depending on the metadata field. The way in which results are shown and ranked can also be different depending on the metadata field.
- Institutions are not restricted to the vocabularies provided through the node of their domain. Controlled vocabularies can contain information originating both from within the area expertise of the node as well as from other fields outside of the node. One example of this is the geographical section of many Dutch cultural thesauri. Indeed, the information about locations is often more scarce than in specialised geographical sources such as the Historical Geothesaurus[3] or Geonames. Institutions of a specific heritage domain can also use these external sources.
- A collection manager must be able to propose new possible terms, for example if during the annotation process it appears that a particular term does not yet exist. The vocabulary manager may then accept, refuse or modify the proposed term.

2.4 Access to collections through the terminology network

One key target of the national strategy is to improve access to collections. This improved access starts with the integration of the data: at the level of the metadata schemes, by making the collections accessible in an interoperable manner, and at the level of collection descriptions, by integrating them through the terminology network. If these conditions are met, aggregation and the use of integrated data become accessible to everyone. The canonical example that is put forward time and again is the aggregation of collections into one national collection that is then made accessible via a portal.

Enriching data with external information furthermore offers heritage institutions the possibility to improve access to their own collection. In our opinion this is equally important. Instutions can make their collections more easily accessible by using enriched information from the terminology network and by adding information from other collections. Several heritage institutions in the Netherlands and in Belgium already use AAT: Openlucht Museum, Zuiderzeemuseum, Belasting en Douane Museum, Bijzondere Collecties Universiteit Leiden, Antwerpse Musea, Musea Oost-Vlaanderen and Erfgoedplus.be. Thanks to the fact that AAT is multilingual – unlike most other internal sources – users can search efficiently in different languages. Other examples: Providing access to items from other collections[4], or creating a website for an exhibition with items from various collections. The information originating from external sources such as Wikipedia can also be used through the terminology network to give an enriched presentation, for example with background information about a person. The integration of collections removes some obstacles that researchers and developers face in their analyses of large cultural datasets that are now only accessible to large labs such as Software Studies Initiative of Lev Manovich.[5]

[3] http://erfgoedenlocatie.nl/

[4] https://www.comsode.eu/index.php/2015/07/linked-open-images/

[5] http://manovich.net/index.php/projects/cultural-analytics-social-computing

The most important requirements for making collections accessible with the help of a terminology network are as follows:
A collection manager can make his or her collection available in an interoperable format (through the collection management system). This system includes the references to the terms from the network.

In a system in which several collections are aggregated, the terminology network is used to search and navigate through the different collections. For this purpose, the system must be able to handle both links between metadata schemes and links between terms.

An institution can use the rich information from the terminology network for its own search engine. Also, it can use other collections linked to the network for new forms of access. In this case, the local search engine must also be able to handle links between metadata schemes and links between terms.

3 The terminology network as a knowledge graph

In this section, we translate the requirements into a vision for a sustainable and user-friendly terminology network and into concrete roles and tasks that are needed to make this vision a reality.

