Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 137 (2018) 120–126

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

Aggregation of cultural heritage datasets through the Web of Data

Nuno Freire[a], Enno Meijers[b], René Voorburg[c], Antoine Isaac[d]

[a]INESC-ID, Lisbon, Portugal
[b]Dutch Digital Heritage Network, The Hague, The Netherlands
[c]National Library of The Netherlands, The Hague, The Netherlands
[d]Europeana Foundation, The Hague, The Netherlands

Abstract

The existence of many digital libraries, maintained by different organizations, brings challenges to the discoverability of cultural heritage (CH) resources. Metadata aggregation is an approach where centralized efforts like Europeana facilitate their discoverability by collecting the resource's metadata. Nowadays, CH institutions are increasingly applying technologies designed for the wider interoperability on the Web. In this context, we have identified the Schema.org vocabulary and linked data (LD) as potential technologies for innovating CH metadata aggregation. We present the results of an analysis using the case of the Europeana network of aggregators and data providers as basis. We have conducted a survey of the available linked data technology, and we defined a solution, which we have put into practice in a pilot implementation within the Europeana network. In this pilot, the National Library of The Netherlands fulfils the role of data provider, with the Dutch Digital Heritage Network, as national aggregator, supporting the provision of several datasets from the national library to Europeana. The metadata is published using LD practices, having Schema.org as the main vocabulary. The national library also implements all the necessary semantic web mechanisms, defined in our solution, for making the datasets discoverable and harvestable by Europeana. Our proposal involves the use of vocabularies for description of datasets, and their distributions, namely DCAT, VoID and Schema.org. Europeana implements the LD harvester side of the solution and applies it to harvest the Schema.org data from the national library.

Keywords: cultural heritage, aggregation, VoID, DCAT, Schema.org

10.1016/j.procs.2018.09.012

1. Introduction

In the World Wide Web, a very large number of online cultural heritage (CH) resources is made available through digital libraries websites. The discoverability of these resources through Internet search engines, and their re-use in other domains, are still underdeveloped. Many CH resources are not of a textual nature (e.g., images, video or sound) and those that are, often lack machine readable full-text for search engine indexing. They consist of digitized images where the application of optical character recognition (OCR) was not performed, due to lack of funding or a suitable OCR technology (e.g., for manuscripts or early printed materials). For discoverability, CH Institutions have always relied on the creation of data records describing the resources (metadata).

These metadata records are the basis for accessing and retrieving the resources through each institutional digital library, which are specifically built for retrieval of this kind of data. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability of the resources by potential users. Across institutions, the discoverability problem is addressed by an organizational architecture based on a central organization (a role often fulfilled by a CH institution, but not always), who approaches discoverability of the resources by collecting their associated metadata. The central organization has the possibility to further promote the usage of the resources by means that cannot be efficiently undertaken by each digital library in isolation. They typically provide Web portals that contain CH focused search engines, also specifically built for metadata.

The data aggregation technologies used within CH are not the same as for Internet search engines or the Web of Data. OAI-PMH [1] has been the embraced solution, since it is highly specialized in fulfilling the requirements for the aggregation of metadata datasets. However, the technological landscape around CH has changed. Nowadays, with the technological improvements accomplished by network communications, computational capacity, Internet search engines, and semantic data interoperability, the motivation for adopting OAI-PMH is disappearing.

In the last years, the CH domain has been able to create sustainable aggregation initiatives, with self-sustaining business models. Examples are Europeana, DPLA, DigitalNZ, Trove and Digital Library of India, which are collecting and providing access to the public digitized cultural assets from Europe, United States of America, New Zealand, Australia and India, respectively. However, the costs related to the implementation of the technical solution for aggregation are high for data providers. For these initiatives, reducing the effort required for data providers would bring more participants to their networks and lower the overall costs, therefore increasing the sustainability of the whole network [2]. In this context, if aggregators were able to re-use the published linked data (LD), data providers could benefit from several advantages. In particular, it would give them the following motivations:

- For those already publishing LD in their digital libraries, the process for sharing their data with CH aggregators would become extremely simple.
- For those that do not yet publish LD in use, implementing the technical requirements for CH aggregation based on LD, would be more rewarding, since wider interoperability with other domains than CH would come as a valuable extra benefit.

