Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices
L. Chan and F. Loizides (Eds.)

doi:10.3233/978-1-61499-769-6-235

Web Technologies: A Survey of Their Applicability to Metadata Aggregation in Cultural Heritage

Nuno FREIRE[a,1], Hugo MANGUINHAS[b], Antoine ISAAC[b], Glen ROBSON[c] and John Brooks HOWARD[d]
[a] *INESC-ID*
[b] *Europeana Foundation*
[c] *National Library of Wales*
[d] *University College Dublin*

Abstract. In the World Wide Web, a very large number of resources are made available through digital libraries. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability and usage of these resources by potential users. A widely-used approach is metadata aggregation, where a central organization takes the role of facilitating the discoverability and use 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. This paper focuses in the domain of cultural heritage, where OAI-PMH has been the embraced solution, since discovery of resources was only feasible if based on metadata instead of full-text. However, the technological landscape has changed. Nowadays, with the technological improvements accomplished by network communications, computational capacity, and Internet search engines, the motivation for adopting OAI-PMH is not as clear as it used to be. In this paper, we present the results of our initial analysis of available potential technologies, in particular, the following: IIIF (International Image Interoperability Framework); Webmention; Linked Data Notifications; Sitemaps; ResourceSync; Open Publication Distribution System (OPDS); and the Linked Data Platform.

Keywords. metadata, cultural heritage, linked data, web technology, standards

1. Introduction

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 and usage of the resources by potential interested users.

An often-used approach is metadata aggregation, where a central organization takes the role of facilitating the discovery and use of the resources by collecting their associated metadata. Based on these aggregated datasets of metadata, the central

[1] Corresponding author, Rua Alves Redol 9, 1000-029 Lisboa, Portugal; Email: nuno.freire@tecnico.ulisboa.pt.

organization (often called aggregator) can further promote the usage of the resources by means that cannot be efficiently undertaken by each digital library in isolation. This scenario is widely applied in the domain of cultural heritage, where the number of organizations with their own digital libraries is very large. In Europe, Europeana has the role of facilitating the usage of cultural heritage resources from and about Europe, and although many European cultural heritage organizations do not yet have a presence in Europeana, it already holds metadata of resources originating from more than 3,500 providers[2].

This domain is also characterized by users that often have very specific information needs, which cannot be easily fulfilled by the Internet search engines. The retrieval of resources based on metadata, in combination with the hypertext documents of the World Wide Web, has been a challenge that the search engines have not yet been able to provide an effective solution for, therefore the retrieval of cultural heritage resources via search engines is ineffective.

The technological approach to metadata aggregation has been mostly based on the OAI-PMH protocol, a technology initially designed in 1999. OAI-PMH was meant to address shortcomings in scholarly communication by providing a technical interoperability solution for discovery of e-prints, via metadata aggregation. The cultural heritage domain embraced the solution offered by OAI-PMH, however, the technological landscape around our domain has changed. Nowadays, cultural heritage organizations are increasingly applying technologies designed for the wider interoperability on the World Wide Web. Particularly relevant for our work are those related with the social web, the web of data, internet search engine optimization, and the IIIF (International Image Interoperability Framework).

In this paper, we present the first results of our work in attempting to rethink our technological approach for metadata aggregation, with the goal of finding a solution to make the continuous operation of the aggregation network more efficient and to lower the technical barriers for data providers to give their contribution to Europeana. This paper makes the following contribution to the digital libraries community:

- An analysis of requirements for metadata aggregation based on a large network of data providers – the Europeana Network.
- A functional analysis for innovative use of state of the art technologies.
- A real-world application experience of open standards, thus contributing for their future improvement.

The paper will describe, in Section 2, the technological approach to metadata aggregation most prevalent in cultural heritage. Specific requirements, which guided our technological survey, are presented in Section 3. The Web technologies that were analyzed are presented in Section 4, and Section 5 concludes and introduces potential options for future work.

