Europeana DSI 2 – Access to Digital Resources of European Heritage

DELIVERABLE

D6.7: Report on LOD and alternative data acquisition mechanisms for Europeana

Revision	Draft for internal project review
Date of submission	
Author(s)	Nuno Freire
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
0.1	01.08.2017	Nuno Freire	INESC-ID, Europeana Foundation	First draft
0.2	08.08.2017	Nuno Freire	INESC-ID, Europeana Foundation	Revised after first round of comments
0.9	14.08.2017	Nuno Freire	INESC-ID, Europeana Foundation	Revised after a round of comments. Ready for internal review.

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

‘Europeana DSI is co-financed by the European Union's Connecting Europe Facility’

Executive Summary

DSI-2 included activities on rethinking Europeana's technological approach for metadata aggregation, with the aim to make the operation of the aggregation network more efficient and lower the technical barriers for data providers to contribute to Europeana. This document presents the key outcomes and conclusions of the work in DSI-2 subtask 6.9.3, which included technological surveys, prototyping and case studies with data providers and related DSI tasks innovating the aggregation network. Note that our work merely complements other DSI efforts to further develop and innovates Europeana's aggregation process, namely Metis and Operations Direct. We do not make any assumption on the organization of the aggregation process, especially on whether the data is 'pushed' by Cultural Heritage Institution (CHIs) to Europeana or pulled by Europeana from CHIs or data aggregators. We rather explore components, which could be used by either framework - some of our proposals have actually been already tested or included in them.

Since its very beginning, Europeana's approach for supporting the discovery of cultural heritage (CH) resources has been metadata by aggregation, based on the OAI-PMH protocol, a technology initially designed in 1999. However CHIs are increasingly applying technologies designed for wider interoperability on the World Wide Web. Particularly relevant are those related with search engine optimization, social networks, and the International Image Interoperability Framework (IIIF). Regardless of the metadata aggregation process for Europeana, CHIs are already interested in developing their systems' capabilities in these areas. If Europeana also uses these technologies, their participation in Europeana may become less demanding.

We have identified a variety of technologies that could be applied, and among these, we consider three technologies that are the most suitable as a first step: IIIF, Sitemaps and Schema.org.

IIIF is getting increasing traction in CH. Moreover, it is a community developed, open framework, thus our requirements for metadata aggregation may be incorporated into future versions of IIIF. Our case studies 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. We chose Sitemaps because of its wide usage within the Europeana data providers. In addition, Sitemaps is a very simple technological solution, with a very low implementation barrier to those providers that are not currently using it. Schema.org is complementary to IIIF and Sitemaps, since it addresses the area of data modelling and representation.

We also consider two additional technologies: ResourceSync and Linked Data Notifications. ResourceSync presents itself as the best technical solution for aggregation of very large datasets and content. Notification mechanisms, such as Linked Data Notifications or WebSub (used in ResourceSync), would enable a rich exchange of data/information between Europeana and data providers that could increase the value given back to data providers from Europeana. These two technologies are less widely applicable since they have higher implementation costs for data providers, but they may enable Europeana to conceive new discovery services for CH.

In conclusion, our work has shown that several technological solutions are available to make the Europeana aggregation network more efficient and with lower barriers for data providers. While some solutions still require further research, some can be adopted in the short term. The adoption of an initial solution is mainly dependent on aligning the technical solution with the business objectives of Europeana for the coming years, then in establishing a best practice within the network, and in equipping Europeana with the necessary software tools and internal workflows.

Table of Contents

1. Introduction

In the World Wide Web, a very large number of cultural heritage (CH) 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 the resources by potentially interested users.

In Europe, Europeana has the role of facilitating the usage of CH resources from and about Europe, and although many European CH Institutions (CHIs) do not yet have a presence in Europeana, it already holds metadata from over 3,500 providers[1].

Since its very beginning, Europeana's approach for supporting the discovery of CH resources has been metadata aggregation. Europeana, acting as a central organization, collects the metadata associated with the CH resources. Based on these aggregated datasets of metadata, Europeana is in a position to promote the usage of the resources by means that cannot be efficiently undertaken by each institution in isolation.

DSI-2 subtask 6.9.3 is one of the ongoing lines of work that aim to contribute to Europeana's strategy to "*make it easy and rewarding for Cultural Heritage Institutions to share high-quality content*"[2]. Its contribution focus on the technological aspects of supporting Europeana's Network of metadata aggregation, aiming to make it more efficient and lower the technical barriers for data providers to contribute to Europeana.