This paper presents the first conclusions and outcomes of a pilot experiment involving three members of the Europeana network of aggregators and data providers: The National Library of The Netherlands (KB), the Dutch Digital Heritage Network (NDE), and Europeana. We follow, in Section 2, by describing earlier and related work in from the CH domain involving metadata aggregation and LD. The methodology applied in pilot is presented in Section 3. Section 4 presents our analysis of the requirements for LD metadata aggregation in Europeana. Section 5 presents the survey of available LD technology with potential to fulfil our requirements. Section 6 presents our solution designed in light of the requirements and available technology. Section 7 concludes.

2. Related work

Although the use of LD in CH has been the focus of much research, most of published literature addresses mainly the aspect of the publication of LD [3, 4, 5, 6, 7] and do not fully address how the common aggregation approach of CH can be based on the existing published LD.

The most similar work to ours is that of the Research and Education Space project (RES). This project has finalized in 2017 but its results are still available. It has successfully aggregated a considerable number of LD resources from CH sources. The resulting aggregated dataset can be accessed online, but an evaluation of its aggregation procedures and results was not published. From the technical documentation available [8], we can see that RES managed to give significant steps in the specification of key tasks to enable the aggregation of LD. Some tasks however were not fully specified by the end of the project, and no further information has been published afterwards.

Generic technical solutions have been proposed by others for enabling aggregation of LD (for example [9]). However, a standards-based approach has not yet been put into practice within CH.

The work presented in this paper is done in the context of the research activities, being carried out within the Europeana Network, for improving the network's efficiency and sustainability. LD has been identified in our past work as one of the technical solutions with application potential [10]. The work described in this paper is part of a series of experiments addressing several Internet technologies for this purpose [11, 12].

Regarding CH metadata based in Schema.org, the KB has created and published a LD version of the Dutch National Bibliography with extensive use of the Schema.org vocabulary. Europeana has engaged in the research of Schema.org in several aspects related to CH. The most significant results of Europeana's research on Schema.org are in best practices for publishing of Schema.org CH metadata [13] and in evaluating its usage in CHIs for its aggregation use case [12].

3. Methodology

Having in mind our objective in innovating CH metadata aggregation networks with sustainable and low technical requirements, we have set up a group of CH institutions, with who we discussed viable solutions, re-use of existing knowledge in CH, and effort required for implementation and regular operation. The three roles in the organizational structure of the Europeana network, were represented. The KB fulfils the role of data provider. NDE, as national aggregator, supports the KB in the implementation of the process for the provision of the datasets to Europeana Foundation, the central aggregator.

Before our pilot, the KB had taken its initial steps in the publication of its datasets using LD practices, having Schema.org as the main vocabulary. For the pilot, KB will revise its usage of Schema.org to ensure compliance with the data requirements of Europeana. The KB also implements all the necessary mechanisms that are required for making the datasets harvestable (described in Section 6). Europeana harvests the datasets, analyses their compliance with its data requirements, and ingests them in its production dataset. To achieve this, Europeana implements a metadata harvester based on the agreed mechanism and applies it to harvest the Schema.org data of the KB.

The harvested Schema.org data is converted to the Europeana Data Model[1] (EDM) prior to ingestion into Europeana. In Europeana's aggregation process, EDM is the interoperability data model to deal with the data heterogeneity of CH data resources in Europe, and its definition follows the LD architecture.

4. Requirements to be addressed in the pilot

The solution adopted by the pilot must fulfil the same functional requirements as the current aggregation solution of Europeana, which is based on OAI-PMH and EDM. It is also required that some aspects of LD are supported, so that a standard LD solution for the Europeana network can be established. The solution must provide the following:

R1 - Data providers must be able to provide a LD resource of their dataset.

R2 - All data transmissions between data providers and Europeana must be built on standard technologies of the Web of data and LD.

