h e g

Haute école de gestion
Genève

Enabling better aggregation and discovery of cultural heritage content for Europeana and its partner institutions

Master's thesis in Information Science carried out by:

Julien Antoine RAEMY

Master's thesis supervised by:

Arnaud GAUDINAT, Associate Professor

Geneva, Switzerland, 14 August 2020

Information Science Department

Master of Science HES-SO in Information Science

Haute école de gestion de Genève

Hes·so GENÈVE
University of Applied Sciences and Arts
Western Switzerland

Acknowledgements

I would like to thank all of the people who have helped me in any way possible over the course of this master's thesis, which also marks the end of my higher education at the Haute école de gestion de Genève, where I completed a bachelor's degree.

First of all, I would like to thank Professor **Arnaud Gaudinat** for his sound advice, his availability and for agreeing to oversee this thesis.

I am very grateful to **Emmanuelle Bermès**, Deputy Director for Services and Networks at the Bibliothèque nationale de France, for having agreed to be the external expert assessing this master's thesis.

I am very much indebted to **Antoine Isaac**, R&D Manager at Europeana, for involving me in the team's discussion, for facilitating exchanges and for his countless suggestions, feedback as well as his patience. Many thanks as well to all Europeana R&D team members, with whom I interacted on a weekly basis and who offered me a great deal of assistance: **Nuno Freire**, **Mónica Marrero**, **Albin Larsson**, and **José Eduardo Cejudo Grano de Oro**.

In this respect, I would also like to mention and thank the following people working within the Europeana Network for their ongoing support: **Valentine Charles**, **Gregory Markus**, **Andy Neale**, **Hugo Manguinhas**, **Henning Scholz**, **Sebastiaan ter Burg**, **Erwin Verbruggen**, **Enno Meijers**, **Haris Georgiadis** and **Cosmina Berta**.

I would also like to extend my deepest appreciation to **Anne McLaughlin** who kindly agreed to give some of her time to proofread this dissertation.

Special thanks also go to everyone who took part in the online survey, and to all of those with whom I was able to correspond through email for possible pilots or simply to get additional information.

Lastly, I would like to express my deep appreciation to my girlfriend, friends, colleagues and family who supported me not only during the last semester of this master's degree in Information Science, but throughout the course of my studies.

Abstract

Europeana, a non-profit foundation launched in 2008, aims to improve access to Europe's digital cultural heritage through its open data platform that aggregates metadata and links to digital surrogates held by over 3700 providers. The data comes both directly from cultural heritage institutions (libraries, archives, museums) as well as through intermediary aggregators. Europeana's current operating model leverages the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Europeana Data Model (EDM) for data import through Metis, Europeana's ingestion and aggregation service.

However, OAI-PMH is an outdated technology, and is not web-centric, which presents high maintenance implications, in particular for smaller institutions. Consequently, Europeana seeks to find alternative aggregation mechanisms that could complement or supersede it over the long-term, and which could also bring further potential benefits.

In scope, this master's thesis seeks to extend the research on earlier aggregation experiments that Europeana successfully carried out with various technologies, such as aggregation based on Linked Open Data (LOD) datasets or through the International Image Interoperability Framework (IIIF) APIs.

The literature review first focuses on metadata standards and the aggregation landscape in the cultural heritage domain, and then provides an extensive overview of Web-based technologies with respect to two essential components that enable aggregation: data transfer and synchronisation as well as data modelling and representation.

Three key results were obtained. First, the participation in the Europeana Common Culture project resulted in the documentation revision of the LOD-aggregator, a generic toolset for harvesting and transforming LOD. Second, 52 respondents completed an online survey to gauge the awareness, interest, and use of technologies other than OAI-PMH for (meta)data aggregation. Third, an assessment of potential aggregation pilots was carried out considering the 23 organisations who expressed interest in follow-up experiments on the basis of the available data and existing implementations. In the allotted time, one pilot was attempted using Sitemaps and Schema.org.

In order to encourage the adoption of new aggregation mechanisms, a list of proposed suggestions was then established. All of these recommendations were aligned with the Europeana Strategy 2020-2025 and directed towards one or several of the key roles of the aggregation workflow (data provider, aggregator, Europeana).

Even if a shift in Europena's operating model would require extensive human and technical resources, such an effort is clearly worthwhile as solutions presented in this dissertation are well-suited for data enrichment and for allowing data to be easily updated. The transition from OAI-PMH will also be facilitated by the integration of such mechanisms within the Metis Sandbox, Europeana's new ad-hoc system where contributors will be able to test their data sources before ingestion into Metis. Ultimately, this shift is also expected to lead to a better discoverability of digital cultural heritage objects.

Keywords: API, Cultural heritage, Data aggregation, Digital transformation, Discovery, Europeana Common Culture, EDM, IIIF, LOD, OAI-PMH, RDF, ResourceSync, Schema.org, SEO, Sitemaps, Social Web Protocols

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AACR2	Anglo-American Cataloguing Rules, 2nd edition
ABCD	Access to Biological Collections Data
AP	ActivityPub
API	Application programming interface
ArCo	Architecture of Knowledge
AS2	ActivityStreams 2.0
BIBFRAME	Bibliographic Framework
CARARE	Connecting Archaeology and Architecture in Europe
CC	Creative Commons
CCO	Cataloguing Cultural Objects
CC0	Creative Commons Zero Public Domain Dedication
CH	Cultural heritage
CHI	Cultural heritage institution
CHO	Cultural heritage object
CIDOC-CRM	CIDOC Conceptual Reference Model
CLI	Command-line interface
CMS	Content management system
CSV	Comma-separated values
DACS	Describing Archives: A Content Standard
DAL	Data Aggregation Lab
DC	Dublin Core
DCAT	Data Catalog Vocabulary
DCAT-AP	DCAT Application Profile for data portals in Europe
DCT	Dublin Core Terms
DCMES	Dublin Core Metadata Element Set
DEA	Data Exchange Agreement
DPLA	Digital Public Library of America
DSI	Digital Service Infrastructure
EAC-CPF	Encoded Archival Context - Corporate Bodies, Persons, and Families

EAD	Encoded Archival Description
EAG	Encoded Archival Guide
ECC	Europeana Common Culture
EDM	Europeana Data Model
EAF	Europeana Aggregators Forum
EF	Europeana Foundation
ENA	Europeana Network Association
EPF	Europeana Publishing Framework
ESE	Europeana Semantics Elements
FRBR	Functional Requirements for Bibliographic Records
GLAM	Galleries, Libraries, Archives, Museums
HDT	Header Dictionary Triples
HTML	Hypertext Markup Language
HTTP	HyperText Transfer Protocol
IIIF	International Image Interoperability Framework
IIIF-C	IIIF Consortium
ISAD(G)	International Standard Archival Description (General)
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KB	Koninklijke Bibliotheek (The Royal Library of the Netherlands)
LD	Linked Data
LDF	Linked Data Fragments
LDN	Linked Data Notification
LDP	Linked Data Platform
LIDO	Lightweight Information Describing Objects
LOD	Linked Open Data
MADS	Metadata Authority Description Schema
MARC	Machine-Readable Cataloging
METS	Metadata Encoding and Transmission Standard
MODS	Metadata Object Description Schema

MTSR	Metadata and Semantics Research
NDLI	National Digital library of India
NDE	Netwerk Digitaal Erfgoed (Dutch Digital Heritage Network)
NISO	National Information Standards Organization
NISV	Netherlands Institute for Sound and Vision
NT	N-Triples
N3	Notation 3
N/A	Not applicable
OAI-ORE	Open Archives Initiative Object Reuse and Exchange
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ONIX	Online Information Exchange
OPDS2	Open Publication Distribution System 2.0
OWL	Web Ontology Language
PID	Persistent identifier
PSNC	Poznan Supercomputing and Networking Center
PuSH	PubSubHubbub
RDA	Resource Description & Access
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	Resource Description Framework Schema
REST	Representational state transfer
RiC	Records in Contexts
RS	ResourceSync
R&D	Research and Development
SEO	Search Engine Optimization
SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organization System
SOCH	Swedish Open Cultural Heritage
SPARQL	SPARQL Protocol and RDF Query Language
SRU/SRW	Search and Retrieve URL/Web Service

TBD	To be determined
Turtle	Terse RDF Triple Language
UCD	University College Dublin
URL	Uniform Resource Locator
URI	Uniform Resource Identifier
VoID	Vocabulary of Interlinked Datasets
VRA	Visual Resources Association
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations
W3C	World Wide Web Consortium

Terminology

This terminology lays out the definitions of the most relevant concepts discussed in this master's thesis. The concepts identified in Table 1 provide a general overview and are not an exhaustive list of the topics covered in this dissertation. Above all, it allows readers to have a good sense of the study's rationale.

Table 1: Glossary of Terms

Term	Definition	Source
Application programming interface (API)	An API is an abstraction implemented in software that defines how others should make use of a software package such as a library or other reusable program. APIs are used to provide developers access to data and functionality from a given system.	(Hyland et al. 2013)
Conceptual Model	Conceptual Models provide a high-level approach to resource description in a certain domain. They typically define the entities of description and their relationship to one another. Metadata structure standards typically use terminology found in conceptual models in their domain.	(Riley, Becker 2010)
Content Standard	Content Standards provide specific guidance on the creation of data for certain fields or metadata elements, sometimes defining what the source of a given data element should be. They may or may not be designed for use with a specific metadata structure standard.	(Riley, Becker 2010)
Controlled Vocabulary	Controlled Vocabularies are enumerated (either fully or by stated patterns) lists of allowable values for elements for a specific use or domain.	(Riley, Becker 2010)
Data Modelling	Data modelling is a process of organising data and information describing it into a faithful representation of a specific domain of knowledge.	(Hyland et al. 2013)
Digital transformation	Digital transformation is an umbrella term that captures the impact of digital innovation on the ground in different sectors. [For Europeana, it's not about simply applying technology, but by doing it] sensibly and with serious consideration to implementing [Europeana's] values.	(D'Alterio 2018)[1]
Discovery	Discovery is the ability for automated processes to find harvestable content for the purposes of aggregating it, thus allowing that content to be subsequently retrieved on a search engine which is used by either humans with a user interface or machines via an API. *NB: This is a very specific view on discovery as it is here conceptually understood as a process.*	Robert Sanderson[2]
Ingestion	The process of collecting, mapping and publishing the data from a data provider.	(Europeana 2015)

[1] Interview with Harry Verwayen, Europeana Foundation Executive Director

[2] Direct message from Robert Sanderson, Cultural Heritage Metadata Director at Yale University, IIIF Slack instance, 4 June 2020

Term	Definition	Source
Linked Data (LD)	A pattern for hyperlinking machine-readable data sets to each other using Semantic Web techniques, especially via the use of RDF and URIs. Enables distributed SPARQL queries of the data sets and a browsing or discovery approach to finding information (as compared to a search strategy). Linked Data is intended for access by both humans and machines. Linked Data uses the RDF family of standards for data interchange (e.g., RDF/XML, RDFa, Turtle) and query (SPARQL).	(Hyland et al. 2013)
Linked Open Data (LOD)	Linked Data published on the public Web and licensed under one of several open licenses permitting reuse.	(Hyland et al. 2013)
Markup Language	A formal way of annotating a document or collection of digital data using embedded encoding tags to indicate the structure of the document or data file and the contents of its data elements. It also provides a computer with information about how to process and display marked-up documents. [Markup Language] are unlike other "metadata" formats in that they provide not a surrogate for or other representation of a resource, but rather an enhanced version of the full resource itself.	(Baca 2016a; Riley, Becker 2010)
Metadata	Information used to administer, describe, preserve, present, use or link other information held in resources, especially knowledge resources, be they physical or virtual. Metadata may be further subcategorized into several types (including general, access and structural metadata). Linked Data incorporates human and machine-readable metadata along with it, making it self-describing.	(Hyland et al. 2013)
Metadata Aggregation	Metadata aggregation is an approach where centralized efforts like Europeana facilitate the discoverability [of resources] by collecting [ingesting] their metadata.	(Freire, Meijers, et al. 2018)
Metadata Mapping	An expression of rules to convert structured data from one format or model to another such as the Europeana Data Model (EDM).	(Europeana 2015)
Record Format	Record Formats are specific encodings for a set of data elements. Many structure standards are defined together with a record format that implements them.	(Riley, Becker 2010)
Structure Standard	Structure Standards are those that define at a conceptual level the data elements applicable for a certain purpose or for a certain type of material. These may be defined anew or borrowed from other standards. This category includes formal data dictionaries. Structure standards do not necessarily define specific record formats.	(Riley, Becker 2010)

Term	Definition	Source
Uniform Resource Identifier (URI)	A global identifier standardized by joint action of the World Wide Web Consortium (W3C) and Internet Engineering Task Force. A Uniform Resource Identifier (URI) may or may not be resolvable on the Web. URIs play a key role in enabling Linked Data. URIs can be used to uniquely identify virtually anything including a physical building or more abstract concepts such as colours.	(Hyland et al. 2013)
User/end-user	A person or entity making use of the services offered by Europeana through the Europeana Portal, Europeana API, third party services or social networks.	(Europeana 2015)

1. Introduction

The presentation of digital objects by cultural heritage institutions on their respective web platforms is a great opportunity to showcase resources, facilitate access for researchers as well as engage with new audiences. This is all the truer for unique digitised artefacts that are rarely accessible to the general public and which, in most cases, can usually only be consulted by a select few users.

Such democratisation and easy access to digital resources nonetheless faces significant challenges because end users, using Internet commodity search engines, hardly ever discover what has been indexed in digital library catalogues, which are generally built in silos – i.e. restricted access within bespoke applications – and often poorly referenced. Similarly, cultural heritage institutions do not have an equal chance to cope with the pace of digital transformation, and small and medium sized institutions do not necessarily have the necessary tools and resources.

To avoid users looking for a needle in a haystack, federated efforts are critical. In this respect, Europeana, a web portal created by the European Union and officially launched in 2008, has strived to position itself as the main gateway for accessing Europe's cultural heritage.

Europeana is in line with other large-scale digital library initiatives such as the Digital Public Library of America, the National Digital Library of India, or Trove in Australia, which not only want to aggregate and disseminate content on their platform, but, thanks to their expertise in research and development, are able to explore new harvesting approaches.

Nevertheless, digital transformation is far from being an easy task. It is especially the case in the cultural heritage field where libraries, archives and museums are accustomed to working with their own metadata standards and once a technology is implemented, it tends to be used for a long time to justify the investment as any deployed technical solution is usually kept in their infrastructure for an extended period of time. For instance, Europeana has to deal with some technologies to ingest the collections displayed on their open data platform that have been around for twenty years. Indeed, the technology of choice in the context of metadata aggregation is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which is itself an outdated mechanism that pre-dates certain principles of the architecture of the World Wide Web.