Our approach in this task was to conduct the following general activities:

- Technological surveys of the state of the art;
- Case studies to identify requirements with data providers, the Data Partners Services team, and related tasks of DSI-2[3];
- Prototyping and testing technological solutions to evaluate their feasibility and acquire deeper knowledge.

This report starts by describing the context and scope of this task within the varied aspects of innovating Europeana's aggregation network. An analysis of the key aspects of the CH domain and its high level requirements for aggregation is presented in Section 3. The report follows with the description of the major technologies that we have analysed. The main results are presented in Section 5, with a general view of the key positive and negative aspects of the technologies and the stage of investigation that our investigation reached in each technology. Section 6 presents a more focused analysis and discussion of those technologies that we considered the most promising for the Europeana aggregation network, along with a discussion of the key tasks required for their adoption. Section 7 concludes.

2. Data acquisition - scope of the task

In the CH domain, the technological approach to metadata aggregation has been mostly based on the OAI-PMH protocol, a technology initially designed in 1999 [Lagoze et al. 2002]. OAI-PMH was originally meant to address shortcomings in scholarly communication by providing a technical interoperability solution for discovery of e-prints, via metadata aggregation.

[1] Source: http://statistics.europeana.eu/europeana [consulted on 27th of June 2017]

[2] Europeana Strategy 2015-2020 available at http://strategy2020.europeana.eu/

[3] Our work relates with Task 1.2 - Innovate the aggregation infrastructure for the Europeana DSI, and Task 6.10 - Prototype innovative technologies to empower collection owners to publish collections via the Europeana platform services.

The CH domain embraced the OAI-PMH solution as discovery of resources was only feasible if based on metadata instead of full-text [van de Sompel&Nelson 2015]. In Europe, OAI-PMH had one of its largest, and earliest, applications in The European Library [van Veen&Oldroyd 2004], 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 [Pedrosa et al. 2010].

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 [NISO 2008], are less focused on metadata. Within the CH domain, metadata-based discovery remains the most widely adopted approach since a lot of material is not available as full-text. The motivation for adopting 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 [Berners-Lee 2006]. By being centred on the concept of repository, instead of centring on resources themselves, the protocol is often misunderstood and its implementations fail, or are deployed with flaws that undermine its reliability [van de Sompel 2015]. Another important issue is that OAI-PMH predates REST [Richardson&Ruby 2007], therefore it does not follow the REST principles, further bringing resistance and difficulties in its comprehension and implementation by developers in CHIs.

The motivation for new data acquisition mechanisms also comes from organizational aspects. Europeana has mainly collaborated with aggregators to collect its dataset, making it an ‘aggregator of aggregators’. This approach allowed Europeana to scale up quickly, relying on the aggregators as conduits to attract CH data providers to share their collections on the web. But the aggregation landscape has become more and more complex over time and Europeana’s audiences now demand more: bigger and more beautiful images, playable videos or sound recordings and searchable full-text that can be read on any device. Several drawbacks of the existing aggregation model have been identified in recent reports [Scholz&Devarenne 2016; Devarenne 2017] and surveys [Scholz 2015].

Nowadays, CHIs 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 [Stuart et al. 2015]. Regardless of the metadata aggregation process for Europeana, CHIs 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 CH 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, Museums and Galleries.
- 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.
- CHIs 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 across CHIs 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 requirements that enable the aggregator to provide value added services based on metadata.

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 CHIs to the aggregator. The metadata is transferred to the aggregator, but it continues to evolve at the data provider side, 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 CH 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 for the aspects of data modelling and data synchronization have evolved at different rhythms. While for data modelling the Europeana Data Model (EDM) [EDM Definition] has always been under continuous improvement, the solution for data synchronization, based on OAI-PMH, has not been reassessed since its early adoption.

Another important aspect of EDM is that it does not impose any constraint in the choice of Web technologies for data synchronization. This comes from EDM following the principles of the Web of Data, and that it can be serialized in XML and in RDF formats. This aspect gives the Europeana network much choice for technological innovation of the aggregation network.

In the context of all DSI-2 activities on rethinking the Europeana aggregation network, our work addresses an area that constitutes a sub problem of Task 1.2 - ‘Innovate the aggregation infrastructure for the Europeana DSI’, providing data acquisition mechanisms for METIS [Devarenne 2017]. A similar relation exists with Task 6.10 - ‘Prototype innovative technologies to empower collection owners to publish collections via the Europeana platform services', where data acquisition technologies can provide additional mechanisms for data providers using the Operation Direct[4] Self-Service Harvesting functionalities.

3. Requirements of the Europeana Network

Data aggregation is a general information systems 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.