R3 - Data providers must be able to transmit to Europeana machine readable licensing of their metadata. Data providers should be able to specify the licensing at the dataset level and also at the individual metadata record level.

R4 - Data providers must provide a machine-readable specification of how the dataset can be downloaded or harvested by Europeana. Two mechanisms may be used: RDF data dumps; or listings of the URI of the resources that are part of the dataset.

[1] Definition of the Europeana Data Model: http://pro.europeana.eu/edm-documentation

R5 - EDM compliant metadata must be made available by the data provider. Alternatively, Schema.org metadata maybe used, as long as after conversion to EDM, it complies with the EDM schema requirements.

5. Possible Linked Data Technologies

To find possible technological solutions for the requirements listed in the previous section, we have conducted a review of technologies that could be used with LD to allow data providers to comply with all the requirements.

The main open questions presented by the requirements involve the machine-readable description of the LD datasets and the technical details about how they can be harvested by Europeana. To address them, we reviewed several vocabularies for describing datasets. Such vocabularies may be able to address requirements R1, R3 and R4, and we analyzed those listed in Table 1, which also indicates the results of our analysis in light of our requirements.

Table 1. Vocabularies for describing datasets, dataset distributions, and resulting analysis regarding the requirements of the Europeana network,

Vocabulary	Description	R1	R3	R4
VoID - Vocabulary of Interlinked Datasets	“VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets. It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloging and archiving of datasets.”[2].	yes	yes	yes
DCAT – Data Catalogue Vocabulary	“DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. Publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites.”[3]	yes	yes	partial
Schema.org	The Schema.org vocabulary defines classes representing Datasets and their distribution. Although Schema.org has classes and properties that would address all requirements, the machine readability of some properties would be limited, since some of the required properties define their data type as Text.	yes	yes	partial
EDM Datasets Profile	This profile defines the elements used to represent datasets ingested by Europeana. The profile is mainly intended to be used to disseminate dataset level information via the Europeana API[4].	yes	no	no
ADMS - Asset Description Metadata Schema	“ADMS is a profile of DCAT, used to describe semantic assets (or just 'Assets'), defined as highly reusable metadata (e.g. xml schemata, generic data models) and reference data (e.g. code lists, taxonomies, dictionaries, vocabularies)”[5]. Since ADMS addresses a specific type of datasets, which is out of our scope, we considered that ADMS is not applicable to the Europeana use case.	N/A	N/A	N/A
RDF Data Cube Vocabulary	This vocabulary provides a means to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts, using RDF[6]. Since Data Cube addresses a specific type of datasets, which is out of our scope, we considered	N/A	N/A	N/A

[2] VoID - Vocabulary of Interlinked Datasets : https://www.w3.org/TR/void/

[3] DCAT – Data Catalogue Vocabulary: https://www.w3.org/TR/vocab-dcat/

[4] Europeana Dataset Profile: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/EDM_Dataset_Profile_042016.pdf

[5] Asset Description Metadata Schema (ADMS): https://www.w3.org/TR/vocab-adms/

[6] The RDF Data Cube Vocabulary: http://www.w3.org/TR/vocab-data-cube/

that Data Cube is not applicable to the Europeana use case.

Three of the analyzed vocabularies are suitable to fulfil the requirements for LD aggregation: VoID, DCAT, and Schema.org. These allow data providers to specify machine readable licenses and describe their datasets (requirements R1 and R3). The main aspects that distinguish them is their capability to support all the options for allowing data providers to provide the technical details about how their datasets can be downloaded or harvested (requirement R4).

The three vocabularies allow the specification of the distribution of the datasets as downloadable RDF files. VoID, however, is the only vocabulary that can be used to specify root resources for the dataset, which allows an entire dataset to be aggregated by crawling. A LD crawler may resolve the root resource(s) and follow the links to other URIs in the retrieved RDF response. By using this crawling method, data providers would not need to create downloadable distributions for their datasets, since these are often hard for data providers to maintain updated. Since only VoID enables crawling, we considered that requirement R4 is only partially fulfilled by DCAT and Schema.org. We have brought this issue into discussion within W3C's Data Exchange Working Group[7], which maintains DCAT, for consideration in future versions of DCAT.