This master's thesis sought to query some Europeana stakeholders on alternative aggregation mechanisms facilitating a transition to other technologies to bring other benefits such as more efficient web referencing, enhanced synchronisation, or even greater interoperability. Improved aggregation holds the promise of greater discoverability for both human and machine users. An upgraded workflow is also key for cultivating new pathways for organisations and users alike to further engage with digital cultural heritage resources.

Beginning with a presentation of the **Context**, this dissertation then follows a relatively standard structure of any scholarly paper: the essential state-of-the-art components are highlighted within the **Literature review**, the overall **Methodology** is listed and briefly described, the outcomes are showcased and analysed in the **Results**, some **Recommendations** are drawn and a **Conclusion** reflecting on the achievements and outcomes as well as establishing future work completes the master’s thesis.

2. Context

This master's thesis is part of the final examination requirements of the Haute école de gestion de Genève (HEG-GE), for obtaining the Master of Science HES-SO in Information Science.

This chapter outlines the background of the master's thesis which was conducted by the author in collaboration with the Europeana Research and Development (R&D) team from 20 February to 14 August 2020[3]. In this regard, it provides a few key insights on Europeana, the rationale and background, as well as the research scope.

2.1 Europeana

Europeana is a non-profit foundation based in The Hague that supports the Europeana service, launched in 2008 an initiative of the European Commission[4]. The Europeana Foundation (EF) serves as facilitator for a community of 2400 experts in digital cultural heritage (the Europeana Network Association - ENA). Their mission is to improve access to Europe's digital cultural heritage through their open data platform which aggregates metadata and links to digital surrogates held by over 3700 providers (Isaac 2019) from cultural heritage institutions (CHIs) such as libraries, archives, museums.

The data comes both directly from organisations, also called data providers, as well as through aggregators, which are intermediaries in the aggregation process who collect data from specific countries or regions, and from specific domains (audio heritage, fashion, photography, etc.). These aggregators advise their providers, for instance, on formats, licenses or any technical conditions under which data can be aggregated.

Since the beginning of Europeana, the data import has been based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The resources, which need to comply with the Europeana Data Model (EDM), are currently ingested by Metis, Europeana's ingestion and aggregation service.

2.2 Rationale and background

This section summarises the motivations for revising the aggregation workflow at Europeana, previous and current R&D projects, as well as the current aggregation strategy.

2.2.1 Motivations for revising the aggregation workflow at Europeana

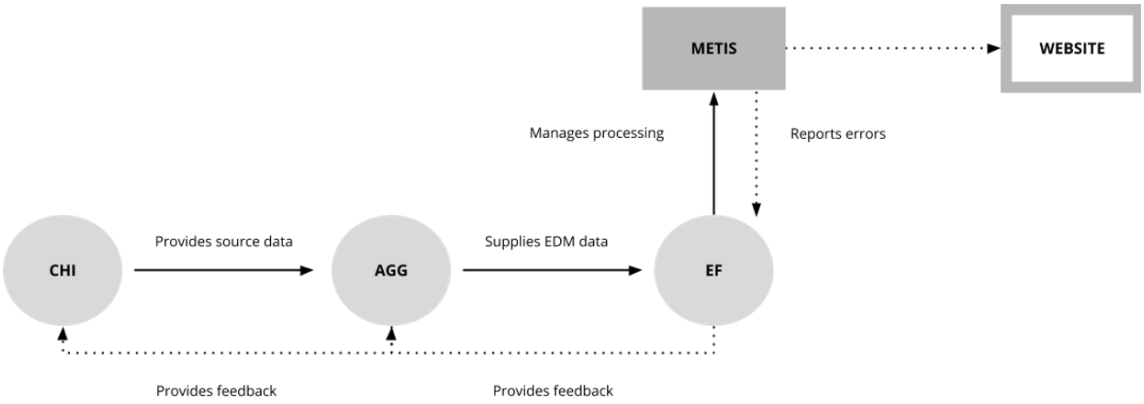
Europeana would like to use technologies other than OAI-PMH, which began development in 1999 and has been stabilised in its second version since 2002 (Lagoze et al. 2002), to aggregate metadata. There are many arguments to support discontinuing the use this protocol. For instance, OAI-PMH is an outdated technology that is not very efficient as data must be copied in several places, its scalability is not optimal, it is not web-centric (Van de Sompel, Nelson 2015; Bermès 2020, p. 52). Furthermore, it is also rather expensive to maintain - especially for institutions that use it only for data consumption by Europeana.

[3] In Appendix 1, there is a table listing the research stakeholders within the ENA who had a direct or indirect impact to the research addressed by this master's thesis (cf. Table 17).

[4] The Europeana service is funded by individual EU member states and by the European Commission via the Europeana Digital Service Infrastructure (DSI), which is in its fourth iteration (DSI-4). Europeana DSI-4 is intended to fulfil Europeana's 2020 strategy and "*provides access to Europe's cultural and scientific heritage*". See (Europeana 2018; 2019a) and project documentation at https://pro.europeana.eu/project/europeana-dsi-4

Europeana also seeks to limit the burden of (semi-) manual labour (such as the scheduling and frequency of updates or the granularity of these updates) when registering and managing the collections that shall be aggregated from its partners' data services. For instance, the updating of cultural heritage objects (CHOs) and associated metadata by content providers is complex to perform because, at the moment, data partners need to flag Europeana manually, when their data has been amended. Below, Figure 1 illustrates Europeana's current operating model.

Figure 1: Europeana's current operating model



(Neale, Charles 2020)

Europeana therefore seeks to find alternative mechanisms to OAI-PMH. It would also help in pushing technologies to Europeana's content providers, which can impact their digital transformation independently of their contribution to Europeana (i.e. these technologies can have a benefit with respect to more general data publication and exchange processes). In particular, Europeana is interested in technologies that are essentially geared towards exchanging higher quality data, namely data with more semantics or better standardisation.

2.2.2 R&D projects and pilot experiments

Europeana has already carried out quite a few tests with different Web technologies[5] to aggregate metadata and links to digitised objects in different ways. Among the aggregation pilots, the following three can be mentioned (Freire, Isaac, Raemy 2020):

- **The Rise of Literacy Project**, which consisted of evaluating the application of Linked Data and the Schema.org data model. It was carried out by the Royal Library of the Netherlands (KB) as a data provider, the Dutch Digital Heritage Network (NDE) as an intermediary aggregator, and Europeana.
- **IIIF aggregation pilots** with the University College Dublin (UCD) and the Wellcome Library where the first dataset was ingested via Sitemaps pointing to IIIF Manifests and the second where the crawling was done via IIIF Collection using the Data Aggregation Lab (DAL).
- **Evaluation of Wikidata for data enrichment** where the usability of Wikidata as a Linked Data source for acquiring richer descriptions of CHOs was evaluated.

Finally, it is also important to highlight that from January 2019 to December 2020 the **Europeana Common Culture** (ECC) project is being carried out. Within this project, there is a Linked Data aggregation functional application led by the Netherlands Institute for Sound

[5] Cf. 3.3 to get more information on these technologies.

and Vision (NISV) to bring the Linked Open Data (LOD) to EDM aggregation route into practice (Freire 2020a; 2020b). The NDE has then developed the LOD-aggregator, a specific pipeline to harvest the data and convert Schema.org into EDM (Freire, Verbruggen, et al. 2019).

2.2.3 Aggregation strategy

In May 2020, Europeana released a new aggregation strategy to "*provide long-term direction of the aggregation of European cultural heritage metadata and content*" (Neale, Charles 2020). This strategy has been adopted with the intention of supporting Europeana's technical infrastructure, in particular Metis, with a view to providing more optimal and swifter publishing options, facilitating data onboarding as well as ensuring improved data quality in a very complex landscape where the three (groups of) stakeholders (CHIs, aggregators, and the EF) have varying motivations, resources and technological skills (Neale, Charles 2020).

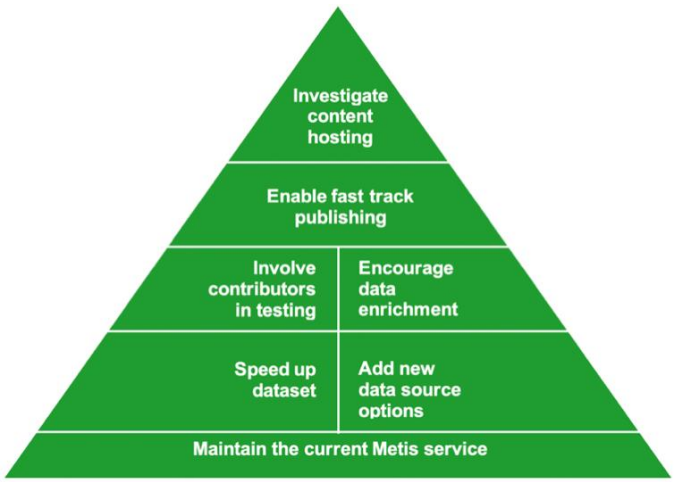
Furthermore, it should also be stressed that the strategy has been produced to be in line with Europeana's global strategy 2020-2025 and more specifically with Objective 1A: "*Develop a more efficient aggregation infrastructure*" (Europeana 2020).

As part of this aggregation strategy, the following seven outcomes have been articulated:

1. Maintain the current Metis service
2. Speed up dataset updates
3. Involve contributors in testing
4. Enable fast track publishing workflow
5. Add new data source options
6. Encourage data enrichment
7. Investigate content hosting

These different outcomes have also been designed to be represented as a conceptual solution evolving over time, as shown in Figure 2 below with the top elements of the pyramid symbolising a longer-term approach.

Figure 2: Aggregation strategy's conceptual solution



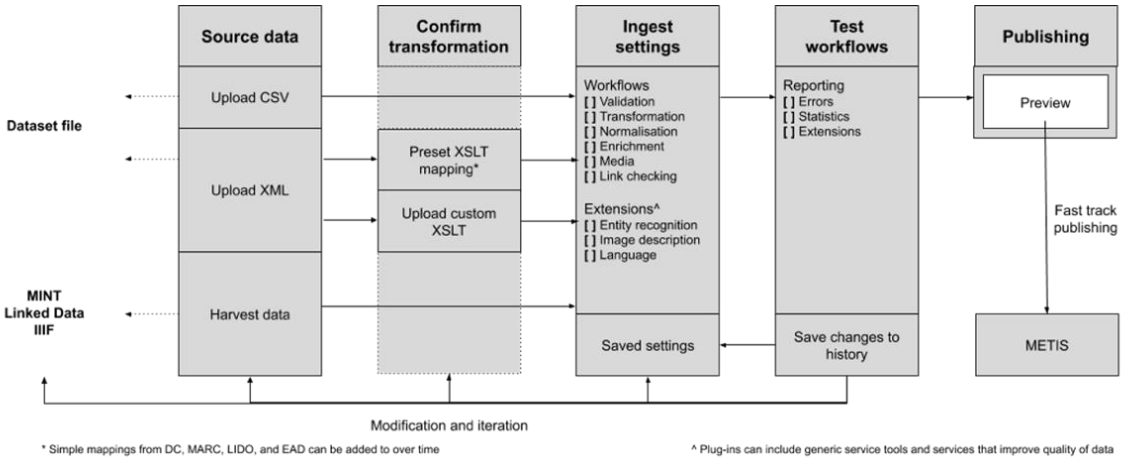
(Neale, Charles 2020)

Some of the outcomes identified in the conceptual solution have a greater impact on this dissertation, particularly those that propose a number of alternative mechanisms for aggregators and CHIs.

For instance, within the third outcome, which aims to involve contributors in testing their data sources before ingestion into Metis, the idea of a Sandbox has been devised. The different functionalities of this tool would range from data import, to data conversion to EDM, and from data enrichment, to preview.

In addition, the fourth, fifth and sixth outcomes of the aggregation strategy, which involve the possibility to have a fast track publishing workflow, to add new data source options (such as IIIF and Linked Data) as well as to encourage data enrichment, all rely on an extended Metis Sandbox concept (cf. Figure 3). Encouraging of data enrichment latter would, for instance, allow the reduction of human intervention, upload data and transform it from common standards, as well as improve the overall data quality (Neale, Charles 2020).

Figure 3: Extended Metis Sandbox Concept



(Neale, Charles 2020)

Moreover, within the strategy itself, a three-stage roadmap was planned with a view to implementing the seven different outcomes. This sequential planning, outlining different tasks, is foreseen to take place over a period of two years (Neale, Charles 2020).

2.3 Research scope

The scope of this research is briefly explained in this section in terms of expectations, constraints as well as research questions and objectives.

2.3.1 Expectations

One of the main expectations of the research was to support Europeana's decisions on the directions for improved data aggregation, notably, but not exclusively, in terms of compliance with EDM prior to ingestion in Metis, traceability upon data update, as well as guidance to data providers and aggregators.

This master's thesis would allow the work already carried out by Europeana's experts in this field to be continued while making new technical and strategic recommendations.

2.3.2 Constraints

The two identified areas that can constrain the realisation of this master's thesis were the dependencies in relation to the management of Europeana's activities as well as the limitations resulting from the current technological landscape and aggregation operating model.

2.3.3 Research questions

This master's thesis sought to address two research questions (in **bold**). The latter led to other interrogations (in *italics*) that necessitated further investigation.

Are there suitable alternatives to OAI-PMH in terms of scalability, ease of use and which covers the requirements of the ENA?

- *Are those technologies covering all the requirements currently supported by OAI-PMH?*
- *What additional features do these technologies bring that are not or are badly supported by OAI-PMH?*
- *Could those technologies complement or replace the existing technologies?*

What are the feasibility conditions to deploy these technologies in the Europeana context?

- *How would adoption of these technologies impact Europeana's current operating model?*
- *How these new alternatives should be presented to CHIs so that they are interested in investing in them?*
- *How could institutions start using those technologies without too much investment?*

2.3.4 Objectives

Three main objectives, themselves divided into several specific ones, were identified.

1) **To provide a brief historical background since the creation of OAI-PMH as well as a comparison between the different methods of aggregating cultural heritage resources and their associated metadata.**
    a. To carry out a state-of-the-art study on the different methods and technologies applied to metadata aggregation
    b. To provide a comparative overview of the various technologies identified for aggregation
    c. To help identify the requirements of the most promising technology that can handle updates and variety of metadata models
2) **To participate in the design and evaluation of prototypes and pilot experiments with the technologies identified**
    a. To gather representative data from Europeana's partner institutions
    b. To establish a procedure with Europeana's R&D team
    c. To conduct and refine tests on different technologies with the help of DAL
    d. To analyse and extrapolate the acquired outcomes
3) **To investigate what services Europeana could offer to encourage adoption of technologies that will gradually be used in place of OAI-PMH**
    a. To assess the impact of leveraging Web technologies to aggregate metadata from Europeana's partner institutions
    b. To suggest different scenarios that conform to the strategy for Metis
    c. To establish recommendations and guidelines to Europeana and partner institutions to reduce non-automated labour
    d. To determine the next steps to be carried out within Europeana and its Network

3. Literature review

This literature review focuses on the different (meta)data aggregation mechanisms in the cultural heritage (CH) domain and how specific technologies enhance resource discoverability and sharing.