The Europeana Network is a network of data providing CHIs. When addressing aggregation across organizations, the technological capacity of the participating organizations is a key determinant of the solution to be applied. We define the requirements for the solution by considering the characteristics of the CH domain along with some particularities of the metadata aggregation carried out in the Europeana Network by data providers and aggregators, and the legacy of the current established aggregation practices.

This section presents the requirements separated between the two sub-problems of data aggregation: the synchronization of data sources, and data modelling/representation.

3.1. Synchronization of data sources

The type of solution for synchronization of data sources across organizations is greatly influenced by the requirements for data consistency guarantees, and synchronization latency. For the Europeana Network the solution must allow an aggregator to collect structured metadata about the digital resources that a CHI (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.
- 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.
- A 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 can be more technologically challenging on the aggregators side than on the data providers', since the aggregators are often better prepared to address more complex technical implementation issues of information systems.

[4] http://pro.europeana.eu/share-your-data/innovation-projects/operation-direct

- The solution must be compatible with the ingestion business workflows of Europeana (described in [Devarenne 2017]). Note that some technologies listed in this deliverable may impact the current Europeana data aggregation workflow and more particularly the definition of the dataset entity. More flexibility will be required as the set of resource or Information Package[5] specified for aggregation by a data provider may vary a lot from one technology to another.

In the context of the above requirements, Section 4 presents the Web technologies that we identified as possible solutions for data synchronization.

3.2. Data modeling and representation

In the current Europeana aggregation network, EDM is the technology that supports data sharing efforts in the aspect of data modeling and representation. It is a solution that allows Europeana to become ‘a big aggregation of digital representations of culture artefacts together with rich contextualization data and embedded in a linked Open Data architecture’ [Gradmann 2010].

EDM also has a key role in many other parts of the Europeana Network. EDM has been a collaborative effort from the very start, involving representatives from all the domains represented in Europeana: libraries, museums, archives and galleries. It supports several of the core processes of the Europeana’s operations, and contributes to the access layer of the Europeana Platform[6], where it supports the data reuse by third parties [Charles&Isaac 2015]. EDM’s influence and usage also reaches beyond the Europeana Network, with notable cases such as the Digital Public Library of America that defined its metadata application profile based on EDM [DPLA 2015].

Although reducing the data conversion effort required within the Europeana aggregation infrastructure is a relevant aspect, in our work we considered that any innovative mechanisms for data modelling/representation to be feasible for application in Europeana, should not impact the others areas where EDM is used. We therefore consider that any new technological solution should address the following requirements:

- It must have the capacity to represent the required information for the minimal requirements of EDM and high quality CH data;
- It should be flexible, especially not committed to the vision of only one of the CH domains Europeana is serving, and ideally offering an easy implementation/learning curve (e.g., allowing Dublin Core-level expression of metadata)
- It should show signs of significant adoption and/or interest

Given the above requirements, searching for innovative technologies in this area has proven to be very hard in the course of our work. Only later in second half of the DSI-2 period one viable solution was identified - Schema.org, which is described in the Subsection 4.3.

4. Overview of 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.

[5] https://www.iso.org/standard/57284.html

[6] Europeana Strategy 2015-2020 available at http://strategy2020.europeana.eu/

Our work has not explored all technologies to the same level of detail, however, in this section, we describe all those that we have identified as being applicable. Sections 5 and 6 present the assessment of the technologies, as well as the application of the most suitable ones.

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 CHIs. It specifies several HTTP based web services [Stuart et al. 2015] covering access to images, the presentation and structure of complex digital objects, composed of one or more images, and searching within their content.

IIIF's 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, we have undertaken several experiments and case studies whose detailed results can be found in [Freire et al. 2017a; Freire et al. 2017b]. Our results have indicated that data acquisition via IIIF is feasible and we have identified the requirements that need to be fulfilled by data providers (these will be further discussed in Section 6).

A second concern identified in the case studies, was the efficiency of the harvesting process over IIIF. Efficiency becomes relevant in very large collections, with sizes in the hundreds of thousands of digital objects[7]. 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 have been 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.

Europeana is engaged with the IIIF community, aiming to establish a standard, or guidelines, that will facilitate the metadata aggregation process of IIIF services. Currently, Europeana co-chairs the IIIF Discovery Technical Specification Group[8] and has its use-case under discussion there[9].

4.2. Sitemaps

Sitemaps [SitemapsProtocol] 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

[7] The typical size of the collections delivered to Europeana is within the thousands or tens of thousands. In such cases, we do not consider the loss in efficiency to be significant, due to nowadays high availability of bandwidth and computational capacity.