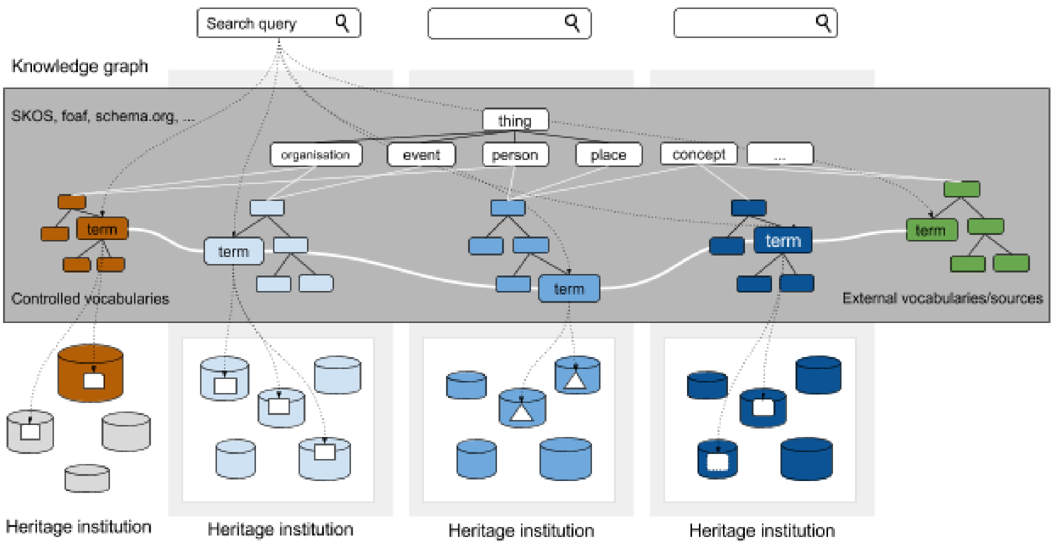

Fig. 3. Vision for a knowledge graph for heritage information. Controlled vocabularies within the heritage sector are linked to each other and to external sources. Collection managers can use different sources to describe their collection items.

The vision we want to put forward consists in an approach to the terminology network as *a knowledge graph for heritage information*. In this vision, *knowledge graph* refers to the initative of Google in which it gathers the knowledge of the world in a structured manner.[6] Google uses this knowledge graph in its search engine and other contexts. The knowledge graph is composed of information taken from various sources, among which Wikipedia. From the outside, the knowledge graph is accessible as an entity to people through the search engine and to computing machines through the API[7]. The uniformity that Google can realise in its knowledge graph is due to the centralised organisation of data and the possibility for Google to form an integrated entity on the basis of the different underlying sources, without interference by third parties. At the scheme level, information is structured using schema.org[8].

A knowledge graph for heritage information can be realised in a similar way. The heritage sector lacks a central authority that is able to integrate all knowledge from the sector.

So we propose to set up and manage the knowledge graph for heritage information in a distributed and decentralised way. Fig. 3 shows a schematic representation of this vision in relation to the existing heritage domains. The controlled vocabularies from the heritage sector, and where necessary also external sources such as Wikidata (and/or Wikipedia), play a key role. The sources are integrated into an entity by (horizontally) linking the terms with each other. The various types of concepts are integrated by (vertically) integrating the schemes of the different vocabularies.

[6] https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html
[7] https://developers.google.com/knowledge-graph/?hl=en
[8] http://schema.org/docs/full.html

We recommend to keep the use of schemes as open as possible. However, providing advice about the schemes already available and how to best implement these is recommended. In addition, existing links between schemes can be made available.

The danger of a distributed approach is that it can lead to uncontrolled growth so that the terminology network may no longer be used as an entity. Such an uncontrolled growth can be limited by a number of simple modalities, thereby enabling the development of an integrated network without a central authority.

We consider the following conditions for a decentralised knowledge graph for heritage information:

- Every institution can become a node and register a controlled vocabulary.
- The existing controlled vocabularies from the domain nodes can form the starting point.
- Nodes guarantee the public availability of controlled vocabulary and interoperability with other sources by integrating the respective terminology schemes.
- Managers of controlled vocabularies are responsible for the integration of other vocabularies in the network by aligning the different terms to each other.
- Collection managers describe their collections with terms from the network.
- This way, they are no longer limited to the sources available within a given heritage domain and they can use information from the network as a whole.
- Because everyone can use and link to sources considered to be most appropriate, we expect the importance of sources (or parts thereof) to become more evident. As a consequence, the managers will have to focus on the information within their domain of expertise and leave other aspects to the other experts within the network.

