Availability of Cultural Heritage Structured Metadata in the World Wide Web

Nuno Freire, Pável Calado and Bruno Martins

We would like to acknowledge the supporting work by Antoine Isaac and Valentine Charles, from the Europeana Foundation, for their reviews and discussions regarding our work.

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.

Introduction

1 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 is still a challenge. Many CH resources are not of a textual nature (e.g., images, video or sound). Those that are textual, often lack machine readable full-text, of which search engines are highly dependent, because they consist of digitized images where the application of optical character recognition (OCR) was not performed, due to lack of funding or availability of a mature 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.

2 These descriptive records are the basis for accessing and retrieving the resources through each institutional digital library website, 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 and usage of the resources by potential users, making the adequate indexing of cultural heritage metadata in internet search engines even more relevant.

3 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). These organizations approach discoverability of the resources by collecting their associated metadata descriptive records. 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 this kind of data records [1].

4 In the particular domain of CH, the data aggregation technologies used are not the same as for Internet search engines. OAI-PMH [2] has been the embraced aggregation solution, since it is highly specialized in fulfilling the requirements for the aggregation of metadata datasets. However, the technological landscape around our domain 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 not as clear as it used to be in CH [3].

5 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 aggregation 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 [7]. In this context, if cultural heritage aggregators were able to re-use the technological solutions in use for indexing by Internet search engines, data providers could benefit from several advantages. In particular, it would give data providers the following motivations:

- For those already implementing these technologies in their digital libraries, the process for sharing their data with CH aggregators would become extremely simple.
- For those that do not yet have these technologies in use, implementing the technical requirements for CH aggregation would be more rewarding, since discoverability through Internet search engines would come as a valuable extra benefit.

6 This paper presents a study of the current application by CH data providers of technological solutions in use for making structured data (or metadata, in the CH context) available for re-use in the Internet. We investigated the use of both linked data and technologies related with indexing of resources by Internet search engines. We have conducted a harvesting experiment of the landing pages from websites of CH digital libraries that participate in Europeana, and collected statistics about the usage of these particular technologies. These technologies allow for representing structured data within HTML, or allow for structured data to be referred to by links within HTML or through HTTP headers. An analysis and discussion of the collected statistics is also presented.

7 We conclude with a discussion, based on the outcomes of this study, regarding future work for establishing a solution for CH aggregation based on the current CH scenario and the available technologies.

Related work

8 Although the use of linked data in CH has been the focus of much research, most of published literature addresses mainly the aspect of the publication of linked data [11][12] [13] and do not fully address how the common aggregation approach of CH can be based on the existing published CH linked data.

9 The most similar work to ours is that of the Dutch Digital Heritage Network (NDE) [9] and the Research and Education Space project[1] (RES). NDE is a Dutch national level program aiming to increase the social value of the collections maintained by the libraries, archives and museums in the Netherlands. NDE is still an ongoing project, and its initial proposals are based on specific APIs to enable data providers to centrally register the linked data URIs of their resources [10]. The current proposal of NDE, by being based in its own defined API, does not yet provide a solution purely based on linked data.

10 The Research and Education Space project has finalized in 2017 but its results are still available. It has successfully aggregated a considerable number of linked data resources from CH sources. The resulting aggregated dataset can be accessed online[2], but an evaluation of its aggregation procedures and results was not published. From the technical documentation available [19], we can see that RES managed to give significant steps in the specification of key tasks to enable the aggregation of linked data. Some tasks however were not fully specified by the end of the project, and no further information has been published afterwards.

11 Generic technical solutions have been proposed by others for enabling aggregation of linked data (for example [14]). However, a standards-based approach has not yet been put into practice within CH.

12 The work presented in this paper is done in the context of the research activities, being carried out within the Europeana Network[3], for improving the network's efficiency and sustainability [7]. Linked data has been identified in our past work as one of the technical solutions with application potential [1]. The work described in this paper is part of a series of experiments addressing several Internet technologies for this purpose [15][16].