[8] http://iiif.io/community/groups/discovery/

[9] https://github.com/IIIF/iiif-stories/issues/69

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.

In digital libraries, Sitemaps typically contain all the links to the landing pages of the digital objects within the digital library.

Sitemaps are widely used, and already existing Sitemaps infrastructure could be leveraged on by Europeana for metadata aggregation, using a web crawler 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), a 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, such as IIIF, APIs and linked data. Such combinations would allow for re-use of available data. With the additional use of sitemaps, it may also enable incremental harvesting, through the use of Sitemaps' resource timestamps.

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.

The application of Sitemaps in the Europeana aggregation network is further discussed in Section 6.

4.3. Schema.org

Schema.org[10] was the single new technology we have identified, which could fulfil the requirements of the Europeana aggregation in the area of data modelling and representation.

Schema.org is a cross-domain initiative for structured data in the Internet. Its main application is in web pages, where data can be referenced or embedded in many different encodings, including RDFa, Microdata and JSON-LD. It is developed as a vocabulary, following the Semantic Web principles. It includes entities and relationships between entities.

Web pages containing Schema.org can be processed by search engines and applications using this structured data, in addition to text and links. The Schema.org website reports its usage in more than 10 million sites and Google, Microsoft, Pinterest, Yandex, among others, already provide services and applications that are based on the available Schema.org structured data. They can, for example, know that a web page describes a culinary recipe, its ingredients and preparation method, or that it describes a movie, its actors, user reviews, etc. For CH digital libraries, Schema.org allows the description of books, maps, visual art, music recordings, and many other kinds of cultural resources.

[10] http://schema.org/

Schema.org is a collaborative and community based activity and its main platform of collaboration is the W3C Schema.org Community Group[11]. The Community Group also serves as a hub for discussion with other related communities, at W3C and elsewhere. Other W3C Community Groups exist that are focussed on specific domains, such as health, sports, bibliography, etc. Representatives of the Europeana community may be involved this way, should a need to 'improve' Schema.org for CH aggregation be raised.

The study of the application of Schema.org was one of our main investigations in DSI-2 (our work is described in detail in [Freire et al. 2017c]). Its application in the Europeana aggregation network is further discussed in Section 6.

4.4. ResourceSync

ResourceSync [NISO 2008] 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 Sitemaps’ 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
- specify alternative download mechanisms, such as alternative mirrors.
- send notifications about resource updates

ResourceSync specifies how to ‘enhance’ a ResourceSync enabled data source with a notifications mechanism based on WebSub [Genestoux&Parecki 2017]. WebSub specifies the communication between publishers of any kind of Web content and their subscribers, based on HTTP. Subscription requests are relayed through hubs, which validate and verify the request. Hubs then distribute new and updated content to subscribers when it becomes available.

This detailed synchronization information provided by ResourceSync allows for more efficient ways of keeping resources synchronized between a source and a destination than Sitemaps and any other technology that we have analysed.

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 CH domain. Most applications of ResourceSync are in grey literature repositories.

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 due to two factors: first, the aggregation

[11] http://www.w3.org/community/schemaorg

of content as well as metadata is starting to gain more attention within the Europeana Network; and second, we believe that due to increasing signs of usage in grey literature, in the near future several CHIs, particularly research libraries, may favour the use of ResourceSync.

4.5. Linked Data Platform

Linked Data Platform [Speicher et al. 2015] specifies the use of HTTP and RDF techniques for accessing and manipulating resources exposed as Linked Data [Berners-Lee 2006].

Several CHIs publish, as linked data, the metadata regarding their resources. Although linked data publication allows for a standard way to reach the metadata in an automatized way, the standard practices do not address all the requirements needed for aggregation by Europeana. Mainly two aspects need to be addressed for linked data sources: first, a mechanism for allowing CHI to indicate to Europeana which metadata resources are to be aggregated; and second, an efficient mechanism for allowing efficient incremental harvesting.

Within the many aspects specified by the Linked Data Platform, some provide the necessary standardization for an efficient aggregation based on linked data sources. In particular the Linked Data Platform Containers[12] and the specification of the usage of HTTP 1.1[13] could fulfil the requirements for aggregation by Europeana.

4.6. 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 [Parecki 2016]. 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 CHIs;
- 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;

[12] Linked Data Platform Containers section of [Speicher et al. 2015], available at: https://www.w3.org/TR/ldp/#ldpc

[13] HTTP 1.1 section of [Speicher et al. 2015], available at: https://www.w3.org/TR/ldp/#specs-http

- 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.

We have conducted exploratory discussions, across teams within Europeana Foundation, regarding the possible application of Webmention[14]. 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.