In addition to the already existing nodes within the Network for Digital Heritage various heritage institutions will have to take the role of node. Vincent de Keizer’s report gives an overview of controlled vocabularies in the cultural heritage sector and potential nodes. It is obvious that large institutions will play a major role. For example, the Netherlands Institute for Art History (RKD) manages several sources in the field of art, among which *RKDartists&*, RKDimages, Nederlandse AAT and Iconclass. Naturalis manages the Dutch Species Catalogue. Smaller institutions with particular expert knowledge of a specific domain can choose to become a node or to make their knowledge available through an existing node. Furhtermore, thematic initiatives such as the project of a visual thesaurus for online fashion heritage and the war sources network[9], can complement an existing node with particular knowledge about a given domain or lead to the creation of a new node. It is important to guarantee the sustainability of temporary projects and to organize the final responsibilities right from the start.

Roles and tasks

We identify four roles necessary to create a sustainable and user-friendly knowledge graph for the heritage sector. Fig. 4 shows these four roles, their respective tasks and relationships.

[9] http://www.oorlogsbronnen.nl/

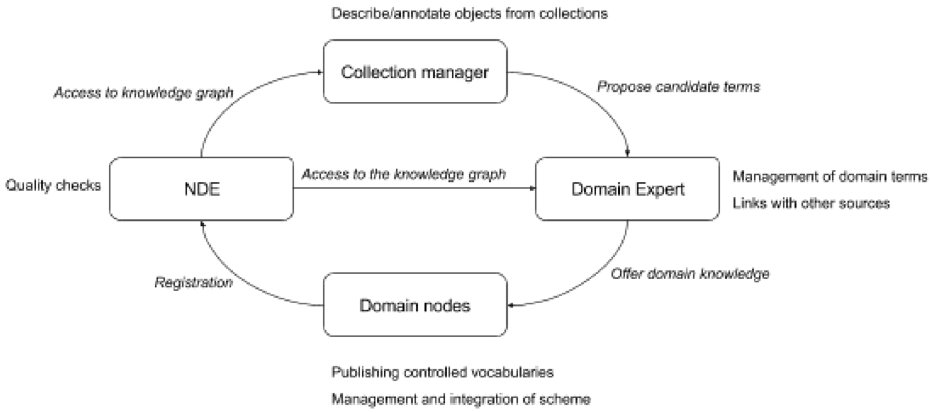
Describe/annotate objects from collections
Collection manager
Access to knowledge graph
Propose candidate terms
Quality checks
NDE
Access to the knowledge graph
Domain Expert
Management of domain terms
Links with other sources
Registration
Domain nodes
Offer domain knowledge
Publishing controlled vocabularies
Management and integration of scheme

Fig. 4 Roles necessary to maintain a sustainable knowledge graph for heritage information. The NDE makes the knowledge graph accessible for collection managers and domain experts. Collection managers use the knowledge graph as a tool to describe their collection items and propose new terms if they are not yet available. Domain experts manage terms and link them with other terms in the knowledge graph. The node publishes the controlled vocabularies and registers itself in the network.

Collection manager

Collection managers integrate their collections at national level by describing them with concepts from the knowledge graph (annotation). If it appears during the annotation process that concepts are necessary that are not yet available in the knowledge graph, the collection manager proposes them to a domain expert.

Domain expert

Domain experts are heritage institutions, associations, foundations or private individuals with a particular expert knowledge of a given domain. This knowledge can be used to complement an existing source of a node. The responsibility for the publication of the source lies with the node. The domain expert and the node coordinate the distributed management. The domain expert integrates its knowledge with other sources by linking concepts. A domain expert can also contribute knowledge through an open platform such as Wikidata. If a domain expert would rather manage their own knowledge, it can become a node itself and assume the related tasks.

Node

A node is responsible for publishing controlled vocabularies. The node manages the scheme of a controlled vocabulary and links it to other schemes. As far as content is concerned, the node coordinates itself with domain specialists. The node registers itself and its sources in the network.