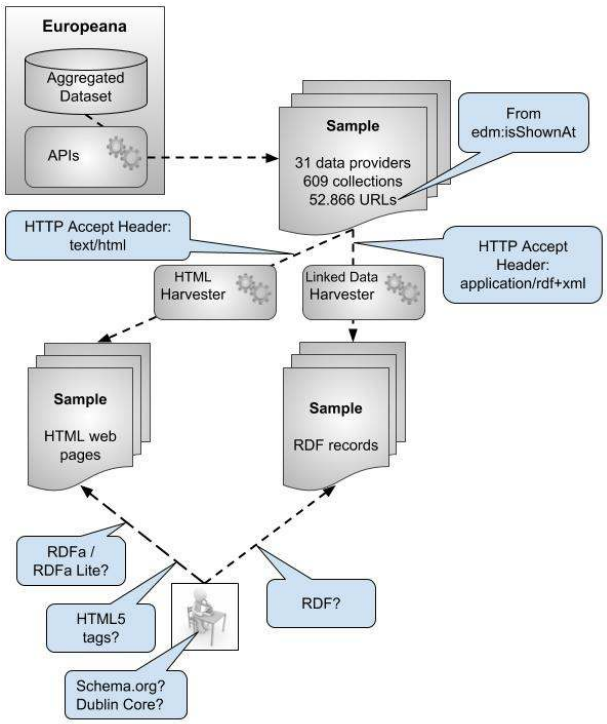
The experimental setup

13 In our harvesting experiment from the landing pages of resources from Europeana data providers, we have harvested samples from 31 different sources. In order to setup this test sample, we used the Europeana Search[4] and Record[5] APIs. The Search API was used first to discover the existing data providers of Europeana and their collections. Afterwards, on a second set of requests, Search API was used to discover a list of records from each collection. In subsequent requests, made on the Record API, we requested the complete metadata records of a sample of records per collection of each data provider. At most 100 records per collection were obtained. From these records we collected the URLs of the landing pages on the data providers' digital libraries. The metadata records were obtained in the Europeana Data Model (EDM) [8] format and the URLs were obtained from the EDM isShownAt[6] property of the ORE[7] Aggregation[8] element. In total, the sample comprehended URLs from 31 data providers, 609 collections and 52.866 resources.

14 We issued two requests on each of the 52.866 landing pages: one request for the human readable version in HTML and a second request for machine readable representation of the resource using HTTP content negotiation [5]. We then processed the responses and collected statistics on the usage of three possible ways that these digital libraries could be encoding the metadata descriptions of the cultural heritage objects: HTML5 meta tags, RDFa/RDFa lite and RDF data (in any of the commonly used serialization formats). For the analysis of the HTML5 meta tags, we have excluded the standard HTTP tags, since they are not meant to provide any descriptive data regarding the content of HTML pages.

15 An additional aspect is addressed in the experiment - the data model or namespaces of the structured data encoded in the landing pages. In particular, we are interested in gathering statistics regarding the use of two data models: Dublin Core Metadata Elements [4]; and Schema.org[9] (used by Google and several other companies). These data models are the most likely to be nowadays in use by CH institutions to represent the metadata of their resources.

Figure 1. The experimental setup



Results

16 The totals responses obtained from the requests issued to the sample of 52.866 URLs are shown in Table 1. None of the responses for linked data resulted in valid RDF. The most frequent response was the HTML page, instead of RDF, therefore hinting that HTTP content negotiation was not even supported. In some cases an error "Unsupported content-type" was received, and for some sporadic cases a JSON response was received, but it was not in a JSON-LD form, therefore, no RDF triples could be obtained from them. Our visual inspection of some of these cases detected that the JSON data was under a specific format, probably defined by a particular JSON API of a digital library system.

17 In the responses to the HTML requests, we detected a total of 25.276 HTML pages containing some form of structured data: 14.407 pages with HTML5 meta tags; and 10.869 with Schema.org data, encoded in RDFa, RDFa Lite or JSON-LD (Table 1).

Table 1. Structured data obtained from the requests issued to the sample of 52.866 URLs from Europeana providers

Linked data requests (content negotiation)	**HTML requests**	
RDF	**HTML with HTML5 meta tags**	**HTML with Schema.org (in RDFa, RDFa Lite or JSON-LD)**
0 (no valid RDF responses)	14407 out of 52866 27% (from 17 Europeana collections)	10869 out of 52866 20% (from 6 Europeana collections)

18 Table 2 summarizes the usage of the HTML5 meta tags. Whenever meta tags were present on the HTML pages, at least one of the standard HTML5 meta tags was in use. In some cases, meta tags with properties using prefixes were also present. Although none of the HTML pages specified the namespace of the prefixes in use (it can be done by using RDFa), some of the prefixes are well-known, and typically they refer to the following namespaces:

- "dc" – Dublin Core Metadata Element Set[10]
- "dcterms" - DCMI Metadata Terms[11]
- "og" – The Open Graph Protocol[12]

19 The prefixes "eprints" and "egms" prefixes where found as well, but we cannot be certain to which namespaces they refer.

Table 2. The rdf:type of Schema.org RDF resources present in the HTML pages

HTML5 meta tags		
Meta tag property prefix	**Number of HTML pages**	**Number of distinct Europeana providers**
HTML5 standard tags	14407	17
Dc	3783	7
Dcterms	1377	2
Og	791	3
egms	701	2
eprints	100	1

20 Regarding the use of Schema.org metadata, in Table 3 it is shown the URIs of classes in use, which gives an impression of what is being described there. Schema.org can be used to describe many aspects related to the HTML page and its content. Therefore although Schema.org data may be present in the HTML, it may not be describing the CH object. And, in fact, we observed the usage of instances of ListItem, BreadcrumbList, SearchAction, Website and ViewAction, in several cases, which indicates that the CH object metadata was not part of the existing Schema.org data.