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 an IIIF crawler to harvest the metadata from the IIIF endpoint.

4.7. Linked Data Notifications

Linked Data Notifications [Capadisli&Guy 2017] (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 have provided feedback to the LDN specifications, considering the metadata aggregation use case. We also developed a pilot of LDN harvester in the IIIF context[15], which was listed among LDN implementation when that technology was published as a formal W3C Recommendation [Capadisli&Guy 2017].

4.8. 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 [Openpub 2011] 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. Only one case was identified[16] .

[14] https://docs.google.com/presentation/d/1PWX3A5zfI8FcEPdBWhabHHPw3FjDFmbR7J2CsLsspVw/

[15] https://github.com/nfreire/LDN4IIIF

[16] http://gallica.bnf.fr/blog/27042017/retrouvez-tous-nos-livres-au-format-epub-dans-votre-application-de-lecture-favorite

5. Results

Given the high number of technological options identified, during the course of our work not all technologies were studied to the same level of detail. We present the main results of our work by presenting a general view of the key positive and negative aspects of all the technologies, and the stage of investigation that was reached in each technology. We also assess the suitability of the technologies for Europeana's aggregation, and indicate the priorities for technology adoption.

5.1. Analysis of the technologies

Due to the large number of technological options available, our method of evaluation was to perform case studies, and prototyping of possible solutions. In some cases, these activities were carried out in cooperation with data providers from the Europeana Network [Freire et al. 2017b] or with other CHIs when no case studies were found in the Europeana Network [Freire et al. 2017c].

With other DSI-2 tasks engaged in innovating the Europeana aggregation network, our work produced significant results in two cases. First, the IIIF and Sitemaps prototype was integrated in the Operation Direct Self-Service Harvesting service. Second, we performed some case studies with the Data Partners Services team for better understanding the workflows of aggregating IIIF sources. Some of these IIIF case studies resulted in the integration of IIIF collections in the production Europeana dataset.

Five software prototypes were developed during the course of our work. Some for support of the case studies and others for evaluating the functionality of the technologies and influence their ongoing specifications to include the aggregation case of Europeana. The prototypes are listed in Table 1, and their source code is openly available in a GIT repository[17].

We did not reach an experimental stage for all of the technologies. In some cases because the initial stages of their analysis considered them clearly less suitable than the other alternatives, or because they were identified too late in the project. Their evaluation was done through analysis of their documentation, functional specifications, and/or by engaging in discussions with both the CH community and digital library research community.

We used three main criteria for ranking the technologies:

- Usage in CHIs: if the technology was already in use and know-how exists in CHIs. For example, we privileged technologies that were already being used for search engines optimization by data providers.
- Complexity for data providers: We evaluated the complexity for data providers in the adoption of the technology. Several factors were analysed: technical complexity of the technology; availability of open and reusable implementations; existence of supporting user communities; availability in commercial IT products; and need for in-house software development.
- Complexity for Europeana: The same factors mentioned in the previous point for data providers were taken in consideration for the adoption of technologies by Europeana. In this case, however, the requirements were considered from the point of view of the aggregator role. One additional factor was considered for Europeana - the need for definition and maintenance of Europeana specifications for the usage of a technology for an aggregation application. For example, some technologies would require Europeana to

[17] The data acquisition experimental software developed in the course of DSI-2 can be consulted and obtained at https://github.com/nfreire/Open-Data-Acquisition-Framework

undertake extensive work in complementing the technology specification with specification of how the technology should be used for Europeana aggregation.

Technology	Experimented with data providers	Experimented with ingestion at Europeana	Prototype software status
Sitemaps	Yes	Yes	Ready for technology transfer. A few minor features not implemented.
IIIF	Yes	Yes	Ready for technology transfer, but (official) IIIF recommendations for discovery are not available yet. Has potential applications to support ingestion in Europeana in some cases.
LDN	No	No	For experimental purposes only.
Webmention	No	No	For experimental purposes only.
Schema.org	Yes	No	For experimental purposes only.

Table 1: Software prototypes developed during DSI-2

A summary of our assessment is presented in Table 2, and detailed results from the experimental work can be found in [Freire et al. 2017a; Freire et al. 2017b; Freire et al. 2017c]. The next section of this report presents a deeper discussion and analysis of adoption of these technologies.

