

White paper: A distributed network of heritage information - June 2017

by Enno Meijers (enno.meijers@kb.nl) and Sjors de Valk (sjors@sjorsdevalk.nl)

The Dutch Cultural Heritage sector is working together in a joint effort to improve the usability of its cultural heritage collection data. Our challenge is to develop a digital heritage infrastructure that overcomes the necessity of aggregating and post-process data. Instead we aim to realize a true distributed network of digital heritage information. This paper describes our approach for developing a new, cross-domain discovery infrastructure for the Dutch heritage collections. With this new infrastructure we expect to improve the usability of the collection data maintained by the heritage institutions. Implementing Linked Data principles in the collection registration systems is one of central building blocks of this approach. We urge the maintainers of the collections to align their data with formal Linked Data resources, like thesauri (people, places, periods, concepts) and to publish data as Linked Open Data.

The Dutch Digital Heritage Network (NDE) program is a national program aimed at increasing the social value of the collections maintained by the libraries, archives and museums in the Netherlands. The partners in the NDE network are the Ministry of Culture, the National Library, the National Archives, the Institute for Sound and Vision, the Cultural Heritage Agency and a number of Research Institutes for Dutch Culture and History. These partners will formalize their co-operation through the installment of a new organization that will be responsible for realizing a joint strategy program for the Dutch cultural heritage network. The goal is a distributed network build by the institutes and their stakeholders (including commercial parties), each contributing from their own perspective. The program consists of three layers with a functional division between the management of data collections ('sustainability'), facilities for connecting that data ('usability'), and applications for presentation and use of the data ('visibility').

Our work at the usability layer is focussed on the development of a lightweight cross-domain infrastructure that is build on a distributed architecture. The core functionality consists of a network of terms that references all common definitions for places, people, concepts, time periods. These terms are made accessible through an SKOS API that collection registration systems can implement in order to search for relevant terms when annotating their cultural heritage objects. As a result the URIs of the terms will be added to the object descriptions. The NDE program works on getting all relevant terminology sources available as Linked Data and provides facilities for term alignment and even support for building new thesauri. Several tools for this work (CultuurLink [1], PoolParty [2]) are being provided by the NDE network.

Having cultural heritage institutions publish their data as Linked Open Data with references to established definitions for places, persons, time periods and concepts is one part of the challenge. The other part is to provide means for browsing in a cross-domain, user centric fashion. Based on possible relevant URIs identified in the user query we want to be able to browse the available linked data in the cultural heritage network. In general the concept of 'browsable linked data' is still a challenging concept. Although Tim Berners-Lee describes the concept of Browsable Graphs and even states that statements which relate things in two

documents must be repeated in each, this is not a common practice in the Linked Data world [2]. If browsable Linked Data is offered then it is limited to the ‘follow your nose’ principle which is only based on using *forward links*. In order to really navigate in a bidirectional way through the Linked Open Data cloud, support for navigating using *back links* is needed as well. To our knowledge little research has been done so far on this topic [3], [4].

Most Linked Data projects make bidirectional navigation work by aggregating Linked Data dumps and loading them in a triplestore where both sides of the triples can be queried. Since our quest is developing a distributed network that avoids replication of data and building very large central infrastructures this direction is undesirable.

An alternative approach is supporting federated queries. In general this would mean that every institution in the network should provide a SPARQL endpoint, which is a big challenge for small organizations and leads to major performance issues. Even a lightweight solution as Linked Data Fragments, developed by Ruben Verborgh and colleagues [6], which we hope to implement in many institutions, still leaves us at the question which endpoints have relevant data for a specific user question. The Dutch Digital Heritage Network consists of about 1500 institutions that hold collections. Random querying all the endpoints in this network using Linked Data Fragments would be impractical and unrealistic. Even in this case a preselection of most relevant endpoints to query would be required.

These arguments have lead to our decision to build a (preferably distributed) registry that records the back links for all the terms used in the Digital Heritage Network. The registry contains the formal Linked Data definitions of all the institutions and a high level description of the datasets they maintain, similar to the general CKAN registries for Linked Data (like datahub.io). In addition to this, we will also record *object profiles* that describe the relations between the object in the collection and the term URIs used in the object description. This information provides the back links and makes it possible to navigate from a term URI to the objects that have a relation with this term.

For implementing the synchronization of these *object profiles* to the registry we are investigating the work of Herbert Van de Sompel, Sarven Capadisli and others on protocols like ResourceSync [7], Linked Data Notifications [8] and Webmention [9]. With this new approach we hope to move away from a traditional *repository centric* approach to a more *web centric* approach where optimizing the usability of resources in their source environment is the starting point. We are currently developing a Proof-of-Concept for the distributed network of digital heritage information and we demonstrate the first results in December 2017 at the SWIB conference.

[1]: <http://cultuurlink.beeldengeluid.nl>

[2]: <https://www.poolparty.biz/>

[3]: <https://www.w3.org/DesignIssues/LinkedData.html>

[4]: M. Stefanidakis et al., *Linking the (un)linked data through backlinks* (2012)

[5]: M. Salvadores et al. *Domain-Specific Backlinking Services in the Web of Data* (2010)

[6]: <http://linkeddatafragments.org/>

[7]: <http://www.openarchives.org/rs/1.1/resourcesync>

[8]: <https://www.w3.org/TR/Idn/>

[9]: <https://www.w3.org/TR/webmention/>