Table 3. The rdf:type of Schema.org RDF resources present in the HTML pages

Schema.org Classes used	
Class URI	**Number of instances**
http://schema.org/Thing	9770
http://schema.org/ListItem	988
http://schema.org/Person	555
http://schema.org/BreadcrumbList	509
http://schema.org/SearchAction	282
http://schema.org/WebSite	282
http://schema.org/Organization	282
http://schema.org/VideoObject	229
http://schema.org/ImageObject	227
http://schema.org/Book	199
http://schema.org/ViewAction	197
http://schema.org/Article	127
http://schema.org/VisualArtwork	100

Conclusion and future work

21 The results of this experiment make it evident that in spite of the numerous activities, in cultural heritage, for making available linked data, reaching it through automated means based on the WWW (i.e. the Web of Documents) is no yet feasible. The digital object entry pages, whose links are sent to Europeana, could not be automatically linked to their respective linked data representations, since they did not support linked data through content negotiation, and the structured metadata we found encoded within the HTML pages was very limited, or even non-existent in the majority of cases.

22 In order to make use of CH linked data for metadata aggregations, less automated approaches need to be employed to discover, link and adapt the aggregation systems to each dataset of the participating CHI data sources (SPARQL end points, data dumps, etc.). Alternatively, aggregators may start to define the technical mechanisms for making linked data automatically discoverable, accessible and usable for aggregation.

23 Another aspect we also conclude from the experiment, is that it supports the beneficial value of CH aggregation initiatives, such as Europeana and DPLA, for promoting the discoverability of the CH objects through both the WWW and linked data. The activities of aggregators in the publication of open linked data, such as [17], are likely to be the most interoperable source of CH linked data currently available. The results of the experiment provide further motivation for the development of Europeana's activities towards Schema.org publication of its dataset and CH metadata in general [18].

24 The next steps of our work will be to survey technologies of the Semantic Web, linked data and vocabularies for the description of datasets. We will analyze these technologies in search for a solution that will enable the aggregation of linked data in fully automatized ways or requiring very little human intervention. Table 4 shows a list of those technologies that we have identified at this stage of our work.

Table 4. Technologies of the Semantic Web, linked data, and vocabularies for the description of datasets, which may enable the aggregation of linked data in fully automatized ways or with little human intervention

Technology	**Description**
Linked Data Platform [20]	*"Linked Data Platform (LDP) defines a set of rules for HTTP operations on web resources, some based on RDF, to provide an architecture for read-write Linked Data on the web"* [20].
VoID - Vocabulary of Interlinked Datasets [21]	*"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."* [21].
DCAT – Data Catalogue Vocabulary [22]	*"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." [22]*
Schema.org[13]	The Schema.org vocabulary defines classes representing Datasets[14] and their distribution[15].
EDM Datasets Profile [23]	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.

25

BIBLIOGRAPHY

References

[1] "Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1: Reference Description." (2012). *DCMI Recommendation*. Available from http://www.dublincore.org/documents/dces/

[2] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T. (1999). "Hypertext Transfer Protocol—HTTP/1.1." Available from: http://www.w3.org/Protocols/rfc2616/rfc2616.html

[3] Freire, N., Isaac, A., Robson, G., Howard, J.B., Manguinhas, H. (2018). "A survey of Web technology for metadata aggregation in cultural heritage." *Information Services & Use*, vol. 37, n. 4: 425–436. Available online: http://content.iospress.com/articles/information-services-and-use/isu859

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

[5] *Sitemaps XML format*. Available from http://www.sitemaps.org/protocol.html

[6] van de Sompel, H., Nelson, M.L. (2015). "Reminiscing About 15 Years of Interoperability Efforts." *D-Lib Magazine* vol. 21, n. 11/12. http://doi.org/10.1045/november2015-vandesompel

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

[8] "Definition of the Europeana Data Model v5.2.8." (2017). *Europeana Foundation*. http://pro.europeana.eu/edm-documentation

[9] Meijer, E., Valk, S. (2017). *A distributed network of heritage information. White paper*. Available online: http://github.com/netwerk-digitaal-erfgoed/general-documentation/blob/master/Whitepaper%20A%20distributed%20network%20of%20heritage%20information.md

[10] *Netwerk Digitaal Erfgoed. High level functional design*. (2017). Available online: http://github.com/netwerk-digitaal-erfgoed/high-level-design

[11] 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*, 201–223. Cham: Springer. http://doi.org/10.1007/978-3-319-49607-8_7

[12] 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*. http://doi.org/10.2200/S00452ED1V01Y201210WBE003

[13] Jones, E., Seikel, M. (Eds.). (2016). *Linked Data for Cultural Heritage*. Facet Publishing.