Network for Digital Heritage

The Network for Digital Heritage enables nodes to register in the knowledge graph. Nodes are responsible for making controlled vocabularies available and for integrating them. The responsibility of the network is to monitor all of these operations. This means that the network must check if the sources made available by a node indeed meet the necessary quality requirements. One aspect of this quality control is to set up an overview of available sources and statistics on content, quality and use. The network furthermore plays the role of advisor to the nodes: How to make a controlled vocabulary accessible in an interoperable manner? Which schemes are available and which are relevant? Which other sources exist that could be linked and how can this be achieved?

Establishing roles

In the Network for Digital Heritage, a collection manager can play the role both of domain expert and of node. This is the case when a heritage institution decides to make an internal controlled vocabulary accessible to the public through the terminology network. In order to determine which of the two roles mentioned above is actually applicable, the heritage institution makes an assessment of the internal vocabulary in relation to the knowledge graph. Three cases can arise from this exercise:

1. The concepts of the internal vocabulary are already available in the knowledge graph and therefore have no real added value. The internal vocabulary is resigned and replaced by links to the knowledge graph.
2. The internal source defines a sub-domain of an existing source that has not been developed much or at all. The heritage institution then becomes domain expert within the existing node.Heritage institition and node together determine how the sources will be integrated.
3. There is no overlap with the knowledge graph, or there is overlap with the knowledge graph but the source provides a different (complementing) image to the knowledge graph. The collection manager makes the internal source available to the public, integrates it in the terminology network and is responsible for its management. In this case, the heritage institution also plays the roles of domain expert and node.

Heritage institutions can also collectively carry out the assessment of internal vocabularies. A fine example hereof is the project of a visual thesaurus for online fashion heritage. In this project, the internal vocabularies are first compared to each other to then the common overlap with the AAT is determined.

Work and realisation

An interesting aspect of setting up a knowledge graph for heritage information is that it is basically not self-evident where the boundaries lie for what can be considered as relevant heritage information. For example, we can ask ourselves whether or not collection items themselves must be part of the knowledge graph. The answer will depend on the type of information that is useful for the heritage sector. For publications, it can be useful to make a distinction between a “work” and the “realisation” of a work. When a library user is looking for a particular book, he or she searches for “the work” and is usually not interested in a specific edition of this book (the realisation). If he or she wants to borrow this work, a specific realisation of the work becomes relevant. Is a particular realisation available at this location? The user must be able to search for “the work” which in turn must be linked to its realisations. To enable this procedure, it is useful to make a source with all works available in the knowledge graph. There are similar examples in other domains, for instance the TV news as “programme series”, and The Thinker by Rodin as “the work”. Wikipedia, for instance, contains precisely this interpretation: https://nl.wikipedia.org/wiki/NOS_Journaal and https://nl.wikipedia.org/wiki/De_Denker.

4 Service concept

In the previous section we have identified the tasks necessary for managing and using the knowledge graph. This section will describe the services necessary to carry out these tasks. We only examine tasks with a technical aspect, not advisory and coordination tasks.

4.1 Describing collection items (annotating)

In collection management systems, annotation fields are usually linked to an internal source. In order to use an external knowledge graph containing heritage information, a collection management system can work in two ways:

(i) using an external API for searching terms, or
(ii) aggregating relevant sources and making them available in the system.

Within the heritage sector, both types of approaches are developed in different systems.

Beside the technical integration of a knowledge graph in a collection management system, new aspects will be reavealed in the course of a term search: To each heritage institution different sources are relevant. The ranking of the search results and the presentation of concepts may also be subject to different requirements that vary depending on heritage institution and annotation field. In this regard, the research into annotation with several thesauri, carried out by Hildebrand and Ossenbruggen at the Rijksmuseum[10], gives insight into the requirements and possible solutions. For heritage institutions to have an efficient setup, the collection management systems must offer the possibility to better adapt the search process and the presentation of results to the requirements of various heritage institutions. Furthermore, collection managers must be able to configure the search system for the terminology network at their discretion.

The creation and implementation of the knowledge graph by the institutions will require effort and time. Also, it will be necessary to adapt workflows and the way in which collection management systems are used. It will not be possible to use the knowledge graph immediately. It can be assumed that the internal vocabularies will be replaced progressively. Until then, institutions can already facilitate the integration of their collections by linking the internal vocabularies to the knowledge graph.