Technology	Usage in CHIs	Complexity for data providers	Complexity for Europeana
Sitemaps	High	Very Low	Very low
IIIF	High	Medium, if already in use High if not in use	Medium
LDN	None known	Medium	High
Webmention	None known	Low	Medium
Schema.org	Low (but with signs of growing)	Medium	Medium
ResourceSync	Low	Medium	Medium
WebSub	None known	Not investigated	Not investigated
Linked Data Platform	High	No conclusive results yet	No conclusive results yet
OPDS	Unknown	Low	Low

Table 2: Summary of the feasibility analysis of each technology

5.2. Recommended technologies and further research

The final results of this task consist in a final assessment on the suitability of the technologies for choosing where work on a adopting a technology may be started, those that showed promising applications but further research is still needed, and those that were identified as inferior options.

From our assessment of all the identified technologies, we consider three technologies to be the most suitable: IIIF, Sitemaps and Schema.org.

- IIIF - We have chosen IIIF because it is gaining traction in CH. Moreover, it is a community developed, open framework. Our requirements and suggestions for metadata aggregation may thus be incorporated into future versions. Our case studies indicated 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.

- Sitemaps - The choice for Sitemaps was motivated by its wide usage within the Europeana data providers. In addition, Sitemaps also provides a very simple technological solution, with a very low implementations barrier.
- Schema.org - Schema.org is complementary to IIIF and Sitemaps, since its main benefit is in simplifying or bringing more value to data providers from their data conversion efforts.

Technology	Short term application feasibility for the Europeana Network	R&D status
Sitemaps	Yes	Ready for knowledge transfer
IIIF	Yes	Ready for knowledge transfer
LDN	No	More R&D required
Webmention	No	Closed
Schema.org	Yes, if usage by data providers keeps growing	Results supported its feasibility, but additional R&D required for putting it into practice
ResourceSync	No	More R&D required
WebSub	No	More R&D required
Linked Data Platform	No	More R&D required
OPDS	No	Closed

Table 3: Summary of the technological assessment

Our assessment also considered two additional technologies to be relevant: ResourceSync and Linked Data Notifications. ResourceSync because it presents itself as the best technical solution for aggregation of very large datasets and for aggregation of content. Notification mechanisms, such as Linked Data Notifications, would enable an agile exchange of data/information between Europeana and data providers that could increase the value given back to data providers for their participation in Europeana. These technologies are less widely applicable since they have higher implementation costs for data providers, but they may enable Europeana to conceive new discovery services for CH.

All these technologies are compatible with the ingestion workflows business requirements of Europeana [Devarenne 2017], however some technologies may offer extra functionality that can benefit the ingestion at Europeana. This aspect may be considered for future research at Europeana.

6. Application scenarios and adoption of the recommended technologies

The three recommended technologies are very distinct in terms of application scenarios and adoption requirements. Even for the same technology, different application scenarios may present their own specific adoption requirements, and these have different impact for Europeana and for providers. In this section we present our view for the three recommended technologies.

6.1. International Image Interoperability Framework

The core value of IIIF for data providers resides in the presentation possibilities it provides for end-users in interacting with digital objects. Its applicability for Europeana aggregation is therefore only for cases where data providers are already engaged in IIIF implementation for end-users. Since IIIF is showing clear signs of growing usage among Europeana providers, its reuse for Europeana aggregation is a strong motivation for providers. The high usage of IIIF also means that extensive IIIF knowledge is available within data providers, which will greatly reduce their effort for implementing the Europeana aggregation requirements by IIIF.

Europeana is also very engaged in IIIF activities beyond aggregation. The Europeana Foundation and partners from the Europeana Network have IIIF knowledge and a IIIF Task Force (initiated by DSI Task 6.9.2) has worked to even better position our community with IIIF.. Also a recommendation is available for how to represent IIIF enabled resources in EDM, making them available for the Europeana Collections portal [Isaac&Charles 2016]. One final factor is that IIIF's community goals are well aligned with the business objectives of Europeana for improvement of end-user interaction of content from digital objects in its portal.

6.1.1. Application scenarios

IIIF aggregation is feasible when data providers are already implementing IIIF for end-users.

6.1.2. Adoption by data providers

For using their IIIF services to provide data to Europeana, data providers typically need to implement some additional functionality such as the provision of a Sitemap of IIIF resources, or a IIIF Collection. These inform Europeana of the IIIF resources that are intended for aggregation. In addition, links to EDM metadata must also be available in the IIIF Manifests.

6.1.3. Adoption by Europeana Foundation

For aggregating metadata through IIIF, Europeana would require IIIF harvesting software to be setup in its aggregation systems. The harvesting software would need to support harvesting of IIIF sources via Sitemaps and IIIF Collections.

Europeana would also need to provide guidelines for data providers on how to prepare a IIIF service for Europeana aggregation.