First, it gives an overview of some important metadata standards in the CH domain. Then, a dedicated section is devoted to the current aggregation process. Lastly, the literature review covers the different Web-based technologies that can enable (meta)data aggregation.

3.1 Metadata in the cultural heritage domain

If CHIs have common information management goals and interests (Lim, Li Liew 2011), such as providing access to knowledge and ensuring the sustainability of CHOs, they are also characterised in their diversity with respect to the metadata landscape. Each domain has distinct ways of describing the resources they collect, preserve, and showcase. Even within a particular domain, significant differences can still be observed (Mitchell 2013; Freire, Voorburg, et al. 2019).

The purpose of this section is to succinctly present the metadata types and standards used by CHIs, both with respect to their distinct natures, but also in what brings them together, such as through the application of LOD technologies.

3.1.1 Types of metadata

Metadata can be divided into several categories, or types, to serve different data management purposes. Traditional library cataloguing focuses, for example, on the identification and description of resources, but there are obviously other types of metadata that carry valuable insights (Hillmann, Marker, Brady 2008).

According to Zeng and Qin (2016), five key purposes can be distinguished: administrative, technical, descriptive, preservation, and use. These types of metadata can either coexist within the same standard or be the subject of a specific one.

The metadata typology can facilitate the extrapolation of future actions to be performed by individuals in charge of data curation. For instance, technical metadata can be leveraged for collection profiling or format validation (Lindlar 2020) as well as use metadata which can provide indications when or if a given resource can enter the public domain or be freely accessible (Whalen 2016).

Metadata are not fixed statements and can be created and maintained incrementally throughout the data's lifecycle (Baca 2016b; 2016c).

3.1.2 Metadata standards

Metadata standards are critical to establishing structured consistency of information, thereby enabling a common interpretation between different stakeholders, both those who own and those who use the resources. Within the CH domain, the use of metadata "*aids in the identification, assessment, and management of the described entities [users] seek*" (Zeng, Qin 2016, p. 3).

Standards derive and have evolved on the basis of the different cultures of each respective subdomain and the underlying (typical) application focus. In libraries, the value is in the content

and the objects are rather regarded as carriers, in archives, the value is in the collection and in its description, and in museums which have many unique artefacts, the value is in the object (Sanderson 2020a). In addition, the automation of museum and archives collection management took place later than for libraries. As such, interoperability and information exchange between libraries has therefore progressed more rapidly than in other types of CHIs (Jacquesson, Roten, Levrat 2019).

Each subdomain has accordingly created and maintained their own metadata standards, rules and models. Many specifications developed for information resources have also been endorsed by standards bodies (Greenberg 2005) and some of these standards are solely used within a specific domain community (Hillmann, Marker, Brady 2008).

Although this dissertation does not specifically focus on CH metadata standards, Table 2 presents some notable examples of some of the most common and widely used standards[6] along with a short description and the main functions they fulfil. For the latter, the choice of classification was made on the basis of the comprehensive standard visualisation and six of the seven functions of Riley and Becker (2010): *conceptual model, content standard, controlled vocabulary, markup language, record format,* and *structure standard*[7].

Table 2: A selection of metadata standards in the CH domain

Standard	Short description	Functions
Anglo-American Cataloguing Rules, 2nd edition (AACR2)	*AACR2 is a data content standard for describing bibliographic materials* (Baca 2016a).	Content Standard
Bibliographic Framework (BIBFRAME)	*BIBFRAME is a data model for bibliographic description designed to replace the MARC standards and to use the principles of linked data to make bibliographic data more useful within the library community as well as in the broader universe of information* (Baca 2016a).	Conceptual Model Structure Standard Content Standard
Cataloguing Cultural Objects (CCO)	*CCO is a manual for describing, documenting, and cataloguing cultural works and their visual surrogates* (Coburn et al. 2010).	Content Standard Controlled Vocabulary
CIDOC Conceptual Reference Model (CIDOC-CRM)	*CIDOC-CRM is an object-oriented model for the publication and interchange of cultural heritage information* (Baca 2016a).	Conceptual Model Structure Standard
Dublin Core (DC): • **Dublin Core Metadata Element Set** (DCMES) • **Dublin Core Terms** (DCT)	Originally, the Dublin Core Metadata Initiative proposed a set of fifteen metadata elements (DCMES) as a common denominator for metadata mapping. The more recent Dublin Core Terms (DCT) include additional metadata elements for greater precision. Both namespaces can be used for a Linked Data application since the terms are expressed as RDF vocabularies (Baca 2016a; Jaffe 2017).	Structure Standard
Encoded Archival Description (EAD)	*EAD is a data structure standard for encoding archival finding aids in SGML or XML*	Record Format Markup Language

[6] Most of the descriptions come from *Introduction to Metadata*'s glossary edited by Baca (2016a): https://www.getty.edu/publications/intrometadata/glossary/

[7] All of these functions are described in the Terminology (cf. Table 1)

Standard	Short description	Functions
	according to the EAD document type definition (DTD) or XML schema that makes it possible for the semantic contents of a finding aid to be machine processed (Baca 2016a).	Structure Standard
Linked Art	*Linked Art is both a community and a data model based on LOD to describe art* (Delmas-Glass, Sanderson 2020).	Conceptual Model Record Format Structure Standard
Lightweight Information Describing Objects (LIDO)	*LIDO is a simple XML schema for describing and interchanging core information about museum objects* (Baca 2016a).	Record Format Markup Language Structure Standard
Machine-Readable Cataloging (MARC)	*MARC is a set of standardized data structures for describing bibliographic materials that facilitates cooperative cataloging and data exchange in bibliographic information systems* (Baca 2016a).	Structure Standard Record Format Content Standard
Metadata Encoding and Transmission Standard (METS)	*METS is a standard for encoding descriptive, administrative, and structural metadata relating to objects in a digital library, expressed in XML* (Baca 2016a).	Structure Standard Record Format
Resource Description & Access (RDA)	*RDA is a cataloguing standard for libraries which has begun to replace AACR2* (Baca 2016a).	Content Standard Structure Standard
Visual Resources Association (VRA) **Core**	*VRA Core is a data standard for the description of works of art and architecture as well as the digital surrogates that document them* (Riley, Becker 2010; Baca 2016c).	Structure Standard Record Format Controlled Vocabulary

Moreover, the classification of metadata standards according to their functionality is not crisp, i.e. classification decisions may vary depending on perspective. Other classifications of metadata standards in the CH domain are known to sort them according to each subdomain, for instance, the taxonomy of Elings and Waibel (2007) and the metadata blocks' clustering of Mitchell (2013).

3.1.3 Metadata convergence and interoperability

Among the first interoperability efforts in the CH sector were the development of MARC standards in the 1960s to facilitate the exchange of bibliographic data between libraries (Bermès 2011; Baca 2016c). The manner in which libraries could exchange records with each other was facilitated with the establishment of Z39.50, a technology that predates the Web (client-server standard from the late 1970s), which allows one to query different library catalogues (Alexander, Gautam 2004).

At the end of the 1990s, the development of OAI-PMH, founded in the open access movement (Gaudinat et al. 2017), which relies on DC and XML to be a simple denominator for achieving interoperability, is a well-established protocol within the CH domain (cf. 3.2.2 for more information) that bridges the gaps within the broader GLAM (Galleries, Libraries, Archives, Museums) community; even though the protocol is not based on the architecture of the Web (Freire et al. 2017). Alternatively, SRU/SRW (Search and Retrieve URL/Web Service), which

was specified in 2006, is a "*web service for search and retrieval based on Z39.50 semantics*" (Lynch 1997; Reiss 2007), but it is limited to the library domain (Bermès 2011).

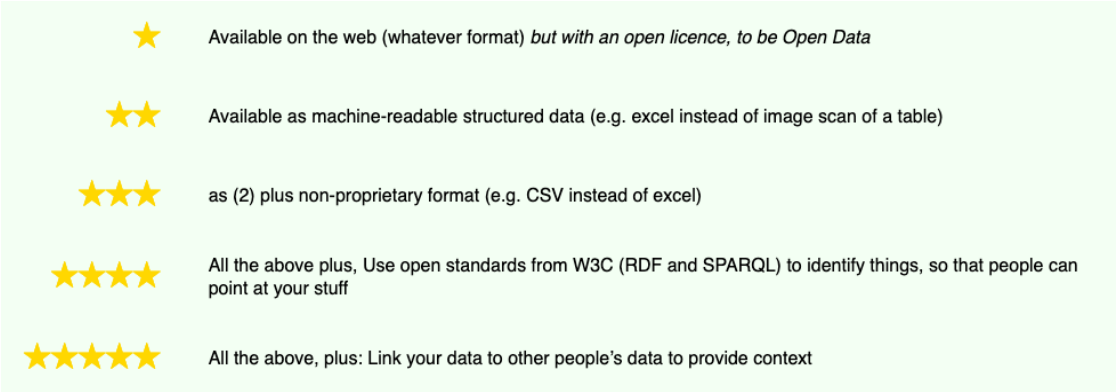
As all of these efforts (with the exception of SRU/SRW) are not based on web standards, they presuppose that end-users will end up on CHI's platforms to discover CHOs (Bermès 2011).

In the last 30 years, the advent of the World Wide Web in 1989 (Berners-Lee, Fischetti 2001) and the emergence of online catalogues in the decades that followed have challenged the traditional functions of cataloguing as well as the metadata standards in use (Bermès, Isaac, Poupeau 2013, p. 20). The Web is a tremendous enabler for CHIs to share information and showcase their collections to a wider spectrum of users. However, in order to facilitate a rich user experience where individuals can navigate seamlessly from one resource to another on CH platforms without being concerned about their provenance, some common denominators must be found, as CHIs have historically maintained siloed catalogues, disconnected from the broader web ecosystem (Bermès 2011).

The Semantic Web, which can be seen as an extension of the Web through World Wide Web Consortium (W3C) standards (Berners-Lee, Hendler, Lassila 2001), enables interoperability based on URIs and the creation of a global information space. It is based on Resource Description Framework (RDF) assertions that follow a subject-predicate-object structure (Bermès, Isaac, Poupeau 2013, p. 37).

The publication of data as part of the Semantic Web requires following certain steps such as using URIs to designate resources, dereferencing HTTP URIs, using W3C standards (RDF, SPARQL), as well as linking its dataset to other endpoints. A generic deployment scheme is the five stars of LOD (as shown in Figure 4 below) initiated by Tim Berners-Lee (2009). This scheme can, for instance, be used to indicate the compliance level and to assess the effort required to reach LOD.

Figure 4: Five stars of LOD



(Berners-Lee 2006)

Historically, the CH domain has been interested in the Semantic Web from the very start. Among others, DC is worth mentioning, which was heavily inspired by RDF (Wolf 1998).

For the past few years, most of the LOD projects in the CH domain have been carried out to expose data to larger audiences, for metadata enrichment, or to facilitate data interlinking (Smith-Yoshimura 2018).

While an increasing number of CHIs, essentially research libraries or national libraries, are involved in LOD publishing (Smith-Yoshimura 2018), it should be noted that, according to a

survey performed in 2018 by the ADAPT Centre at Trinity College Dublin, among 185 information professionals, the greatest challenge lies in the difficulty of integrating and interlinking datasets and that the mapping between traditional CH metadata standards and native LOD models still causes problem (McKenna, Debruyne, O'Sullivan 2018; 2020).

3.2 Aggregation landscape in the cultural heritage domain

This section gives a brief overview of the aggregation landscape in the CH domain, outlining the main national or transnational aggregation initiatives, then providing an overview of OAI-PMH, as well as providing Europeana's requirements in terms of publication on their platform.

3.2.1 Cultural heritage aggregation platforms

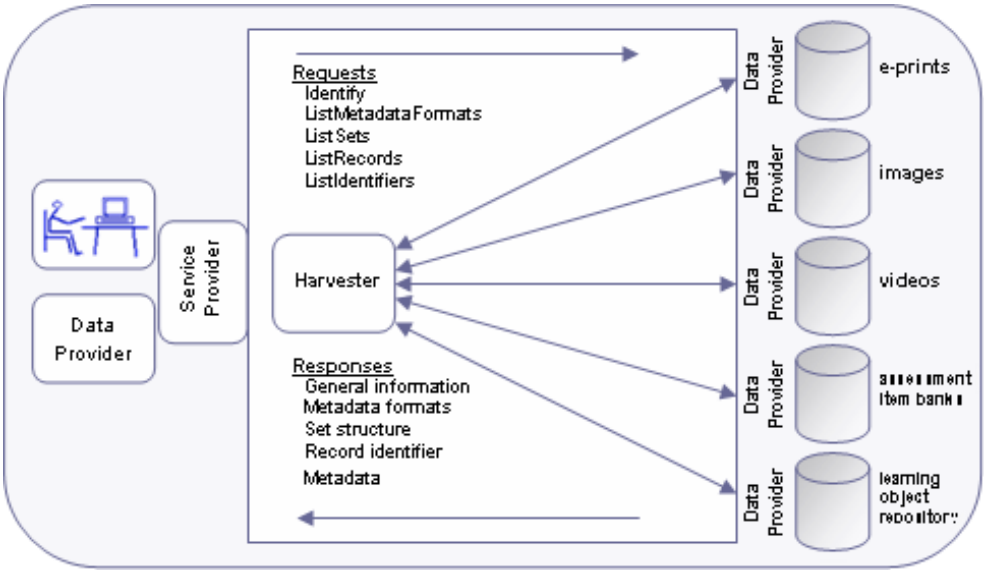
Recent years have seen several national and transnational initiatives set up scalable and sustainable platforms to support resource discoverability in the CH domain, such as DigitalNZ (New Zealand) launched in 2006, Europeana in 2008, Trove (Australia) in 2009, the Digital Public Library of America (DPLA) in 2013, as well as the National Digital Library of India (NDLI) in 2016. However, it should also be noted that all these initiatives still partially or heavily depend on OAI-PMH to harvest metadata.

3.2.2 Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH is an XML-based specification that started in 1999 to improve the discoverability of e-prints through metadata aggregation. The second version of OAI-PMH is the latest stable release of the protocol and was defined in 2002 The data modelling through OAI-PMH relies, but is not restricted to, DC (Lagoze et al. 2002).

OAI-PMH divides the framework's actors into data providers, which provide access to metadata, and service providers, which use harvested metadata to store and enrich their own repositories (Alexander, Gautam 2004). As shown in Figure 5, the protocol defines six requests (or "verbs") that can be issued as parameters for HTTP GET or POST requests: *Identify, List Metadata Formats, List Sets, Get Record, List Record, List Identifiers*. OAI-PMH defines five possible types of responses, encoded in XML: *General Information, Metadata Formats, Set Structure, Record Identifier, Metadat*a (Lagoze et al. 2002).