Given the availability of three possible solutions, we evaluated the feasibility of allowing data providers to apply any of the vocabularies, in order to design a solution that would allow them to use any existing expertise that they may have in-house, on one of the vocabularies. Supporting several vocabularies may bring higher costs for Europeana for the operation of the aggregation network. We investigated the community involvement and contributions related with theses vocabularies and we identified that Europeana's implementation and operation of an LD harvester may be leveraged on the comprehensive alignments that exists between the vocabularies. Table 2 describes the most relevant alignments that we have identified.

Table 2. Some existing alignments between DCAT, Schema.org, and VoID.

Vocabularies aligned	Description
DCAT and Schema.org	One of the activities carried out by the W3C Dataset Exchange Working Group (DXWG), is the preparation of the next version of DCAT. DCAT v1.1 is currently in a draft stage and includes the alignment of DCAT with Schema.org. A provisional version of the alignment is available at: https://github.com/w3c/dxwg/blob/gh-pages/dcat/rdf/schema.ttl
DCAT, ADMS and VoID	This alignment was the original work that lead to the creation of the Schema.org Dataset class. It resulted from a collaboration around the DCAT, ADMS and VoID vocabularies. The details of the alignment are available at http://www.w3.org/wiki/WebSchemas/Datasets
DCAT and Schema.org	An alignment initiative from the W3C Spatial Data on the Web Working Group, that included also the alignment with ISO 19115 - 'Geographic information - Metadata'. This work was concluded in 2016, and is available at: https://webgate.ec.europa.eu/CITnet/stash/projects/ODCKAN/repos/dcat-ap-to-schema.org/browse

6. Design of the adopted solution for the pilot

The solution adopted for the pilot was to support all three vocabularies: VoID, DCAT and Schema.org. Data providers may describe their datasets using any of the vocabularies, and also use terms from more than vocabulary when necessary. To inform Europeana about the existence of a LD dataset, data providers communicate to Europeana the resolvable URI of the RDF resource for the dataset.

[7] https://www.w3.org/2017/dxwg/wiki/Main_Page

Europeana's LD harvester supports the interpretation of the three vocabularies, by implementing the alignments mentioned in Section 5. By interpreting the dataset RDF resource, it triggers the appropriate harvesting mechanism:

- Based on crawling through root resources – in this case, data providers use the property *rootResource* (VoID) pointing to a RDF resource that must contain *hasPart* properties (Dublin Core Terms) with the URI's of the metadata about CH objects.
- Based on downloading a distribution – in this case, data providers may use the classes and properties of any of the vocabularies, since all three support this mechanism. The downloadable distribution needs to be available as RDF files using one well known encoding for RDF.

The metadata about CH objects may be provided using either Schema.org or EDM. The LD harvester integrates the implementation of the conversion from Schema.org to EDM [12] and the datasets are processed for ingestion into Europeana in EDM, through the whole process.

Guidelines for supporting data providers to prepare their implementation were prepared by Europeana[8].

The LD harvester is being developed as open source. Currently, the software is still under active development, but its future outcomes and source code can be consulted online[9].

7. Conclusion and future work

The first phase of the pilot provided positive perspectives for innovating metadata aggregation in CH. We have identified and put into practice a solution with low implementation barriers and use of existing expertise in CH institutions. Although the pilot is conducted just with three partners, we believe it provides a strong support for further research and progress towards operational adoption of LD-based metadata aggregation. The final stage of the pilot will focus in evaluating the suitability of Schema.org for describing CH objects and supporting the Europeana data requirements for aggregation.

At this point of our work, we have identified a key problem to be addressed in future work: LD-based aggregation of very large datasets. We have already identified, as possible solutions, the HDT (Header, Dictionary, Triples) data structure for RDF [14] and the solution provided by Linked Data Fragments [9]. We expect to conduct in future work, an analysis of the adoption feasibility of these technologies by data providers, and conduct case studies.