6.2. Sitemaps

Sitemaps is a simple technology with very low implementation barriers, and is widely used by CHIs for Internet search engines. It enables or simplifies aggregation by other technologies. Europeana also uses Sitemaps (in its portal), and therefore, possesses knowledge about the technology.

6.2.1. Application scenarios

For providers that have no other technology in place for providing data to Europeana, Sitemaps can be applied more easily than OAI-PMH, to make EDM metadata available to Europeana.

Another application of Sitemaps would be to complement other technologies that need a mechanism for data synchronization. We consider the following cases to be relevant for Europeana:

- IIIF services - sitemaps make references to IIIF Manifests.
- Schema.org (web pages) - sitemaps include URLs of web pages which include Schema.ord embedded data or references.
- Schema.org (APIs) - sitemaps include URLs of provider specific APIs, which output Schema.ord data.
- Schema.org (linked data) - sitemaps include URIs of linked data resources.

6.2.2. Adoption by data providers

The sophistication of a Sitemap may vary considerably. In its simplest form it requires just the creation of a file listing URLs. The main disadvantage of simple Sitemaps is computational inefficiency of harvesting, which can affect the aggregation of large collections. We analyzed the size of Europeana collections and the majority of collections are of sizes that could be harvested using these simple Sitemaps, therefore, simple Sitemaps be suitable for many small and medium data providers.

For more sophisticated Sitemaps, data providers must provide resource modification and deletion timestamps. With these Sitemaps, the harvest processes can be as efficient as OAI-PMH.

6.2.3. Adoption by Europeana Foundation

For aggregating metadata through Sitemaps, Europeana would require a Sitemaps harvesting software to be setup in its aggregation systems. The harvesting software would need to support harvesting also of the other technologies used in conjunction with Sitemaps, as described in the application scenarios..

Europeana would also need to provide guidelines for data providers. The guidelines would need to indicate what kind of applications scenarios are supported and how providers should prepare their Sitemaps for Europeana aggregation.

6.3. Schema.org

Of the three selected technologies, Schema.org is the one less likely to provide a considerable number of implementations in the short term, since we have not identified any production level implementations within the Europeana Network. Also, although our assessment of Schema.org was successfully concluded [Freire et. al 2017c], we consider that the use of Schema.org in Europeana aggregation is not at production level, since it still requires some research effort to fully prepare its application.

Schema.org looks very promising in the medium term, however. It has the potential to reduce the effort on data conversion by data providers, since it becomes a shared solution for discovery through Europeana and through Internet search engines. The Sitemaps+Schema.org approach used by Internet search engines could be leveraged on for the purposes of Europeana aggregation. In some cases, it could completely eliminate the extra effort of delivering data to Europeana.

6.3.1. Application scenarios

Schema.org aggregation is feasible when data providers are already implementing it for search engine discovery purposes, or wishes to implement Schema.org from the start with both cases in mind.

6.3.2. Adoption by data providers

Assuming that Schema.org metadata is already being prepared for search engines, only minor adaptations to the Schema.org metadata may be necessary, to comply with Europeana aggregation requirements.

6.3.3. Adoption by Europeana Foundation

Adoption of Schema.org by Europeana would be much more demanding than for data providers. Europeana would require metadata mappings and conversion tools. Given the size of the Schema.org vocabulary, this effort could be significant. Schema.org is also evolving constantly, requiring Europeana to keep monitoring it, and maintaining its mappings and tools.

Europeana would also need to provide guidelines for data providers. The guidelines would need to indicate how the EDM requirements and recommendations for delivery of high quality should be applied under a Schema.org representation.

7. Conclusion

Our work has shown that several technological solutions are available to make the Europeana aggregation network more efficient and with lower barriers for data providers.

While some solutions still require further research, some can be adopted in the short term. The clearest first choices to address are the application of IIIF and Sitemaps for data synchronization between providers and Europeana. These two technologies are already in use in CHIs and can be leveraged on for Europeana’s aggregation purposes.

Next comes the application of Schema.org for data modeling and representation, whose application has been shown to be feasible in our case studies and with potential benefit in simplifying or bringing more value to data providers from their data conversion efforts.

The adoption of an initial solution by Europeana is mainly dependent on aligning the technical solution with the business objectives of Europeana for the coming years, then in establishing best practices within the Network, and in equipping Europeana with the necessary software tools and internal workflows.