2. Metadata Aggregation in Cultural Heritage – Past and Present

In the cultural heritage domain, the technological approach to metadata aggregation has been mostly based on the OAI-PMH protocol, a technology initially designed in 1999 [1]. OAI-PMH was originally meant to address shortcomings in scholarly

[2] http://statistics.europeana.eu/europeana (consulted on 4th of January 2017)

communication by providing a technical interoperability solution for discovery of e-prints, via metadata aggregation.

The cultural heritage domain embraced the solution offered by OAI-PMH, since discovery of resources was only feasible if based on metadata instead of full-text [2]. In Europe, OAI-PMH had one of its largest, and earliest, applications in The European Library [3], which aggregated digital collections and bibliographic catalogues from 48 national libraries. It was also the technological solution adopted by Europeana since its start, to aggregate metadata from its network of data providers and intermediary aggregators [4].

However, the technological landscape around our domain has changed. Nowadays, with the technological improvements accomplished by network communications, computational capacity, and Internet search engines, the discovery of resources, such as e-prints, is largely based on full-text processing, thus the newer technical advances, such as ResourceSync [5], are less focused on metadata. Within the cultural heritage domain metadata-based discovery remains the most widely adopted approach since a lot of material is not available as full-text. The adoption of OAI-PMH for this purpose is not as clear as it used to be, however. OAI-PMH was designed before the key founding concepts of the Web of Data [6]. By being centered on the concept of repository, instead of centering on the resources, the protocol is often misunderstood and its implementations fail, or are deployed with flaws that undermine its reliability [2]. Another important factor is that OAI-PMH predates REST [7]. Thus, it does not follow the REST principles, further bringing resistance and difficulties in its comprehension and implementation by developers in cultural heritage organizations.

An additional aspect relevant for our work, is that nowadays, cultural heritage organizations are increasingly applying technologies designed for wider interoperability on the World Wide Web. Particularly relevant are those related with Internet search engine optimization and the International Image Interoperability Framework [9]. Regardless of the metadata aggregation process for Europeana, cultural heritage institutions are already interested in developing their systems' capabilities in these areas. By exploring these technologies, the participation in Europeana of these institutions may become much less demanding and possibly even transparent.

The cultural heritage domain has some specific characteristics, which have heavily influenced how metadata aggregation has been conducted in the past. We consider the following to be the most influential:

- Several sub domains compose the cultural heritage domain: Libraries, Archives and Museums (the term LAM is often used to refer to the three sub domains).
- Interoperability of systems and data is scarce across sub-domains, but it is common within each sub-domain, both at the national and the international level.
- Each sub-domain applies its specific resource description practices and data models.
- All sub-domains embrace the adoption and definition of standards based solutions addressing description of resources, but to different extents. A long-time standardization tradition has existed in libraries, while this practice is more recent in archives and museums.

- Several of the adopted standards tend to be flexible towards data structure. Standards based on relational data models, for example, are rare in cultural heritage, while XML-based data models are common.
- Organizations typically have limited budgets to devote to information and communication technologies, thus the speed and extent of innovation and adoption of new technologies is slow.

In this environment, a common practice has been to aggregate metadata, under an agreed data model that allows the data heterogeneity between organizations and countries to be dealt with in a sustainable way. These data models typically address two main requirements:

- Retaining the semantics of the original data from the source providers
- Supporting the information needs of the services provided by the aggregator

These two requirements are typically addressed in a way that keeps the model complexity low, with the intention of simplifying the understanding of the model by all kinds of providers, and to allow for a low barrier of implementation of data conversion solutions, by both providers and aggregators.

Another relevant aspect of metadata aggregation is the sharing of the sets of metadata from the providing organizations to the aggregator. The metadata is transferred to the aggregator, but it continues to evolve at the data provider, thus the aggregator needs to periodically update its copy of the data. In this case, the needs for data sharing can be described as a cross-organizational data synchronization problem.