4.2 Proposing potential terms

If it appears during the annotation process that a concept is not yet available in the knowledge graph, a collection manager must be able to propose it as a potential term. This means that external users must be allowed to propose terms through the management system of a vocabulary.
The vocabulary manager must then assess this term. The proposer must be informed through the vocabulary management system about the decision taken, for example by receiving a notification. This implies that collection management systems must be fitted with additional features in order to (i) provide the possibility to furnish a potential term through an external system and (ii) to be able to get notified when a potential term has been integrated in the external system.

[10] http://oai.cwi.nl/oai/asset/13989/13989D.pdf

It is up to the collection manager to decide to which domain expert a potential term must be submitted. In the long term, it may become possible to develop a service that suggests (or decides itself) where a potential term must be integrated.

Domain expert

Domain terms management

Different systems for managing controlled vocabularies already exist. Domain terms are also often managed in collection management systems. These systems are usually only accessible internally at a given heritage institution – hence the necessity to have an export function to make the terminology list available to the public. Managing domain knowledge in a collective and distributed way can also be achieved by using existing systems such as Wikipedia and Wikidata. Managing terms also implies actions such as cleaning a controlled vocabulary, in order to avoid redundancy of terms or to standardise the format of names.

Linking controlled vocabularies

Domain experts must link their controlled vocabularies to other sources within the terminology network. For a solution to be user-friendly, the user must be able to have control over the quality of the search results. This is why fully automated solutions usually do not provide enough control.

These technologies must be combined with a means to analyse the results. For the sake of sustainability, it is also important for domain experts that once the term is linked, it remains linked even if the system is updated.

Linking to external sources can also be used to analyse internal sources. The links provide insight, for example, into the overlap between sources. On the other hand, that which is not linked is probably not available in the external source either. Thus, linking can be used to identify and analyse the strengths of internal vocabularies.

Existing tools do not yet support analysis of several sources at once. The lack of these tools becomes clear when considering projects such as the Visual Thesaurus for Online Fashion Heritage in which the process of comparing internal sources of heritage institutions requires a lot of manual labour.

Node

Publishing controlled vocabularies

It is crucial to share controlled vocabularies in an interoperable way. One obvious option is to share this data using Semantic Web standards of the World Wide Web consortium[11], such as RDF. Details about the use of such semantic standards are not covered by this paper.

Beside the format in which a vocabulary is shared, the way in which data is communicated is also important. One obvious way is to simply make the data available in the form of files. There are other initiatives however, that enable various form of access to the published data. For example, the Semantic Web community has come up with the principle of publishing via Linked Open Data[12], making

[11] https://www.w3.org/2001/sw/wiki/Main_Page
[12] http://linkeddata.org/guides-and-tutorials

the data accessible through a SPARQL endpoint[13] and Linked Data Fragments[14]. The “Open Archives Initiative Protocol for Metadata Harvesting” (OAI-PMH)[15] gives the possibility of harvesting periodical updates. Not only does OAI-PMH allow to access all data, but also, for example, the updates carried out on a given date.

Management and integration of schemes

The nodes manage the schemes in which domain experts can represent information. This information is different for places/persons and concepts/events. There are already various schemes for types such as persons and places. The Network for Digital Heritage will have to play a pioneering role, together with the existing nodes, in order to determine which schemes are relevant for the heritage sector.

Registration of nodes

Nodes register in the Network for Digital Heritage. The node indicates which sources can be found and where. The Open Data community has produced different platforms for registering datasets, for example CKAN[16].

Network for Digital Heritage

Quality control

The Network for Digital Heritage controls the quality of the sources proposed by the nodes. To do so, it must aggregate the content regularly. The specific quality aspects must be determined more in detail: for example availability of the source, syntactic accuracy, use of datatypes etc.