Acknowledgments

This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by the European Commission under contract number 30-CE-0885387/00-80.

References

[1] Lagoze, C., H. van de Sompel, M.L. Nelson, and S. Warner. (2002) "The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0." Available from: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[2] Verwayen, Harry. (2017) "Business Plan 2017: 'Spreading the Word'", *Europeana Foundation.* Available online: https://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-business-plan-2017.pdf

[3] Simou, N.,Chortaras, A., Stamou, G., Kollias. S. (2017) "Enriching and Publishing Cultural Heritage as Linked Open Data", in: Ioannides M., Magnenat-Thalmann N., Papagiannakis G. (eds*) Mixed Reality and Gamification for Cultural Heritage* pp 201-223. Springer, Cham. (2017) doi:10.1007/978-3-319-49607-8_7

[4] Hyvönen, E. (2012) "Publishing and Using Cultural Heritage Linked Data on the Semantic Web", in: Ding, Y., Groth, P. (eds), *Synthesis Lectures on the Semantic Web: Theory and Technology*. 2012. doi: 10.2200/S00452ED1V01Y201210WBE003

[5] Jones, E., Seikel, M. (eds) (2016) "Linked Data for Cultural Heritage", *Facet Publishing*.

[8] https://github.com/nfreire/Open-Data-Acquisition-Framework/blob/master/opaf-documentation/SpecifyingLodDatasetForEuropeana.md

[9] https://github.com/nfreire/data-aggregation-lab

[6] Pedro A. Szekely, Craig A. Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E. Fink, Rachel Allen, Georgina Goodlander. (2013) "Connecting the Smithsonian American Art Museum to the Linked Data Cloud." The Semantic Web: Semantics and Big Data, pp. 593-607. doi:10.1007/978-3-642-38288-8_40

[7] Mauro Dragoni, Sara Tonelli, Giovanni Moretti. (2017) "A Knowledge Management Architecture for Digital Cultural Heritage.". Journal on Computing and Cultural Heritage (JOCCH) - Special Issue on Digital Infrastructure for Cultural Heritage, Part 2, vol. 10, issue 3. doi:10.1145/3012289

[8] BBC. (2016) "A guide to the Research & Education Space for contributors and developers", McRoberts, M. (eds). Available online: https://bbcarchdev.github.io/inside-acropolis/

[9] Sande, Miel Vander, Ruben Verborgh, Patrick Hochstenbach, and Herbert Van de Sompel (2018) "Towards sustainable publishing and querying of distributed Linked Data archives.", in Journal of Documentation, vol. 74, n. 1. doi: 10.1108/JD-03-2017-0040

[10] Freire, Nuno, Hugo Manguinhas, Antoine Isaac, Glen Robson, John B. Howard (2018) "Web technologies: a survey of their applicability to metadata aggregation in cultural heritage.", in: L. Chan, F. Loizides (eds.), *Information Services & Use Journal*, volume 37, issue 4, Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices.

[11] Freire, Nuno, G. Robson, J. B. Howard, H. Manguinhas, A. Isaac (2017) "Metadata Aggregation: Assessing the Application of IIIF and Sitemaps Within Cultural Heritage. ", in the proceedings of the International Conference on Theory and Practice of Digital Libraries 2017.

[12] Freire, Nuno, V. Charles, A. Isaac (2018) "Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata.", in the proceedings of the Extended Semantic Web Conference 2018.

[13] Wallis, R., A. Isaac, V. Charles, and H. Manguinhas (2017) "Recommendations for the application of Schema.org to aggregated Cultural Heritage metadata to increase relevance and visibility to search engines: the case of Europeana.", in Code4Lib Journal, Issue 36. ISSN 1940-5758

[14] Fernández, Javier D., Miguel A. Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias (2013) "Binary RDF Representation for Publication and Exchange (HDT).", in Web Semantics: Science, Services and Agents on the World Wide Web, *Elsevier.*