[14] Rietveld, L., Verborgh, R., Beek, W., Vander Sande, M., Schlobach, S. (2015). "Linked Data-as-a-Service: The Semantic Web Redeployed." In Gandon, F., Sabou M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (Eds.) *The Semantic Web. Latest Advances and New Domains.* ESWC 2015. Lecture Notes in Computer Science, vol 9088. Cham: Springer

[15] Freire, N., Robson, G., Howard, J. B., Manguinhas, H., Isaac, A. (2017). "Metadata Aggregation: Assessing the Application of IIIF and Sitemaps within Cultural Heritage." In *21st International Conference on Theory and Practice in Digital Libraries.*

[16] Freire, N., Charles, V., Isaac, A. (2018). "Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata." In P*roceedings of the Extended Semantic Web Conference 2018 (ESWC).*

[17] Charles, V., Freire, N., Isaac, A. (2014). "Links, languages and semantics: linked data approaches in The European Library and Europeana." *Linked Data in Libraries: Let's make it happen! IFLA 2014 Satellite Meeting on Linked Data in Libraries.* Available online: http://ifla2014-satdata.bnf.fr/pdf/iflalld2014_submission_Charles_Freire_Isaac.pdf

[18] Wallis, W., Isaac, A., Charles, V., Manguinhas, H. (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." *Code4Lib Journal*, 36. ISSN 1940-5758. http://journal.code4lib.org/articles/12330

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

[20] Speicher, S., Arwe, J., Malhotra, A. (2015). *Linked Data Platform 1.0. W3C Recommendation*. Available online: http://www.w3.org/TR/ldp/

[21] Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. (2011). "Describing Linked Datasets with the VoID Vocabulary." *W3C Interest Group Note*. Available online: http://www.w3.org/TR/void/

[22] Maali, F., Reikson, J. (Eds.). (2014). "Data Catalog Vocabulary (DCAT)." *W3C Recommendation*. Available online: http://www.w3.org/TR/vocab-dcat/

[23] "Europeana Dataset Profile." (2016). *Europeana Foundation.* Available online: http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/EDM_Dataset_Profile_042016.pdf

NOTES

1. http://bbcarchdev.github.io/res/
2. http://bbcarchdev.github.io/res/collections
3. The Europeana Network is a community of 1,700 experts with the shared mission to expand and improve access to Europe's digital cultural heritage, in the organization they work for and/or by contributing to shape Europeana's services.
4. http://pro.europeana.eu/resources/apis/search
5. http://pro.europeana.eu/resources/apis/record
6. http://www.europeana.eu/schemas/edm/isShownAt
7. ORE - Object Reuse and Exchange: Data Model
8. http://www.openarchives.org/ore/1.0/vocabulary#aggr
9. http://schema.org/docs/about.html
10. http://dublincore.org/documents/dces/
11. http://www.dublincore.org/documents/dcmi-terms/
12. http://ogp.me/

13. http://schema.org/
14. http://schema.org/Dataset
15. http://schema.org/DataDownload

ABSTRACTS

In the World Wide Web, a very large number of resources is made available through digital libraries. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability, sharing and reuse of the resources. A widely-used approach is metadata aggregation, where centralized efforts like Europeana facilitate the discoverability and use of the resources by collecting their associated metadata. The cultural heritage domain embraced the aggregation approach while, at the same time, the technological landscape kept evolving. Nowadays, cultural heritage institutions are increasingly applying technologies designed for the wider interoperability on the Web. This paper presents a study of the current application by cultural heritage data providers of technological solutions in use for making structured metadata available for re-use in the Internet. We investigated the use of both linked data and technologies related with indexing of resources by Internet search engines. We have conducted a harvesting experiment of the landing pages from websites of digital libraries that participate in Europeana, and collected statistics about the usage these particular technologies. These technologies allow for representing structured data within HTML, or for structured data to be referred to by links within HTML or through HTTP headers capabilities. We conclude with a discussion of future work for establishing a solution for cultural heritage aggregation based on the current situation and the available technologies.

INDEX

Keywords: metadata, cultural heritage, search engines, linked data, World Wide Web

AUTHORS

NUNO FREIRE

INESC-ID, Universidade de Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal
Tel. +351.213100300
Fax +351.213145843
nuno.freire@tecnico.ulisboa.pt
https://sites.google.com/site/nfreire
(corresponding author)

PÁVEL CALADO

INESC-ID, IST, Universidade de Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal

pavel.calado@tecnico.ulisboa.pt
http://web.ist.utl.pt/pavel.calado/

BRUNO MARTINS

INESC-ID, IST, Universidade de Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal
bruno.g.martins@tecnico.ulisboa.pt
http://web.tagus.ist.utl.pt/~bruno.martins/