NDE also establishes which sources are available and how they are linked to each other. The resulting data is shared publicly and made accessible visually as a data cloud. In a more extended version, the user will be able to navigate through the schemes of sources. Which types of concepts are available and which are their characteristics and relationships?

Advanced services will be able to map out where future improvements may be possible: Which sources overlap but are not yet integrated? Which (sub-)domains are not yet available?

Access to the knowledge graph

In order to annotate a collection with the help of the terminology network, a collection manager must be able to search for concepts within the whole network. So a service similar to the API of the Google Knowledge Graph[17] becomes necessary for the terminology network. Also, institutions must have the possibility to adjust the configuration of the search system to their needs, i.e. to their collections and specific annotation sections.

In already existing vocabulary management systems, it is possible to search terms throughout the system. However, there is currently no umbrella service that can be used to search a knowledge graph for heritage information as a whole and that enables adaptation of the search system configuration to the requirements posed by the different collections and annotation sections.

[13] https://en.wikipedia.org/wiki/SPARQL
[14] http://linkeddatafragments.org/
[15] https://www.openarchives.org/OAI/openarchivesprotocol.html
[16] http://ckan.org/
[17] https://developers.google.com/knowledge-graph/

5 Final word

5.1 Implementation of a service concept

In this paper, we have carved out the contours of a service concept for a knowledge graph for cultural heritage information. For further development, we recommend the Network for Digital Heritage to support and enhance existing initiatives in the heritage sector. While doing so, it is advisable to connect to the goals of both the projects 'Usability' (work package 2) and 'Visibility' (work package 1). The final objective of the alignment layer is indeed to improve visibility. Examples of existing initiatives in which the role of the alignment layer can be adapted in practice are the visual thesaurus for online fashion heritage, the war sources network, aligning the AAT and the Cultural Heritage thesaurus.

5.2 Access to collections

Integration of heritage collections via a shared terminology network is a requirement for improving access to these collections. In addition, tools and/or services that use the terminology network are necessary in order to provide access to the collections. The exact requirements will depend on the specific use. Basic operations that are indispensable include:

- Searching for concepts from the knowledge graph
- Using links between concepts to communicate various interpretations in the knowledge graph to the user (see: disambiguation pages of Wikipedia)
- Searching and navigating to (collection) items (maybe from different collections) via concepts from the knowledge graph
- Using links between concepts to search via other terminology or language
- Getting an overview of the collections that have items regarding a particular concept
- Obtaining suggestions, for a given collection item (or set of items), of items from other collections via concepts from the knowledge graph

In addition to the possibilities to create uniform access, across collections, we also see significant added value for the heritage institutions themselves. Traditionally, museums, libraries and archives have had the role of guiding people through the physical buildings that are home to our cultural heritage. In an interesting statement, Peter Edsen[18] argues that museums must claim this social role on the Web as well. In order to achieve this, the development of new applications must be simplified considerably. Institutions but also nodes and associations must be able to experiment and quickly develop new forms of user experiences.

Large news platforms, as for instance developed by The Guardian, have been experimenting and strengthening their role on the Web for quite some time now. Although the goals of such platforms are more strongly motivated by commercial aspirations, cultural institutions can learn a lot from them. David Weinberger elaborates on how the development of various platforms within news organisations plays a key role in this context.[19] The platforms were initially only intended to simplify the re-use of the organisation's own content outside its premises: re-use in apps, hackathons and visualisations that would change the world. In retrospect, these platforms seem to have a considerable

[18] https://medium.com/tedx-experience/dark-matter-a6c7430d84d1

[19] http://shorensteincenter.org/open-news-platforms-david-weinberger/

added value for the news organisations themselves. They can now more easily produce their own applications, experiment with new forms of publishing and of course co-operate with interesting new parties.

It is thus particularly important for heritage institutions to share and integrate their raw data. In addition, the institutions must become user of these data themselves. Maybe they too will soon become the most important user of the knowledge graph for cultural heritage.