In the cultural heritage domain, OAI-PMH is the most well established solution to address the data synchronization problem. Since OAI-PMH is not restrictive in terms of the data model to be used, it allows the sharing of the metadata per the adopted data model of each aggregation case. The only restriction imposed by OAI-PMH is that the metadata must be represented in XML.

In the case of Europeana, the technological solutions around the Europeana Data Model (EDM) [8] have always been under continuous improvement. However, the solution for data synchronization based on OAI-PMH has not been reassessed since its adoption.

The Web Technologies, presented in the following sections, address mainly the data synchronization problem, since the common data model based on EDM is intended to remain in usage. EDM does not impose any obstacles in the choice of Web technologies for this purpose, the data synchronization can be addressed with a wide variety of technologies. This comes from EDM following the principles of the Web of Data, and that it can be serialized in XML and in RDF formats.

3. Requirements for Cross-Organizational Data Synchronization

The synchronization of data sources is a general problem, for which computer scientists have provided many possible solutions. The type of solution applicable to each case is greatly influenced by the requirements of the application scenario, mainly in terms of data consistency guarantees and synchronization latency.

We focus on the scenario of data synchronization across data sources from different organizations. We define the requirements for the solution by considering the characteristics of the cultural heritage domain, mentioned in the previous section, along

with some particularities of the metadata aggregation carried out in the Europeana network of data providers and aggregators.

The solution must allow an aggregator to collect structured metadata about the digital resources that a cultural heritage organization (the provider) wants to make available in Europeana. A solution should address the following requirements:

- The set of resources for aggregation is specified by the provider, and may comprehend all the resources of a digital library, or just a subset.
- The set of aggregated resources may evolve over time; therefore, the synchronization process must provide efficient mechanisms for incremental aggregation that will happen over time.
- The synchronization process between the provider and Europeana must be automatic and efficient, in terms of computation and network communication.
- The synchronization mechanism must be scalable to the level of the largest datasets nowadays available in Europeana, which are in the range of 2-5 million resources.
- The solution should be simple to adopt by data providers. One of the following aspects would make a solution simple to adopt:
    - It is based on technologies already in use by data providers;
    - It has very simple technical requirements for implementation;
    - Open source and free tools exist for deploying the solution.
- The solution may be more technologically challenging on the aggregator's side than on the data providers', since the aggregators are often better prepared to address more complex technical implementation issues of information systems.

In the context of the above requirements, the following section will present the Web technologies that we identified as possible solutions.

4. Web Technologies for Metadata Aggregation

Most of the technologies described in this section were designed for fulfilling the needs of general use cases, and are applicable across several domains. Some of these can completely fulfil the requirements of metadata aggregation, while others only do so partially, and need to be combined with other technologies. Not all technologies have been explored, in our work, to the same level of detail, but, in this section, we describe all those that we have identified as being applicable.

4.1. International Image Interoperability Framework

The International Image Interoperability Framework, commonly known as IIIF, is a family of specifications that were conceived to facilitate systematic reuse of image resources in digital image repositories maintained by cultural heritage organizations. It specifies several HTTP based web services [9] covering access to images, the presentation and structure of complex digital objects, composed of one or more images, and searching within their content.

IIIF strength resides in the presentation possibilities it provides for end-users. From the perspective of data acquisition, however, none of the IIIF APIs was specifically designed to support metadata aggregation. Nevertheless, within the output

given by the IIIF APIs, there may exist enough information to allow HTTP robots to crawl IIIF endpoints and harvest the links to the digital resources and associated metadata.