Figure 5: OAI-PMH Structure



(Lovrečić 2010)

As such, it is an asynchronous protocol that predates the modern architecture of the Web (Bermès, Isaac, Poupeau 2013, p. 36). It is technically not in line with REST principles and can be seen, on a conceptual level, as "*repository-centric*" as opposed to "*resource-centric*" (Van de Sompel, Nelson 2015). Another resulting concern is the fact that OAI-PMH "*has rarely been implemented to its full scale, i.e. benefiting from its incremental harvesting features*" and that the CHO metadata of a data provider evolves separately from that hosted on the service provider's side (Freire, Robson, et al. 2018).

3.2.3 Publication requirements of the Europeana Network

This subsection covers metadata and content requirements that CHIs and aggregators have to follow to publish their data onto the Europeana platform.

Once the Europeana Data Exchange Agreement (DEA) has been signed by a CHI or by an aggregator, there is first a test phase with a potential data partner where a sample of their collection has to be provided to Europeana, either by ZIP or via OAI-PMH (Scholz 2019a).

Europeana's overall data contribution workflow can be divided into three phases: data submission, data processing (cf. Table 18 in the Appendices where the nine relevant ingestion steps in Metis are listed and briefly explained) and data publication.

At Europeana, EDM is the solution that has been found to reconcile the different models as well as to publish LOD (Charles, Isaac 2015). EDM is a generic model based on OAI-ORE, SKOS, and DC among others (Doerr et al. 2010),is an improvement of the Europeana Semantics Elements (ESE), and "*aim[s] at being an integration medium for collecting, connecting and enriching the descriptions provided by Europeana's content providers*" (Charles et al. 2017, p. 8). While this dissertation does not go into detail on EDM, it is important to mention that each CHO issued to Europeana leads to the creation of instances of the following main classes of EDM:

> *"[a] Cultural Heritage Object (i.e.,* `edm:ProvidedCHO` *and* `ore:Proxy` *that represent different data sources for objects), one or more digital representations (i.e,* `edm:WebResource`*) and 'contextual' resources (places, persons, concepts, timespans), in compliance with the one-to-one principle"* (Wallis et al. 2017)

For contributing metadata to Europeana[8], a number of EDM elements are mandatory to ensure that the data and associated metadata are of the highest possible quality (Isaac, Clayphan 2013; Charles, Isaac 2015). The labelling of objects with valid rights statements, through the `edm:rights` property, is also required. The 14 available rights statements[9] come from rightsstatements.org, an initiative led by Europeana, DPLA, Kennisland, and Creative Commons (CC) (Fallon 2015; Scholz 2019a). Additionally, to display CHOs in thematic collections (archaeology, art, fashion, etc.) and hence make them more visible on the Europeana platform, a data partner must provide relevant keywords (Scholz 2019a).

The quality of data contributed to Europeana is measured via different tiers for metadata (A, B, C) and for content (1, 2, 3, 4) within the Europeana Publishing Framework (EPF). In essence, the better rated the data related to a CHO, the more that CHO will be visible on the Europeana platform (Daley, Scholz, Charles 2019; Scholz 2019a; 2019b).

[8] Mapping examples in EDM can be found in Appendix 3 starting on page 66.
[9] Available rights statements: https://pro.europeana.eu/page/available-rights-statements

Finally, it bears remembering that EDM is not a fixed model and is updated in accordance with the needs of the Europeana Foundation and the Europeana Network. As an example, Europeana has recently extended EDM in order to extend its data ingestion framework so that it can accept and recognize IIIF resources[10] (Isaac, Charles 2016; Isaac 2019).

3.3 Alternative web-based technologies for (meta)data aggregation

This section looks at the different aggregation components and technologies which can be a part of alternative mechanisms that Europeana could deploy and propose to its aggregators and CHIs. These technologies may potentially one day supersede OAI-PMH either entirely or partially.

The first subsection introduces the two essential components and the subsequent two subsections outline the underlying technologies in terms of their capabilities and their relevance as an aggregation mechanism. The fourth and last subsection gives a high-level overview of the different aggregation mechanisms that are highlighted in this literature review.

3.3.1 Aggregation components

Several components need to be considered in the aggregation process. Based on Freire et al. 2017, two main categories have been identified:

- Data transfer and synchronisation
- Data modelling and representation

In Table 3, the two aggregation components are briefly described.

Table 3: Aggregation components

Component	Short description
Data transfer and synchronisation	One of the essential components of aggregation is the transfer of data sources from a hosting website to a third party platform, in other words finding a way for an aggregator to collect (meta)data from a CHI. Resources are also likely to evolve on the CHI website and consideration needs to be given to the synchronisation of data sources as well, for instance by using an incremental approach which could be achieved by using a notification mechanism based on Semantic Web technologies.
Data modelling and representation	Data transfer and synchronisation needs to rely upon an agreed data model in order to tackle data heterogeneity. This is all the truer for aggregators showcasing on their platforms data from various domains, like the CH sector. In the case of Europeana, the data model and representation chosen, with which data providers and intermediary aggregators must comply, is EDM. Other data models can be explored, as has already been the case with Schema.org, but some metadata mapping would still be required.

3.3.2 Technologies for data transfer and synchronisation

The technologies for data transfer and synchronisation are organized in alphabetical order.

At the beginning of each section, there is a dotted box indicating one or more namespaces depending on the number of mechanisms tied with the protocol, as well as comment on whether or not a pilot experiment has already taken place within the Europeana Network. Then there is a short presentation of the technology and its relevance for (meta)data aggregation.

[10] IIIF to EDM profile: https://pro.europeana.eu/page/edm-profiles#iiif-to-edm-profile

3.3.2.1 ActivityStreams 2.0 (AS2)

https://www.w3.org/TR/activitystreams-core/ https://www.w3.org/TR/activitystreams-vocabulary/	Pilot experiment already conducted in the context of the IIIF Change Discovery API which leverages AS2.

ActivityStreams 2.0 (AS2) is as much a syntax as a W3C vocabulary, being part of the Social Web protocols series, allowing to represent activity flows, actors, objects and collections in JSON(-LD) and to syndicate them within social web applications (Guy 2017; Snell, Prodromou 2017a; 2017b).

AS2 is generally not used on its own and is a valuable adjunct to help with data transfer and synchronisation. For example, its verbs are used by the IIIF Change Discovery API (cf. 3.3.2.3) and AS2 can be combined with other Social Web Protocols such as AP or Linked Data Notifications (LDN) for the payload of notification requests (Guy 2017; Sanderson 2018; Appleby et al. 2020a).

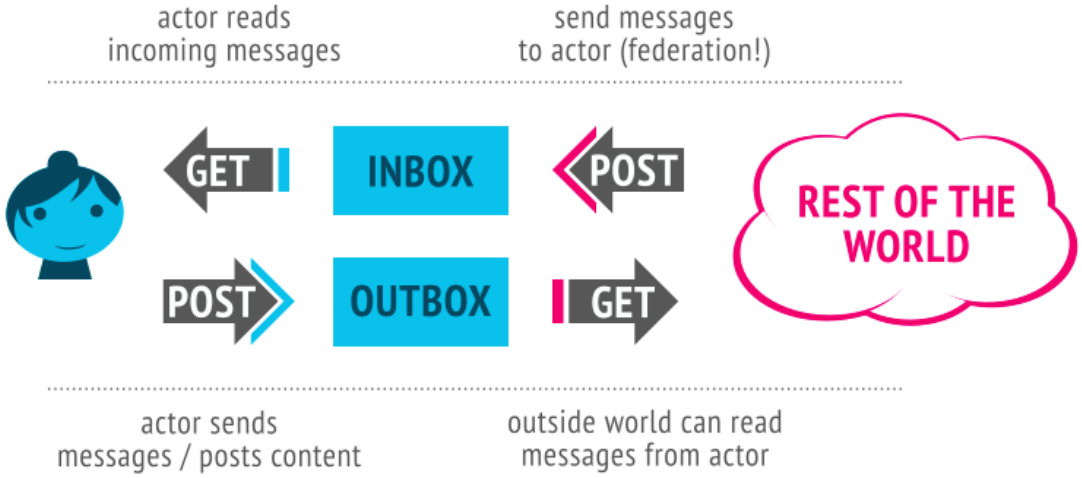
3.3.2.2 ActivityPub (AP)

https://www.w3.org/TR/activitypub/	No previous pilot experiment

ActivityPub (AP) is an open and decentralized W3C standard created in 2018, and is based on the AS2 vocabulary which provides a JSON(-LD) API for client-to-server (publishing) and server-to-server (federation) interactions.

It is part of the suite of Social Web Protocols and its use in metadata aggregation lies in its ability to notify actors across a given network via GET and POST HTTP Requests of each action (or activity) (Lemmer Webber, Tallon 2018). In other words, each actor has an inbox to receive messages and an outbox to send them (Guy 2017). These different boxes are equivalent to endpoints as illustrated by Figure 6.

Figure 6: ActivityPub request examples



(Lemmer Webber, Tallon 2018)

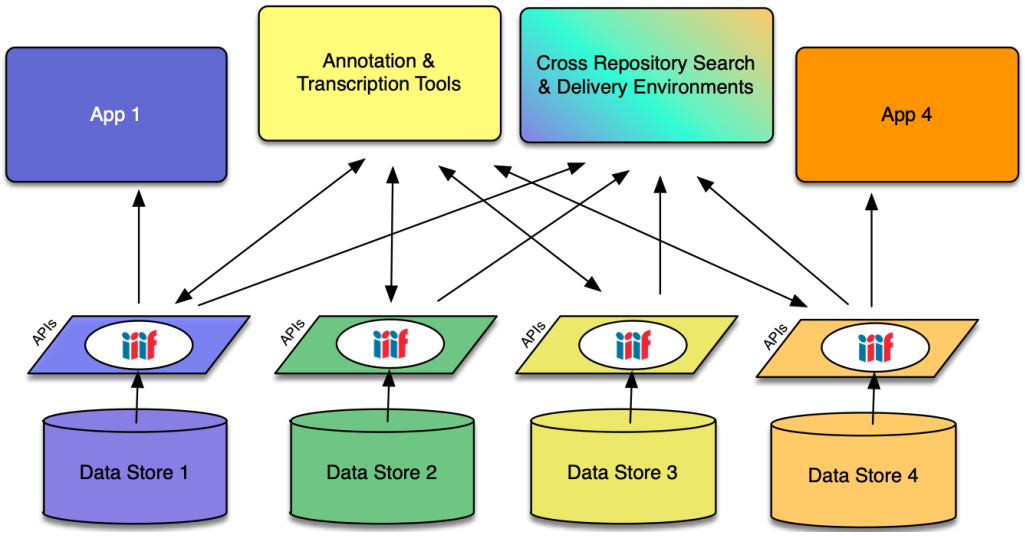
3.3.2.3 International Image Interoperability Framework (IIIF)

https://iiif.io/api/image/2.1/ https://iiif.io/api/presentation/2.1/ https://iiif.io/api/discovery/0.9/	Pilot experiments already conducted with aggregation based on IIIF and Sitemaps, IIIF Collections (collection of IIIF Manifests or collection of IIIF collections), as well as with the Change Discovery API.

https://iiif.io/api/image/3.0/ https://iiif.io/api/presentation/3.0/ https://iiif.io/api/content-state/0.2/	No previous pilot experiment

The International Image Interoperability Framework (IIIF – pronounced '*triple-eye-eff*') is a community, driven-initiative that creates shared APIs to display and annotate digital representations of objects (Hadro 2019). As shown in Figure 7, IIIF has enabled the creation of an ecosystem around web-based images, which consists of various organisations deploying software that comply with the specifications (Snydman, Sanderson, Cramer 2015).

Figure 7: IIIF APIs in the client-server model



(Hadro 2019)

The following are the current stable IIIF APIs, which are all HTTP-based web services serialised in JSON-LD (Raemy, Schneider 2019):

- **IIIF Image API 3.0**[11]: *specifies a web service that returns an image in response to a standard HTTP(S) request* (Hadro 2019; Appleby et al. 2020b).
- **IIIF Presentation API 3.0**: *provides the necessary information about the object structure and layout* (Hadro 2019; Appleby et al. 2020c).
- **IIIF Content Search API 1.0**: *gives access and interoperability mechanisms for searching within a textual annotation of an object* (Appleby et al. 2016; Raemy 2017).
- **IIIF Authentication API 1.0**: *allows application of IIIF for access-restricted objects* (Appleby et al. 2017; Raemy 2017).

While all of these APIs have a strong focus on delivering rich data to end users, they were not specifically designed to support metadata aggregation (Rabun 2016; Warner 2017). For example, there aren't any requirements in terms of metadata standards that accompany IIIF Manifests (which are the representation and description of an object) and there aren't any elements indicating a timestamp for the creation or modification of an object (Freire et al. 2017).

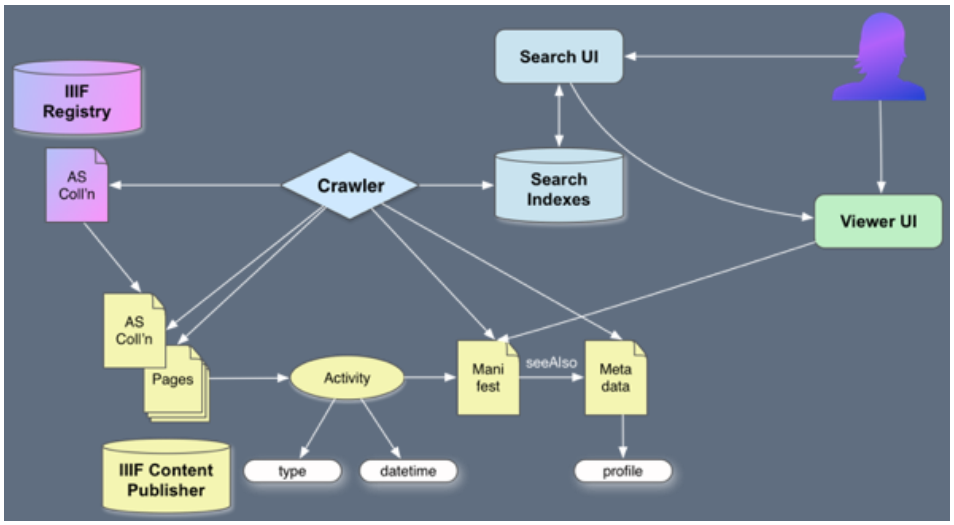
[11] The IIIF Image and Presentation APIs are sometimes referred to as the "core IIIF specifications". They both have been upgraded from 2.1.1 to 3.0 (with breaking changes) to integrate time-based media in June 2020. Most organisations who have implemented IIIF, including Europeana, will need to revamp their infrastructure and their IIIF resources as well to support the newest versions of the IIIF core APIs.