References

[Berners-Lee 2006]
T. Berners-Lee. 2006. Linked Data Design Issues. W3C-Internal Document. <http://www.w3.org/DesignIssues/LinkedData.html>

[Capadisli&Guy 2017]
S. Capadisli, A. Guy (eds.). 2017. Linked Data Notifications. W3C Recommendation.. <https://www.w3.org/TR/ldn/>

[Charles&Isaac 2015]
V. Charles, A. Isaac. 2015. Enhancing the Europeana Data Model (EDM). Project Europeana V3.0
<http://pro.europeana.eu/files/Europeana_Professional/Publications/EDM_WhitePaper_17062015.pdf>

[Devarenne 2017]
C. Devarenne. 2017. MS1.1: Ingestion workflows business requirements update. Project Europeana DSI 2– Access to Digital Resources of European Heritage <http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Milestones/ms1.1-ingestion-workflows-business-requirements-update.pdf>

[DPLA 2015]
Digital Public Library of America. 2015. Metadata Application Profile, version 4.0. <https://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf>

[EDM Definition]
Europeana v1.0. The EDM Definition V5.2.7. <http://pro.europeana.eu/web/guest/edm-documentation>

[Freire et al. 2017a]
N. Freire, H. Manguinhas, A. Isaac, G. Robson, J. B. Howard. 2017. Web technologies: a survey of their applicability to metadata aggregation in cultural heritage. 21st International Conference on Electronic Publishing. <http://ebooks.iospress.nl/publication/46657>

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

[Freire et al. 2017c]
N. Freire, V. Charles, A. Isaac. 2017. Report on the case studies of Schema.Org metadata acquisition for Europeana. Technical report. Project Europeana DSI 2– Access to Digital Resources of European Heritage. <https://docs.google.com/document/d/1ncDjScep73irC_AjAMhTpfb4134L71kh0ynjrxcGwsI>

[Genestoux&Parecki. 2017]
J. Genestoux, A. Parecki (eds.). 2017. WebSub. W3C Candidate Recommendation. <https://www.w3.org/TR/websub/>

[Gradmann 2010]
S. Gradmann. 2010. Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. <http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20White%20Paper%201.pdf>

[Isaac&Charles 2016]
A. Isaac, V. Charles (eds.). 2016. Guidelines for submitting IIIF resources for objects in EDM. <http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/IIIFtoEDM_profile_042016.pdf>

[Lagoze et al. 2002]
C. Lagoze, H van de Sompel, M.L. Nelson, S. Warner. 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

[NISO 2008]
National Information Standards Organization. 2014. ResourceSync Framework Specification. <http://www.niso.org/apps/group_public/download.php/12904/z39-99-2014_resourcesync.pdf>

[Openpub 2011]
The openpub community. 2011. OPDS Catalog 1.1 specification.. <http://opds-spec.org/specs/opds-catalog-1-1>

[Parecki 2016]
A. Pareck (ed.). 2016. Webmention. W3C Candidate Recommendation. <https://www.w3.org/TR/webmention/>

[Pedrosa et al. 2010]
G. Pedrosa, G. Petz, C. Concordia, N. Aloia. 2010. Europeana OAI-PMH Infrastructure. Project Europeana Connect deliverable D5.3.1.

[Richardson&Ruby 2007]
L. Richardson, S. Ruby. 2007. Restful Web Services. O'Reilly.

[Scholz 2015]
H. Scholz. 2015. D1.1: Recommendations to improve aggregation infrastructure. Project Europeana version 3. <http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Deliverables/EV3%20D1_1%20Aggregation%20Infrastructure.pdf>

[Scholz&Devarenne 2016]
H. Scholz, C. Devarenne. 2016. D1.1: Work and implementation plan to innovate the aggregation infrastructure. Project Europeana Data Service Platform. <http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.1-work-and-implementation-plan-to-innovate-the-aggregation-infrastructure.pdf>

[SitemapsProtocol]
Sitemaps XML format. <https://www.sitemaps.org/protocol.html>

[Speicher et al. 2015]
Speicher S, Arwe J, Malhotra A. 2015. Linked Data Platform 1.0. W3C Recommendation. <https://www.w3.org/TR/ldp/>

[Stuart et al. 2015]
S. Stuart, R. Sanderson, T Cramer. 2015. The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. Archiving 2015. <http://purl.stanford.edu/df650pk4327>

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

[van Veen&Oldroyd 2004]
T. van Veen, B. Oldroyd. 2004. Search and Retrieval in The European Library: A New Approach. D-Lib Magazine, vol. 10, n. 2. ISSN 1082-9873.

[Verwayen 2017]
H. Verwayen. 2017. Business Plan 2017: ‘Spreading the Word’. <http://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-business-plan-2017.pdf>

[Wallis et al. 2017]
R. Wallis, 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. Code4Lib Journal, Issue 36. ISSN 1940-5758. <http://journal.code4lib.org/articles/12330>