To study the feasibility of data acquisition via IIIF, several experiments and case studies have been undertaken, and are currently in progress. The early experiments revealed that IIIF contains all the necessary elements for automatic harvesting of metadata. Some of these elements are, however, not of mandatory implementation, thus they will not be available in many IIIF endpoints. The following elements of IIIF APIs must be provided by data providers, to enable Europeana to harvest:

- Structured metadata: the typical metadata available in the output of IIIF is intended for end-user presentation, thus it is unable to fulfil the requirements of ingestion in Europeana. This limitation may however be overcome by using the optional links (i.e. seeAlso) to structured metadata, as specified in IIIF. These enable crawlers to harvest metadata in any format provided, such as EDM, Dublin Core, etc.
- IIIF Collection indicating the resources for Europeana: In IIIF, it is not required that the endpoint implements a mechanism to make publicly known all the digital objects that it makes available. However, such mechanism may be implemented, and, optionally, the IIIF provider may implement a IIIF Collection that lists the digital objects it holds, or just those intended for delivery to Europeana. By making this collection known to Europeana, all the digital objects referenced in the collection can be crawled, and their metadata harvested by Europeana.

There is one piece of information that IIIF does not provide, which is the modification timestamp of the digital objects. This aspect has an impact in the efficiency of the harvesting process, but only becomes relevant in very large collections, with sizes in the hundreds of thousands of digital objects. In the typical size of the collections delivered to Europeana, within the thousands or tens of thousands, the loss in efficiency is not significant nowadays, due to high availability of bandwidth and computational capacity.

To overcome this issue of harvesting efficiency in large collections, other technologies may be used in conjunction with IIIF. Examples are Sitemaps, HTTP Headers, and notification protocols, such as Webmention and Linked Data Notifications, which we are also being evaluated in our work and are described in this document. This issue of harvesting efficiency has been brought to the attention of the IIIF community, and we are engaged in the discussions for achieving a standard mechanism, or recommendations, which will address it within the IIIF community.

The results so far indicate that data acquisition via IIIF is feasible, and presents little technological barriers for data providers that already have an IIIF solution in place for their own purposes. In the Europeana side, once a IIIF crawler tool is integrated with its aggregation management system, ingestion of IIIF data sources can be carried out under the same process of nowadays.

4.2. Webmention

Webmention is a technology that addresses the general problem of allowing Web authors to obtain notifications when other authors link to one of their documents [10].

Webmention is currently published at W3C as a First Public Working Draft. We could not accurately determine how widely adopted Webmention is nowadays, but many resources can be found in the World Wide Web, from software implementations, running services, and many discussions on its use.

The notification mechanisms provided by Webmention, can be used to mediate the communication between the systems of aggregators and the data providers. Webmention presents the following positive aspects:

- A very simple technological solution;
- Any of the parties may initiate the exchange of information.

There are, however, some negative points regarding Webmention:

- No deployments of Webmention are known to exist in CH institutions;
- The notifications do not allow data to be transmitted, so it must be complemented with other technology, such as the example of linked data, which is described further ahead in this section;
- The notifications may lack semantic meaning (e.g. type of notifications) required for some aggregation operations;
- The application of Webmention, for metadata aggregation, diverges somewhat from what Webmention was designed for. If Europeana uses it for this purpose, further elaboration of specifications will be necessary to define how Webmention is meant to be used.

Due to the lack of a mechanism to transmit data in Webmention notifications, we see its application only in combination with other technologies. For example, in combination with existing linked open data (LOD) that data providers already have in place. Webmention would allow data providers to indicate to Europeana, which resources from their LOD dataset should be aggregated by Europeana.

Webmention could also be applied in a similar way to aggregate metadata from IIIF endpoints. The underlying approach may be the same as for LOD. But in this case, the notifications sent by the data providers to Europeana, would contain links to IIIF resources (manifests), and Europeana would use a IIIF crawler to harvest the metadata from the IIIF endpoint.