The first issue can be addressed by adding an optional link to structured metadata (through `rdfs:seeAlso`) and the second matter is currently being addressed by the IIIF community through their Discovery Technical Specification Group and the creation of several specifications, namely the Change Discovery and Content State APIs, as well as the creation of a central IIIF registry (Sanderson 2018; Robson et al. 2020).

- **IIIF Change Discovery API 0.9**: *specifies a machine to machine API that provides the information needed to discover and subsequently make use of IIIF resources. It leverages ActivityStreams to describe changes to resources and facilitates crawling to build search indexes* (Sanderson 2018; Raemy, Schneider 2019; Appleby et al. 2020a).
- **IIIF Content State API 0.2**: *describes the current or desired state of the content that a client is rendering to a user. The API allows for standardised approach to deep-linking into objects and annotation from search results* (Warner 2017; Appleby et al. 2019).

An overview of a IIIF Discovery ecosystem enabling a well-defined harvesting process of IIIF resources is illustrated by Figure 8 below.

Figure 8: Overview of a IIIF Discovery ecosystem



(Sanderson 2018)

Lastly, IIIF is, technically speaking, not LOD, but it is in a conceptual sense as it is somewhat "*a visual support for LOD*" (Cossu 2020) and the two frameworks can work alongside each other. For example, Linked Art can be used to reference IIIF resources or services and IIIF can point to a Linked Art description via a `rdfs:seeAlso` property to leverage semantic discovery (Sanderson 2020b).

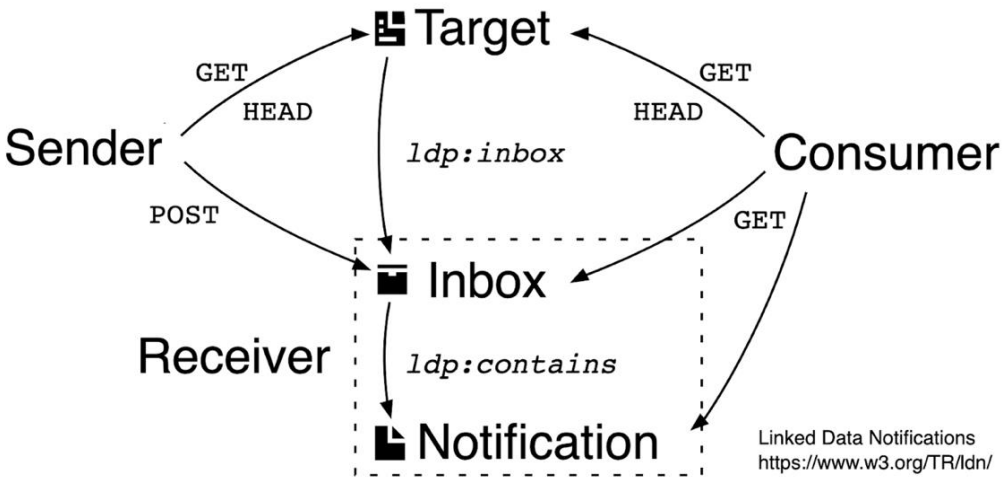
3.3.2.4 Linked Data Notifications (LDN)

https://www.w3.org/TR/ldn/	No previous pilot experiment

Linked Data Notifications (LDN) is a JSON-LD based Social Web Protocol for delivery, facilitating messages sent by servers (*receivers*) to different applications (*senders*) and defining how these applications (*consumers*) can retrieve those messages (Capadisli, Guy 2017). In other words, LDN is a resource-centric protocol where notifications are structured as well as being identifiable and reusable by different applications on the Web. Also, LDN treats notifications as persistent entities (Capadisli 2019).

An overview of the different LDN concepts and the possible HTTP requests are shown in Figure 9.

Figure 9: LDN overview



(Capadisli, Guy 2017)

The interest of LDN in aggregation is its great modularity as it leverages the Linked Data concepts of shared vocabularies and URIs. For instance, the storage of notifications is compatible with LDP, an LDN receiver can understand requests coming from AP federated servers and finally LDN can also draw on AS2 syntax and vocabulary. Finally, a combined implementation with a IIIF ecosystem is also possible and has already been done to connect distributed scholarly discussion (Witt 2017a; 2017b) or as part of exploratory activities carried out by the IIIF Discovery Technical Specification Group[12].

3.3.2.5 Linked Data Platform (LDP)

https://www.w3.org/TR/ldp/	No previous pilot experiment

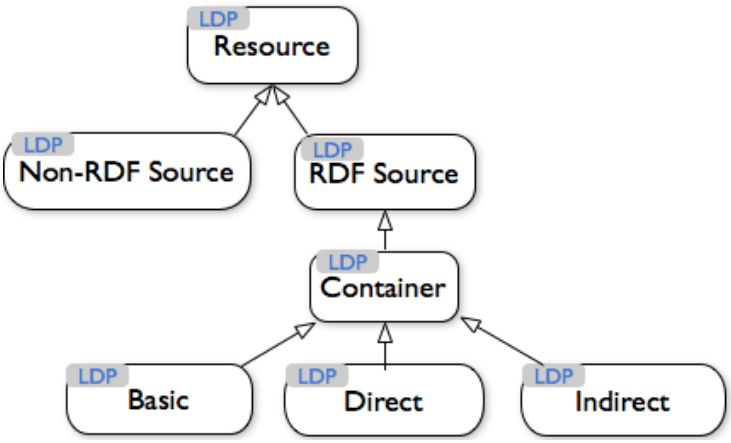
Linked Data Platform (LDP) is a W3C standard based on HTTP requests, some of them on RDF, defining a set of rules allowing a read-write Linked Data architecture. LDP considers everything as resources and can interact with RDF as well as non-RDF sources (Speicher, Arwe, Malhotra 2015).

For RDF sources, the *Container* type has been defined by LDP, representing a collection of linked documents or information resources. Three types of containers have been conceived: a basic one defining a simple link to the information it contains, a direct container adding the notion of membership, and the indirect container that can link to a totally different resource than the one added[13] (Correa 2015). Even though LDP is a W3C standard that predates the Social Web Protocols, an LDP Basic Container can be compared to an LDN inbox (Capadisli 2019).

Figure 10 illustrates the relationships between the different types within the LDP standard.

[12] LDN for aggregation of IIIF Services: https://github.com/nfreire/LDN4IIIF
[13] For more information about LDP Containers, please consult https://gist.github.com/hectorcorrea/dc20d743583488168703

Figure 10: Class relationship of types of LDP Containers



(Speicher, Arwe, Malhotra 2015)

While LDP's potential for aggregation lies in its ability to easily access and manipulate metadata, the issues of pointing to specific harvestable resources and the incrementation part cannot solely be resolved by leveraging this mechanism (Freire et al. 2017).

3.3.2.6 Open Publication Distribution System Catalog 2.0 (OPDS2)

https://drafts.opds.io/opds-2.0	No previous pilot experiment

The Open Publication Distribution System Catalog 2.0 (OPDS2) is a syndication format for electronic publications based on the Readium Web Publication Manifest model and JSON-LD. The second version of the specification is still at the draft level and differs from V1.2 which was based on Atom and an XML serialisation (Freire et al. 2017).

The purpose of this protocol is the aggregation, distribution, discovery and acquisition of electronic publications. If its interest is mostly geared towards e-books, other types of publications can be syndicated (Gardeur 2020). Additionally, the core metadata vocabulary of OPDS2 is Schema.org, which is an advantage over DC that is used in V1.2 since it provides greater expressiveness and can enable better web indexing.

3.3.2.7 ResourceSync (RS)

http://www.openarchives.org/rs/1.1/resourcesync http://www.openarchives.org/rs/notification/1.0.1/notification http://www.openarchives.org/rs/notification/1.0.1/framework_notification	Pilot experiment conducted in the context of aggregation based on extended Sitemaps leveraging elements from the ResourceSync namespace.

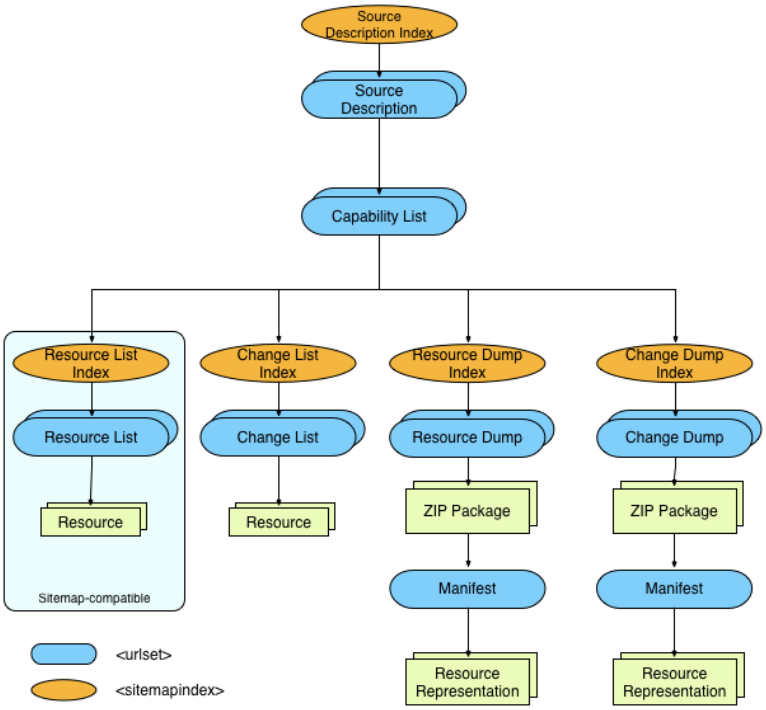
ResourceSync (RS) is a specification issued as a joint effort between the Open Archives Initiative (OAI) and the National Information Standards Organization (NISO). RS, which also known as Z39.99-2017, leverages Sitemaps and adds extensions to the protocol to enable third-party systems to remain synchronised with a data provider's CHOs and their associated metadata. It is also possible to use RS in conjunction with WebSub to establish a notification mechanism (Haslhofer et al. 2013; Klein et al. 2013; Freire et al. 2017).

RS support the following capabilities, which are all serialised in XML as it relies on the Sitemaps protocol (American National Standards Institute, NISO 2017):

- **Resource List**: list and describe the resources that a Source makes available for synchronisation
- **Resource List Index**: a group of multiple Resource Lists
- **Resource Dump**: a link to a package of the resources' bitstreams
- **Resource Dump Index**: a group of multiple Resource Dumps
- **Resource Dump Manifest**: a description of the package's constituent bitstreams
- **Change List**: a document that contains a description of changes to a Source's resources
- **Change List Index**: a group of Change List
- **Change Dump**: a document that points to packages containing bitstreams for the Source's changed resources
- **Change Dump Index**: a group of multiple Change Dumps
- **Change Dump Manifest**: a description of the constituent bitstreams of the package

Figure 11 gives an overview of the framework structure.

Figure 11: ResourceSync Framework Structure



(American National Standards Institute, NISO 2017)

Its relevance for aggregation is that not only can RS synchronize metadata but also content. In addition, RS relies on Sitemaps, a well-known and fairly easily deployable protocol.

3.3.2.8 Sitemaps

https://www.sitemaps.org/protocol.html	Pilot experiments conducted with aggregation based on standard Sitemaps, Sitemaps extended with elements from the IIIF namespace, as well as Sitemaps extended with elements from the ResourceSync namespace

Sitemaps is a protocol that enables webmasters to tell search engines which webpages of a given site are available to be crawled by robots. It consists of an XML file listing URLs and additional metadata (Schonfeld, Shivakumar 2009).

The following three XML tag definitions are required: `<urlset>` which references the Sitemaps protocol and encapsulates the file, `<url>` which encapsulates the other URL entries, as well as `<loc>` which provides the URL of a specific webpage. The other optional tags can provide information about the last update (`<lastmod>`), the change in frequency (`<changefreq>`) and a property value from 0.0 to 1.0 in relation to other pages of the site (`<priority>`).

The relevance of Sitemaps in aggregation is that the protocol is very widespread on the Web and that it can be an entry point to crawl pages that need to be harvested (Freire et al. 2017). It is, in fact, a protocol that can be combined with other mechanisms, or a core technology, for instance like ResourceSync, which is built on it (Haslhofer et al. 2013; Freire, Robson, et al. 2018).

3.3.2.9 Webmention

https://www.w3.org/TR/webmention/	No previous pilot experiment

Webmention is a form-encoding-based protocol for delivery developed by the W3 Social Web Working Group which relies on HTTP and URL Encoded Form (`x-www-urlencoded content`). "*It provides an API for sending and receiving notifications when a relationship is created (or updated or deleted) between two documents*" (a source and a target) (Parecki 2017).

It should also be noted that Webmention and LDN are indeed both intended for delivery and have some overlapping functionality, but they differ in how they handle "*(...) different content types of requests*" (Guy 2017).

The interest that Webmention could have within an aggregation ecosystem is in its ability for a data provider to keep track of when their CHO's URLs are mentioned on a third-party platform, as well as providing a mechanism for a data provider to notify an aggregator which resources should be harvested (Freire et al. 2017).

3.3.2.10 WebSub

https://www.w3.org/TR/websub/	No previous pilot experiment

WebSub, previously known as PubSubHubbub (PuSH), is a W3C standard that is part of the Social Web Protocols which describes an approach for "*(...) subscription of any resource and delivery of updates about it*" (Guy 2017).

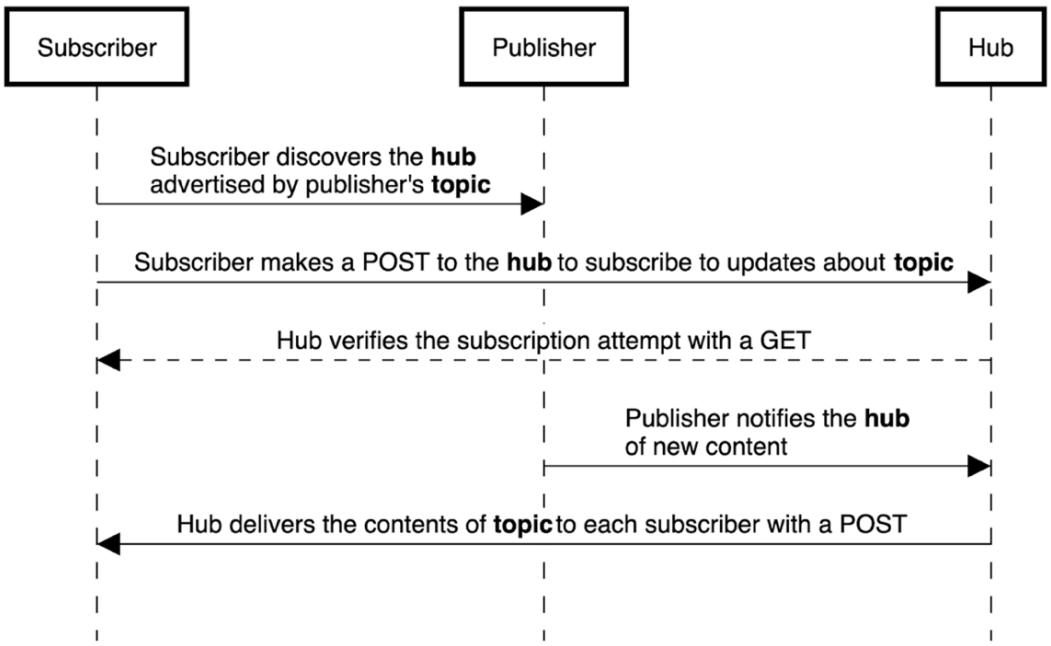
With WebSub, it is the platform hosting the data that itself pushes new content to the aggregators, as opposed, for example, to an RSS feed that must regularly check for updates. To receive these updates though, a subscription over HTTP through a dedicated hub is needed. A hub acts as an intermediary entity relaying "fat ping" notifications[14] (Genestoux, Parecki 2018; Capadisli 2019).