4.3. Linked Data Notifications

Linked Data Notifications [11] (LDN) is similar in functionality to Webmention, but it is built having the Web of Data in mind, while Webmention is focused in the Web of Documents. LDN is being designed on top of the W3C's Linked Data Platform (see below), and its notifications have richer semantics than the simple notifications of Webmention. Another promising aspect of LDN is that the notifications may carry data, thus allowing for a more straightforward way of fulfilling metadata aggregation than Webmention. We engaged with the LDN editorial group, and are currently providing feedback to the LDN specifications, considering the metadata aggregation use case.

4.4. Sitemaps

Sitemaps [12] allow webmasters to inform search engines about pages on their sites that are available for crawling by search engine's robots. A Sitemap is an XML file that

lists URLs of the pages within a website along with additional metadata about each URL (i.e., when it was last updated, how often it usually changes, and how important it is, relative to other URLs within the same site) so that search engines can more efficiently crawl the site. Sitemaps is a widely-adopted technology, supported by all major search engines. Many content management systems support Sitemaps out-of-the-box, and Sitemaps are simple enough to be manually built by webmasters when necessary.

Considering the application of Sitemaps in the context of Europeana, for data acquisition, it presents the following positive points:

- A simple technology with low barriers for implementation, even for small organizations.
- Already in use in several cultural heritage organizations, where it is applied for search engine optimization of their websites and digital libraries.
- It is extensible; thus, it can be adapted to Europeana specific requirements. For example, Google has Sitemap extensions for images and for videos, each one defining a set of metadata elements for its media type.

A Sitemap is an XML file, which is prepared per the Sitemap Protocol [12]. In digital libraries, Sitemaps typically contain all the links to the landing pages of the digital objects within the digital library.

These kinds of Sitemaps are widely used, thus already existing Sitemaps could be used by Europeana for metadata aggregation, using a WebCrawler such as those used by Internet search engines. Starting by following the links in a Sitemap, and processing structured data within HTML (e.g. microdata, Schema.org, linked data available by content negotiation), an Europeana Crawler may discover the digital cultural heritage objects, as well as metadata.

Besides its typical use for Internet crawlers, Sitemaps may also be deployed by Europeana and data providers in conjunction with other technologies, which would allow for simple ways of sharing data. For example, Sitemaps could be made available by data providers, in order to inform Europeana of the digital objects to be aggregated and when they are updated.

Sitemaps, present two clear benefits: a very low technological barrier, and data providing organizations often have in-house knowledge about XML and/or Sitemaps. Sitemaps are a key technology applied for Internet search engine optimization, thus it is already in use within data providers' websites and digital libraries for making their resources discoverable in Internet search engines. Providing metadata to Europeana by using Sitemaps would substantially reduce the implementation effort needed by data providers.

4.5. ResourceSync

ResourceSync [5] is a NISO standard that enables third-party systems to remain synchronized with a data provider's evolving digital objects, supporting both metadata and content. ResourceSync is based on the Sitemaps protocol and introduces extensions that enable its functionality for accurately and efficiently synchronizing the content of digital objects. Additionally, to Sitemap's capabilities, it allows data sources to:

- specify groups of resources, instead of each one individually.

- specify alternative ways to download the resources, as for example, as a bundle in a zip file.
- specify what has changed at a time.
- specify alternative ways to download just a set of changes
- link resources to metadata that describes the resources
- link to older versions of resources
- specifying alternative download mechanisms, such as alternative mirrors.

This detailed synchronization information provided by ResourceSync allows for much more efficient ways of keeping resources synchronized between a source and a destination.

The extra functionality of ResourceSync over Sitemaps, also increases the technical barriers for its adoption. At the time of writing of this document, we have not yet been able to locate a case of ResourceSync deployment in the cultural heritage domain. Most applications of ResourceSync are in grey literature repositories, which are usually out of scope of cultural heritage.

Since the current focus of Europeana is in acquisition of metadata, ResourceSync may offer more than is necessary, and be an unnecessary challenge for implementation by data providers. Still, ResourceSync is an important technology to follow, particularly as the aggregation of content as well as metadata is starting to gain more attention within the Europeana Network.