[14] Fat ping is a "*(...) ping which contains a copy of the content that has been changed*" (*fat ping* 2015)

This standard can be seen as a mechanism for communication between publishers and their subscribers where hub and topic URLs can be discovered by looking at HTTP headers of the resource URL (Genestoux, Parecki 2018). For instance, WebSub is leveraged by RS for its notification mechanism (Haslhofer et al. 2013).

Figure 12 below outlines a high-level protocol flow for WebSub.

Figure 12: WebSub high-level protocol flow



(Genestoux, Parecki 2018)

3.3.3 Technologies for data modelling and representation

The presentation of the technologies follows the same logic as the previous subsection. It should also be noted that of the three elements presented here, only Schema.org could really be considered as a complement to EDM and that the two other methods are pieces that can assist in data modelling and representation.

3.3.3.1 Data Catalog Vocabulary (DCAT)

https://www.w3.org/ns/dcat#	Pilot experiment already conducted with aggregation based on Linked Data

Data Catalog Vocabulary (DCAT) is an RDF vocabulary initially created in Ireland by the Digital Enterprise Research Institute and then transferred under W3C governance. DCAT facilitates interoperability between different data catalogues published on the Web (Albertoni et al. 2020).

In terms of its harvesting options, DCAT allows specifying a downloadable dataset distribution as well as referring to a SPARQL endpoint (Freire 2020b).

3.3.3.2 Schema.org

https://schema.org/	Pilot experiments already conducted with aggregation based on Linked Data and based on Sitemaps and Schema.org in HTML pages

Schema.org is the name of the cross-domain vocabulary as well as the initiative that was created by major Internet search engines (Google, Bing, Yahoo and Yandex) in 2011, which

"*(...) seeks to encourage the publication and consumption of structured data on the Internet*" (Freire, Verbruggen, et al. 2019). It allows indexing crawlers to more accurately identify the meaning of indexed pages, sometimes referred as Semantic Search Engine Optimization (SEO) (Wallis et al. 2017). Schema.org can be serialized in Microdata, RDFa as well as JSON-LD (Freire et al. 2017).

Schema.org became a community-based effort in 2015 with the creation of the W3C Schema.org Community Group and its vocabulary maintenance is done through GitHub repositories.

Its key role in aggregation lies in its vocabulary which provides relatively precise descriptions of CHOs, as well as being able to indicate to crawlers where a downloadable dataset distribution is available (Freire 2020a). Last but not least, Schema.org can also facilitate the referencing of web pages (Freire, Charles 2017; Freire, Charles, Isaac 2018).

3.3.3.3 Vocabulary of Interlinked Datasets (VoID)

https://www.w3.org/TR/void/	Pilot experiment already conducted with aggregation based on Linked Data

The Vocabulary of Interlinked Datasets (VoID) is an RDF vocabulary for discovering and leveraging Linked Data sets (Keith et al. 2011). VoID consists mainly of the following two classes:

- `void:Dataset` to describe datasets issued by a single publisher in RDF and accessible either through dereferenceable URIs, via a SPARQL endpoint, or by other methods such as RDF data dumps or the ability to specify a list of URIs.
- `void:Linkset` (subclass of `void:Dataset`) to specify the links between these datasets.

The relevance of VoID in aggregation is to allow a crawler to point towards the appropriate target in several ways (Freire 2020b).

3.3.4 Overview of aggregation mechanisms

Table 25 in the Appendices provides a high-level overview of all the different technologies that are highlighted here.

This overview contains the following categories[15]: the name of the **technology**, the associated **URL**, **version**, **date**, **aggregation component** (data transfer and synchronisation or data modelling and representation), a **short description**, the **governance** bodies, **HTTP Requests** (such as GET, HEAD, POST, etc.), the **serialisations** (XML, JSON-LD, etc.), as well as its **notification** style of network communication (push or/and pull).

[15] Items that cannot be categorised are labelled "N/A" (not applicable).

4. Methodology

This chapter is divided into four sections, first taking into account the general methodological approach applied throughout the master's thesis[16], then focusing on the data collection stage, followed by the subsequent data analysis as well as the limitations of the different methods that were applied.

4.1 Overall approach

From a methodological point of view, a mixed approach (qualitative and quantitative) was used to address the research questions.

The qualitative methods consisted mainly of regular interviews with collaborators of Europeana, conducting a literature review, taking part in the testing and documentation phase of the Europeana Creative Common's LOD-aggregator functional application, as well as the assessment and conduct of aggregation pilots.

As for the quantitative methods, an online survey provided metrics on the use, interest and awareness of different aggregation mechanisms. In addition, some outputs generated during the aggregation pilots also yielded quantitative figures.

With the aim of validating the research and analysis carried out, informal and punctual interviews were conducted with relevant stakeholders throughout the dissertation. It is worth pointing out, though, that no verbatim accounts were produced, and the minutes were recorded in the author's internal logbook. Most of these meetings tended to turn into constructive discussions, working sessions or even demos. Table 4 outlines the different interactions that were conducted. It lists the dates, the names of the people involved (cf. Table 17 in the Appendices to have more information on the mentioned stakeholders, especially their role) as well as the main meeting objectives.

Table 4: Validation interviews

Date	Stakeholder(s)	Objectives
05.03.2020	Valentine Charles	Giving a demo of Metis and explaining Europeana's current operating model
27.03.2020	Nuno Freire	Providing clarification on LD aggregation and on the functionalities of DAL
17.04.2020	Cosmina Berta, Enno Meijers, Erwin Verbruggen	ECC – defining the next steps and onboarding of the author
20.04.2020	Nuno Freire, Enno Meijers, Erwin Verbruggen	ECC LOD Functional Application – documenting the pilot and its results
23.04.2020	Cosmina Berta, Erwin Verbruggen	ECC – outlining the sustainability aspects of the different ECC functional applications

[16] Trello, a Kanban board software, was used for the management of the overall project: https://trello.com/b/w1Cb85vd/

Date	Stakeholder(s)	Objectives
23.04.2020	Antoine Isaac, Nuno Freire	Comparing the different aggregation mechanisms and discussing the deployment requirements of each mechanism
20.05.2020	Antoine Isaac, Nuno Freire, Albin Larsson	Analysing the survey findings and the future aggregation pilots 1
20.05.2020	Nuno Freire, Enno Meijers, Erwin Verbruggen	ECC LOD Functional Application – assessing the report on documentation and functionalities of the LOD-aggregator pipeline carried out by the author
29.05.2020	Antoine Isaac, Nuno Freire, Albin Larsson	Analysing the survey findings and the future aggregation pilots 2
12.06.2020	Antoine Isaac, Nuno Freire, Albin Larsson	Analysing the survey findings and the future aggregation pilots 3
15.07.2020	Nuno Freire	Taking a decision on the feasibility of aggregation pilots

4.2 Methods of data collection

The data were collected in a variety of ways throughout the master's thesis, but all stemmed from a collaborative effort and mutual understanding with Europeana R&D team.

Indeed, these data collections were carried out thanks to active participation by the author in their weekly catchups, through one-to-one meetings with Antoine Isaac, R&D Manager, as well as by means of ad hoc meetings listed in Table 4.

4.2.1 Reviewing the state-of-the-art

In order to fully capture the technological and strategic stakes of improving (meta)data within the field of cultural heritage, an extensive literature review (cf. 3) has been conducted[17].

Based on this literature review, a comparison of various aggregation mechanisms was then realised (cf. 3.3.4).

4.2.2 Europeana Common Culture's LOD Functional Application

Following a discussion with the members of Europeana R&D team at the beginning of April, it was agreed that there were significant crossovers between this study and the LOD Functional Application carried out within the ECC project (such as the willingness to improve metadata harvesting).

The author was therefore involved in a couple of meetings related to the LOD Functional Application and a few others concerning the ECC project in general. The main outcomes of

[17] The relevant resources, which are not necessarily all included in this dissertation's Bibliography, were recorded on a public library on Zotero: https://www.zotero.org/groups/2446985/ch_aggregation_discovery

this collaboration were the review of the LOD-aggregator (cf. 5.1.1) and the co-authoring of a submitted conference paper (cf. 5.1.3).

4.2.3 Survey on alternative aggregation mechanisms

The survey on alternative aggregation methods (cf. 5.2), which was later retitled "*survey on alternative aggregation mechanisms*" as most technologies have to be combined with one another to provide an aggregation mechanism, was, to some extent, the centrepiece of this master's thesis as it enabled to identify trends and interests of different data partners, and it was also useful for devising potential future pilots.

The following parts succinctly outline the timeline and promotion of the survey, its objectives, structure as well as the hypotheses that were considered.

4.2.3.1 Timeline and promotion

Firstly, a test was carried out with Europeana R&D team in April to validate the questions as well as to correct any grammar and spelling errors and ensure that the flow of questions worked.

After verification, the survey was conducted online through Google Forms[18] and was available from 20 April to 8 May 2020.

The call for participation was published on several channels, including EuropeanaTech's listserv and Twitter account[19], on the author's Twitter account, on a dedicated Europeana channel within IIIF's Slack instance. It was also presented through a lightning talk on the first day of the Europeana Aggregators Forum (EAF) on 6 and 7 May 2020. On EuropeanaTech's listserv, two announcements were sent out, the very first one on 20 April and a reminder on 4 May (see both messages in Appendix 4 on page 68).

4.2.3.2 Objectives

The main objective was to gauge the awareness, interest, and use of technologies other than OAI-PMH for (meta)data aggregation. The main target audiences of the survey were the data providers and the aggregators of the Europeana Network, albeit it was decided to keep it open to other organisations and individuals working in the CH field.

The secondary objective of the survey was to identify possible pilot experiments that Europeana could conduct with interested organisations.

4.2.3.3 Structure and questions

The survey was divided into nine sections with a total of fifteen questions (ten mandatory and five optional). As shown in Figure 31 in Appendix 5, two sections were only shown to participants depending on the answers given to a preceding question (cf. Appendix 6 on page 70 to see all survey questions).

[18] https://forms.gle/iq2fZ8wCqBMGTrDq6
[19] https://twitter.com/EuropeanaTech/status/1252163772652929024

4.2.3.4 Hypotheses

The following three assumptions were made prior to the launch of the survey:

- ResourceSync or W3C's Social Web Protocols (ActivityStreams, Linked Data Notifications, Webmention, WebSub) are relatively unknown and rarely used within the CH domain.
- Header Dictionary Triples (HDT), an RDF binary format, which was created to compress large sets of data and facilitate query scalability (Vander Sande et al. 2018), is still quite recent in the LOD sector and certainly a rarity in the CH field.
- The variety of metadata standards is very important and the number of in-house "flavours" of these standards used within the Europeana Network is quite high. The survey wasn't aimed to get a thorough view of the metadata landscape though, but rather to get an idea of what metadata mappings would be necessary.

4.2.4 Assessment of potential aggregation pilots

On the basis of the survey findings, the interest of the participants and the available data and existing implementations, an assessment was carried out to determine the feasibility of aggregation pilots (cf. 5.3). The following courses of action were identified:

- Carrying out an initial triage among the survey respondents who expressed an interest in an aggregation pilot.
- Selecting the appropriate aggregation routes.
- Contacting the relevant organisations to inform them on the feasibility of a pilot and/or to request additional information if necessary.
- Reaching a decision on whether or not a pilot could be conducted.
- Conducting the aggregation pilots that could be done in the allotted time.

4.3 Methods of data analysis

This section provides further information in terms of the tools and the service design method used during the data analysis phase.

4.3.1 Tools

Three main types of tools were applied for data analysis: spreadsheet software, command-line interface (CLI), as well as web-based prototypes.

4.3.1.1 Spreadsheet software

For the analysis of the different aggregation mechanisms, the results of the online survey as well as for the production of a few charts, standard spreadsheet software, both MS-Excel, for backup and Google's own application, to facilitate easy collaboration on several files, were employed.

4.3.1.2 Text editor

The review of the LOD-aggregator was done by forking the repository from GitHub. Then, the capabilities were tested using a command-line interface (CLI) to test the various functionalities as well as a text editor to display the available datasets produced.

4.3.1.3 Europeana R&D tools as testbed

The DAL as well as the Europeana Metadata Testing Tool, two web-based prototypes set up by Nuno Freire, were utilised to test various aggregation mechanisms.

4.3.2 Service design

The “Opportunity Solution Tree” template, a four-step visual aid which maps out connections to serve a desired outcome (Becker 2020), was chosen to formulate the different scenarios that can enable better aggregation and discovery of CH content.

The propositions stemming from that visual representation were then linked to the Europeana Strategy's priorities and further broken down into suggested steps.

4.4 Limitations

The possible methodological limitations of this study are mainly:

- the possible representation (sample bias) of the survey participants which does not fully reflect the comprehensive nature of all CHIs (cf. 5.2.3 for more details);
- the significant involvement of experts in the field of LOD and IIIF throughout the entire study, as these people are very keen on deploying new protocols at a relatively early stage.

In addition, the time constraints did not allow some potential pilot aggregations to take place because it required too much work on the data partners' side to adjust their data models or protocol implementations (cf. 5.3.3) or for anyone else to help them doing so.

5. Results

This chapter on the results of the master's thesis is divided into three sections and first highlights the author's contribution to the ECC LOD Functional Application, then presents the online survey findings, and, thirdly, features the aggregation pilots.

5.1 Analysis of ECC LOD Functional Application

The participation within the ECC project consisted of gathering stake-holders' considerations on the sustainability of the various project's components, assessing the functionalities and the conformity of the documentation of the LOD-aggregator, the pipeline created for the LOD Functional Application, as well as the submission of a paper for the 2020 Metadata and Semantics Research (MTSR) conference.

Besides the elements presented in this section, it is also worth mentioning that during the EAF, which took place online at the beginning of May, there was a lightning talk co-presented by the author about the activities on alternative aggregation mechanisms carried out by Europeana R&D and its data partners in recent years, a call for participation in the online survey (cf. 5.2), as well as an account of the goals of the ECC LOD Functional Application and its related technical infrastructure (Raemy, Freire 2020).

5.1.1 Sustainability discussions

As the different project outcomes of the ECC project need to be sustained for a period of three years, a sustainability effort was undertaken through a series of meetings to first determine whether or not these outcomes could be further developed into production and what requirements would be necessary. For the LOD Functional Application, the following two sustainability points were addressed by the author:

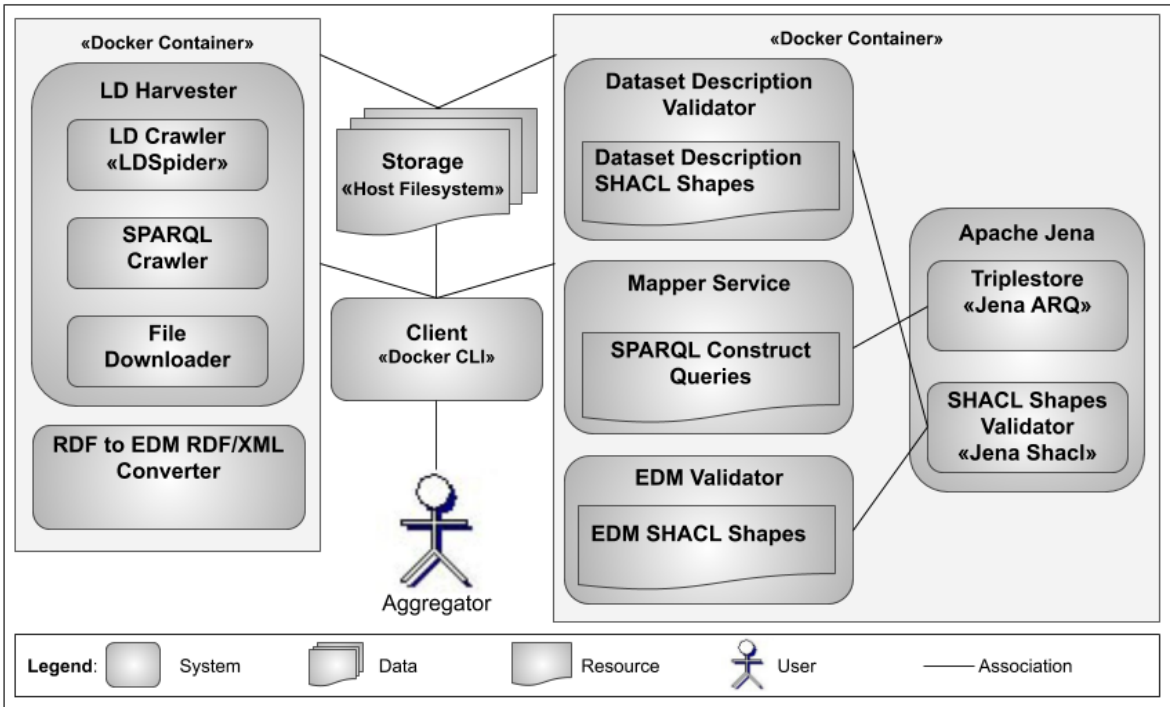
- Looking for new datasets and pilots, in particular as a follow-up to the online survey (cf. 5.2).
- Assessing the usability and possible integration of the system within the Metis Sandbox.

5.1.2 Assessment of the LOD-aggregator

The LOD-aggregator, a generic open-source toolset based on Docker containers for harvesting and transforming LOD for ingest into Europeana, was assessed with respect to its documentation and for testing the various functionalities.

The toolset leverages Docker for flexibility and scalability reasons as it "*allows aggregators to deploy only part of [it], according to their needs*" (Freire et al. 2020). Its main components are the Dataset Description Validator, the LD Harvester, the Mapper Service, the EDM Validator, as well as the RDF to EDM RDF/XML Validator (cf. Figure 13). It should also be mentioned that the toolset works as a CLI.

Figure 13: High-level architecture of the LOD-aggregator



(Freire et al. 2020)

Based on an internal Europeana document on criteria for selecting software components[20] as well as a guide on how to document code hosted on GitHub, the following seven assessment criteria were taken into account: *Value proposition, Licence, Maintenance, Functionality testing, Documentation, Versioning, Quality/Security*.

Although no problems were found regarding the functionalities, and although test runs for all the datasets that went through the pipeline were easily executed, there were some concerns about the terminology, a few typos, as well as the rationale of the toolset which was not clearly stated. Indeed, the value proposition was not sufficiently explicit.

Some aspects, such as maintenance or versioning, were considered minor and were not properly accounted for at this stage of the evaluation. Finally, the criterion concerning quality/security was not addressed by the author, having judged that he did not necessarily have the required expertise. Table 5 provides an exhaustive account of what was assessed.

Colour markings of Table 5			
All in order	It doesn't appear to be a concern at this time.	Some improvements are needed.	This criterion was not assessed

Table 5: Assessment criteria of the LOD-aggregator

Criteria	**Feedback**
Value proposition	The value proposition is not highlighted well enough, some tags should be added as well as a summary in the README. This aspect should be clearly articulated in the report to demonstrate the added value for Europeana, aggregators and data providers.

[20] Internal document under preparation by the Europeana Platform Services

Criteria	Feedback
Licence	European Union Public Licence (EUPL) v.1.2.
Maintenance	There are no open issues, but there aren't many contributors[21]. Also, as it was a functional application part of the ECC project, the maintenance would need to be assessed at a later stage.
Functionality testing	Installation: The *.env* file step should be explained before using the crawl command with an example. Otherwise, it is quite straightforward.
	Crawler, Mapper, Validator, Export, Convertor, Zip: they all work well
Documentation	Functionalities aspects are well-documented.
	There are still a couple of typos though and some labels need to be consistent across the repository. Missing information in the README (project name, a description/summary, a table of content, contributing, credits)
	In the near future, it would be interesting to set up a wiki, containing for example more in-depth tutorials and a FAQ. Important acronyms (ECC, LOD, etc.) should also be fleshed out the first time they appear.
Versioning	No version/no release.
Quality/ Security	Code vulnerabilities and critical issues should perhaps be evaluated through an audit assessment.

To overcome these issues, a couple of pull requests were suggested to developers who incorporated them between 5 May and 20 June 2020[22].

5.1.3 Metadata and Semantics Research (MTSR) paper

A paper written by five individuals, including the author of this dissertation, titled "*Metadata Aggregation via Linked Data in Europeana: results of the Common Culture project*" was submitted on 1 August 2020 to the 14th International Conference on Metadata and Semantics Research. The authors shall be notified at the beginning of September 2020 whether or not this paper has been accepted in the conference proceedings.

[21] Some Europeana recommendations for maintenance recommend the existence of "an active and sufficiently large community". But there is no explicit indication on what this means.

[22] https://github.com/netwerk-digitaal-erfgoed/lod-aggregator/commits/master

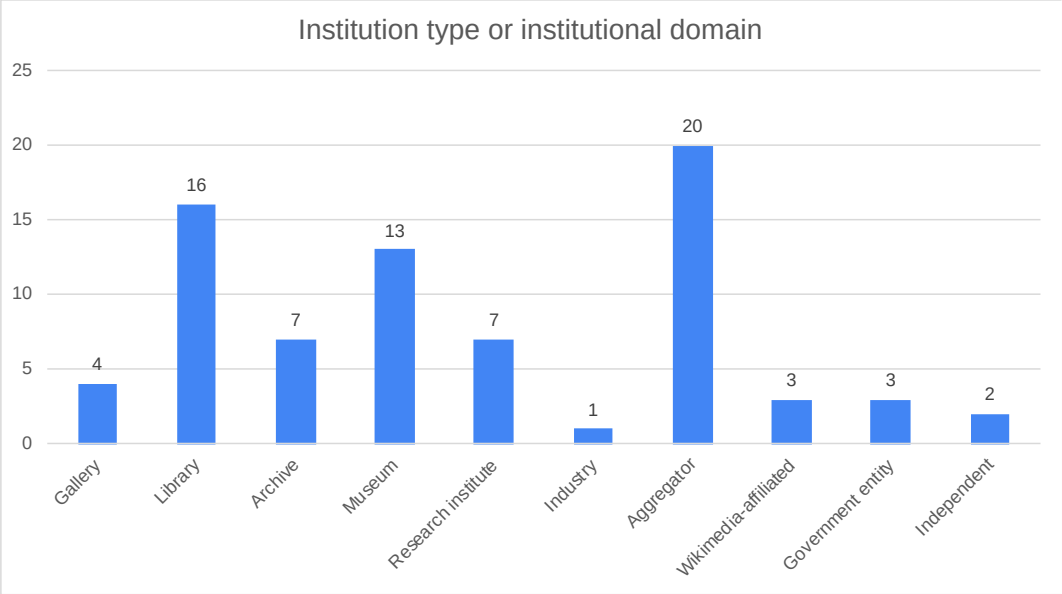
5.2 Survey

This section is divided into three parts starting with a short subsection on the number and provenance of survey participants, followed by a substantive subsection on the findings and closing with any potential biases.

5.2.1 Number and provenance of participants

A total of 52 participants completed the survey. Aggregators (20 occurrences: 38.5%[23]), libraries (16: 30.7%) and museums (13: 25%) were the three most commonly selected types of affiliation[24]. Some of the affiliations mentioned in the "Other" category by respondents were grouped together. These include three organisations or volunteers that were identified as part of the Wikidata community and classified as "Wikimedia-affiliated". Apart from the latter affiliation, each identifiable organisation was only accounted once in the survey (cf. Figure 14).

Figure 14: Typology of survey participants



The country was not asked, but extrapolating from organisation names, one can see that respondents came from 20 different countries and that except in three cases (Brazil, Israel, and England), all participants were from the European Union (cf. Table 6). The highest participation by country was Lithuania (six times), followed by Belgium, Germany and Italy (each four times). In some cases, designating a country was not possible. In total, there were nine instances where a specific country could not be ascertained, and to address this issue, two categories were created: international (seven occurrences) for thematic aggregators and N/A (two occurrences) when extrapolation was not possible.

Table 6: Survey participants' provenance

Country	Occurrences
Lithuania	6
Belgium, Germany, Italy	4

[23] Unless otherwise indicated, the 100% is measured with respect to all participants (N = 52), even for questions where multiple responses were possible.

[24] Note that the choice was not exclusive here. The overlap is one-quarter with 13 survey participants who checked off several options, with a very large majority from aggregators.

France, Sweden, The Netherlands	3
Czech Republic, Greece, Ireland,	2
Brazil, England, Estonia, Hungary, Israel, Latvia, Poland, Romania, Slovakia, Spain	1
International	7
N/A	2

5.2.2 Findings

The sequence of questions outlined in this section highlighting the survey findings does not strictly follow the online survey structure, but clusters the questions thematically, although maintaining a chronological order. In addition, some information such as names of institutions, LOD and IIIF endpoints as well as emails are not disclosed in this dissertation[25].

In addition, a summary of the survey findings presented in this dissertation was included in an official deliverable of the Europeana Digital Service Infrastructure (DSI-4) project (Freire, Isaac, Raemy 2020).

5.2.2.1 Metadata for publishing and exchanging purposes

The survey demonstrates the wide variety of metadata and serialisations used or known by the participants (cf. Figure 15), giving a fairly representative sample of the different sub-domains of the CH field as well as the requirements that national or thematic aggregators expect for ingestion.

The metadata standards that participants are most familiar with (without necessarily using them) are Schema.org (27 occurrences: 51.9%), CIDOC-CRM (26: 50%) and EDM (22: 42.3%).

As for the deployment side, Dublin Core (33: 63.4%), EDM (28: 53.8%), MARC (25: 48%), LIDO (14: 26.9%) and MODS (12: 23.1%) are, in order, the standards most used by survey participants. Schema.org, which was identified as having a sufficient level of expressiveness for CHOs and become a potential complement to EDM for data modelling and representation, is used by 11 survey participants (21.2%).

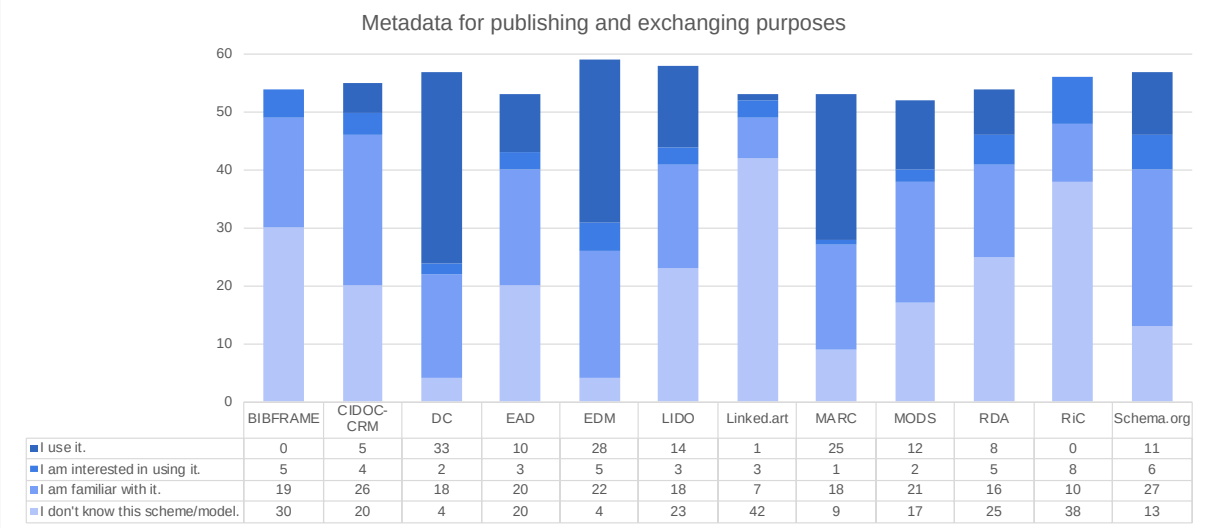
If the survey shows that Schema.org is still in its early phase of adoption within the CH domain, its interest is growing. When Europeana started to investigate back in 2016, they had indeed only been able to find cases of Schema.org usage outside of Europe (Wallis et al. 2017).

On the other hand, BIBFRAME and RiC were never indicated as standards being used for publication and exchange purposes and Linked.art was only mentioned by one participant. The latter three are also the least known standards, which isn't a real surprise, considering that they are fairly new and that each of these standards is rather aimed at a particular subdomain.

The interest in using any of the specific standards is rather limited, out of those mentioned most often, RiC was selected eight times (15.4%), Schema.org six times (11.5%), and BIBFRAME, RDA as well as EDM had each five occurrences (9.6%).