4.6. Open Publication Distribution System

Open Publication Distribution System (OPDS) is a syndication format for digital publications which enables the aggregation, distribution, and discovery of books, journals, and other digital content by any user, from any source, in any digital format, on any device. The OPDS Catalogs specification [13] is based on the Atom syndication format and prioritizes simplicity. OPDS usage can be found in eBook reading systems, publishers, and distributors. Publishers and libraries have been early adopters of OPDS. We could not yet determine how widely used OPDS is within the Europeana network.

4.7. Linked Data Platform

Linked Data Platform [14] specifies the use of HTTP and RDF techniques for accessing and manipulating resources exposed as Linked Data [6]. Several cultural heritage institutions publish as linked data the metadata regarding their resources. Although these data sources can be accessed and processed for aggregation, they are not available in a uniform and standard way. This requires a lot of manual effort for aggregators to processing the data, presenting a serious obstacle to an efficient and sustainable aggregation process. Within the many aspects specified by the Linked Data Platform, some provide the necessary standardization for an efficient aggregation based on linked data sources.

5. Conclusion

In conclusion, several technological solutions from the Web are available and look promising for simplifying the implementation of the metadata aggregation scenario in

cultural heritage. The next steps of this work will aim to assess the actual usage and existing knowledge of these technologies, within the cultural heritage institutions. Future work, on the technical software side, will address how these technologies may be used for designing crawling robots that aggregate the metadata. We expect that with crawling algorithms, which make use of Web technologies, the technical barriers and operational costs may be lowered, leading to more sustainable metadata aggregation networks.

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 the Connecting Europe Facility, telecommunications sector, grant agreement number CEF-TC-2015-1-01.

References

[1] Lagoze C, van de Sompel H, Nelson ML, and Warner S. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. 2002. Available from: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[2] van de Sompel H, Nelson ML. Reminiscing About 15 Years of Interoperability Efforts. D-Lib Magazine. vol. 21, n. 11/12. 2015. doi:10.1045/november2015-vandesompel
[3] van Veen T, Oldroyd B. Search and Retrieval in The European Library: A New Approach. D-Lib Magazine, vol. 10, n. 2. 2004. ISSN 1082-9873.
[4] Pedrosa G, Georg P, Concordia C, Aloia N. Europeana OAI-PMH Infrastructure. Project Europeana Connect deliverable D5.3.1. 2010.
[5] National Information Standards Organization. ResourceSync Framework Specification. 2014. Available from: http://www.niso.org/apps/group_public/download.php/12904/z39-99-2014_resourcesync.pdf
[6] Berners-Lee T. Linked Data Design Issues. W3C-Internal Document. 2006 Available from: http://www.w3.org/DesignIssues/LinkedData.html
[7] Richardson L, Ruby S. Restful Web Services. O'Reilly. 2007.
[8] Europeana v1.0, The EDM Definition V5.2.7. Online from: http://pro.europeana.eu/web/guest/edm-documentation
[9] Stuart S, Sanderson R and Cramer T. The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. Archiving 2015, 2015. Available from: http://purl.stanford.edu/df650pk4327
[10] Parecki A (ed.). Webmention. W3C Candidate Recommendation, 2016. Available from: https://www.w3.org/TR/webmention/
[11] Capadisli S, Guy A (eds.), "Linked Data Notifications", W3C Working Draft, 2016. Available from: https://www.w3.org/TR/ldn/
[12] Sitemaps XML format. Available from https://www.sitemaps.org/protocol.html
[13] The openpub community. OPDS Catalog 1.1 specification. (2011). Available online: http://opds-spec.org/specs/opds-catalog-1-1
[14] Speicher S, Arwe J, Malhotra A. Linked Data Platform 1.0. W3C Recommendation, 2015. Available from: https://www.w3.org/TR/ldp/