[25] The anonymised version of the survey responses is accessible here: https://doi.org/10.5281/zenodo.3966693 (Raemy 2020a)

Figure 15: Awareness, use, and interest in metadata standards for publishing and exchanging purposes[26]



	BIBFRAME	CIDOC-CRM	DC	EAD	EDM	LIDO	Linked.art	MARC	MODS	RDA	RiC	Schema.org
I use it.	0	5	33	10	28	14	1	25	12	8	0	11
I am interested in using it.	5	4	2	3	5	3	3	1	2	5	8	6
I am familiar with it.	19	26	18	20	22	18	7	18	21	16	10	27
I don't know this scheme/model.	30	20	4	20	4	23	42	9	17	25	38	13

Participants also had the opportunity to cite other metadata they use. METS was mentioned four times and ESE as well as in-house variations (of metadata standards mentioned beforehand) were each mentioned (or hinted) three times. A total of 26 different instances of standards were cited by the participants and 15 of these 26 instances were mentioned once (cf. Table 7).

Table 7: Additional metadata standards

Metadata standards	Occurrences
METS	4
In-house variation, ESE	3
ABCD, IIIF, EAC-CPF, EAG, MADS, UNIMARC, CARARE Metadata Schema	2
OAI-PMH, Omeka XML, SOCH/K-samsök, ONIX, SKOS, ArCo Ontology, DCAT, PICA, Z39.50, Datacite, ResourceSync, EN19507 (Cinematographic Works Standard), Spectrum, PLMET, DNZ, JATS	1

5.2.2.2 Metadata serialisations

As shown in Figure 16, the vast majority of participants are aware of or use one or more metadata serialisations (CSV, JSON, MARCXML or MarcXchange, RDFa, RDF serialisations, XML).

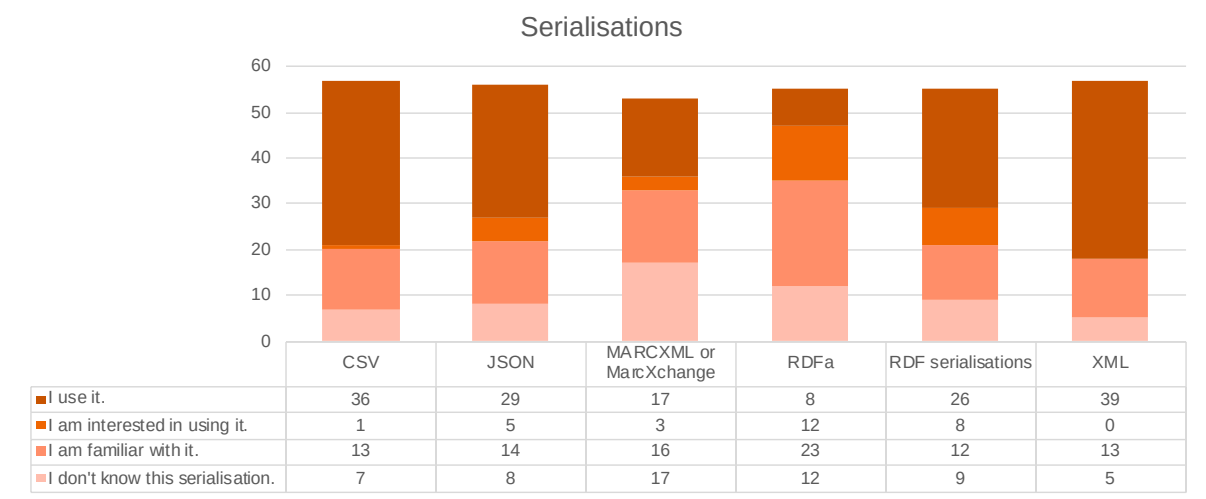
The serialisation that has the highest number of occurrences in terms of awareness (without necessarily being used) is RDFa (23 occurrences: 44.2%). MARC serialisations (MARCXML and MarcXchange) ranks second with 16 instances (30.8%), but it is also the least known among the participants (17: 32.7%). The latter can be explained because this serialisation type is the only one to be almost exclusively used by libraries.

XML (39: 75%), CSV (36: 69.2%) and JSON (26: 50%) are the most commonly used serialisations. Half of the survey respondents use one of the RDF serialisations (RDF/XML,

[26] Participants could select more than one answer per question but were required to choose at least one answer, so the total figure per item amounts to a minimum of 52.

JSON-LD, Turtle, etc.), and while RDFa is the least used serialisation (8: 15.4%), it is also the one that respondents are most interested in (12: 23%).

Figure 16: Awareness, use, and interest in metadata serialisations[27]

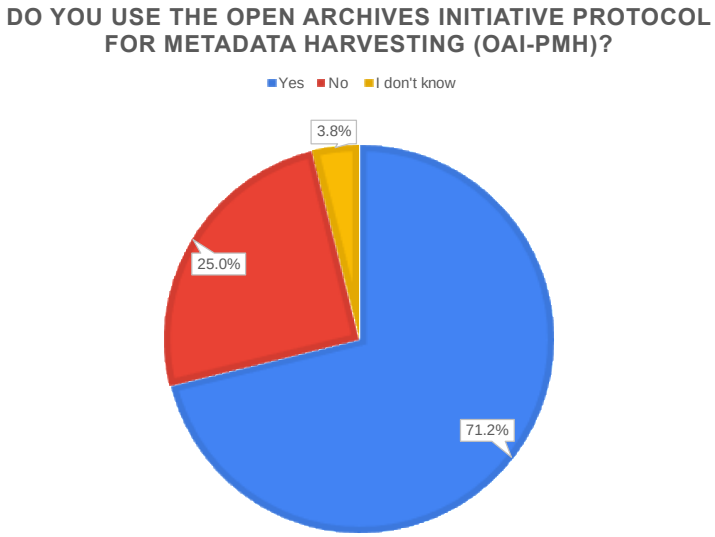


	CSV	JSON	MARCXML or MarcXchange	RDFa	RDF serialisations	XML
I use it.	36	29	17	8	26	39
I am interested in using it.	1	5	3	12	8	0
I am familiar with it.	13	14	16	23	12	13
I don't know this serialisation.	7	8	17	12	9	5

5.2.2.3 OAI-PMH

OAI-PMH, the current technical solution for metadata aggregation into Europeana, is a standard used by several aggregation efforts. 37 survey participants (71.2%) use OAI-PMH for aggregation purposes, 13 don't (25%) and 2 do not know (3.8%) whether they use this protocol (cf. Figure 17).

Figure 17: Use of OAI-PMH

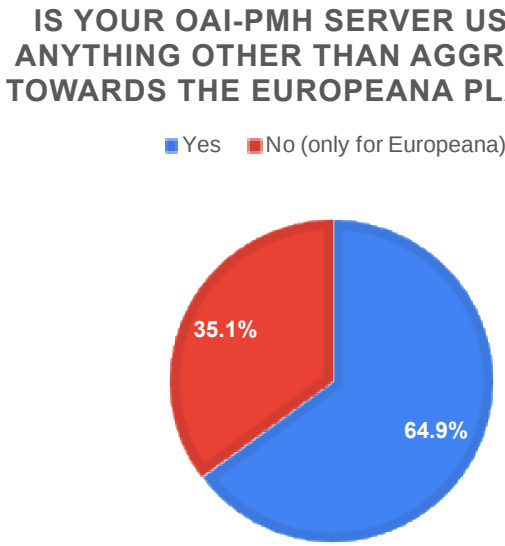


Of these 37 participants who use OAI-PMH, more than a third do so only in the context of aggregation towards the Europeana platform (13 out of 37: 35.1% – cf. Figure 18).

Among these 13 participants who use OAI-PMH only in this context, the great majority of them did not know, test or implement the alternative aggregation methods that were outlined in the survey. However, it is worth noting that two IIIF-related aggregation mechanisms (based on IIIF Collections or on Sitemaps) as well as aggregation via Sitemaps and Schema.org were all mentioned twice (cf. Figure 19 to consult the answers of all participants).

[27] Ibid.

Figure 18: Use of OAI-PMH in the Europeana context

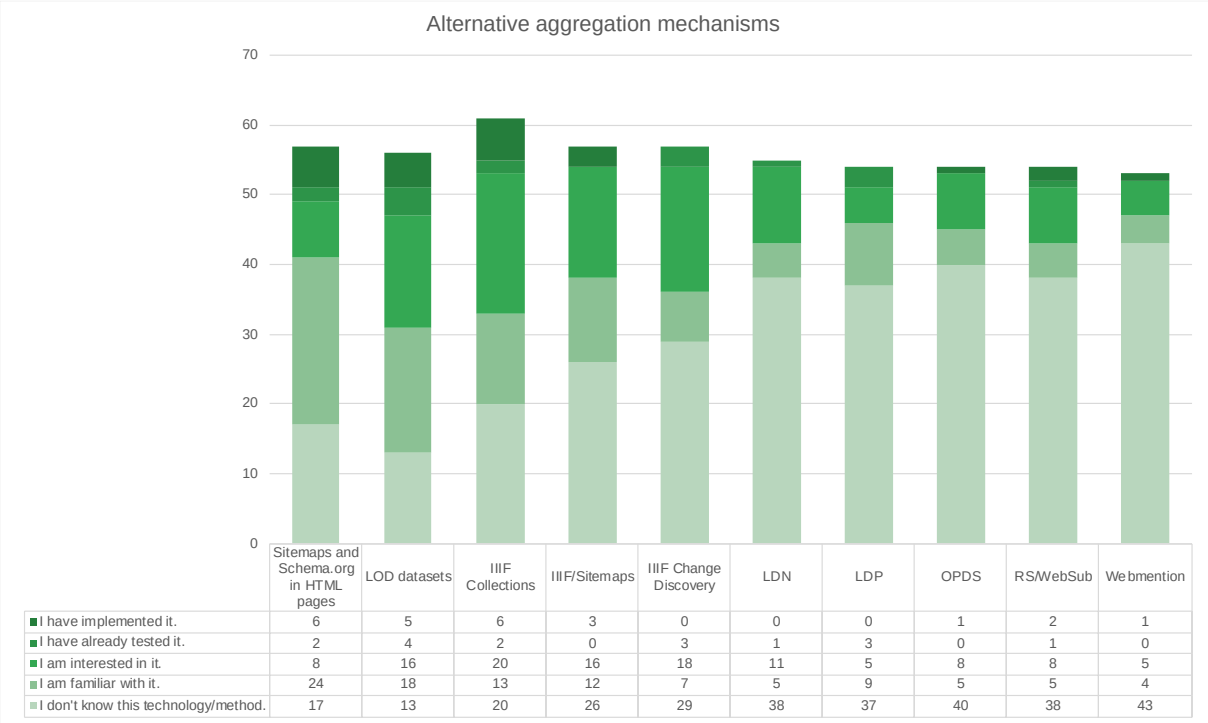


5.2.2.4 Alternative aggregation mechanisms

With respect to alternative aggregation mechanisms, out of the 10 proposed methods, only a very small fraction has ever been tested or implemented by survey participants and most of these methods were also unknown to them. It is worth noting that aggregation via IIIF collections and the mechanism combining Sitemaps and Schema.org are the ones that have been the most implemented (6 occurrences - 11.5% - each).

The participants are particularly interested in all IIIF-related mechanisms (between 16 and 20 participants responded that they were interested in using one of the three methods - between 30.8% and 38.5%), in the aggregation of LOD datasets (16: 30.8%) as well as in LDN (11: 21.2%).

Figure 19: Awareness, use, and interest in alternative aggregation mechanisms[28]



	Sitemaps and Schema.org in HTML pages	LOD datasets	IIIF Collections	IIIF/Sitemaps	IIIF Change Discovery	LDN	LDP	OPDS	RS/WebSub	Webmention
I have implemented it.	6	5	6	3	0	0	0	1	2	1
I have already tested it.	2	4	2	0	3	1	3	0	1	0
I am interested in it.	8	16	20	16	18	11	5	8	8	5
I am familiar with it.	24	18	13	12	7	5	9	5	5	4
I don't know this technology/method.	17	13	20	26	29	38	37	40	38	43

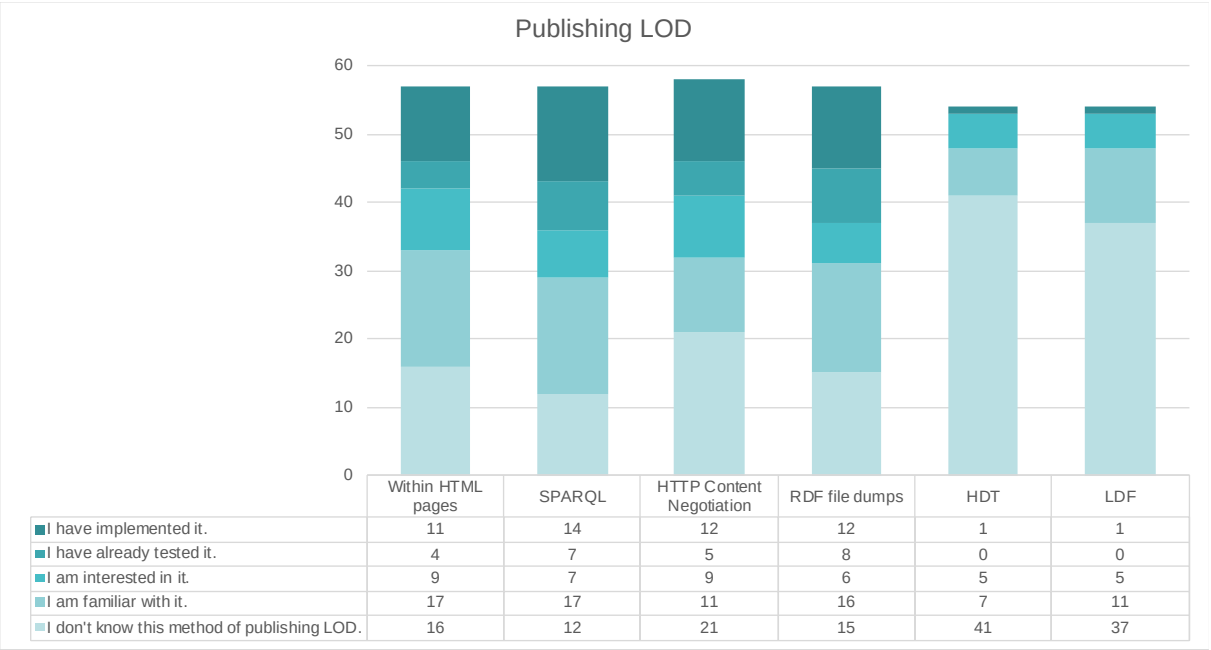
[28] Ibid.

5.2.2.5 LOD

While the methods for publishing LOD are familiar to about a quarter among the survey participants (ranging from 7 to 17 occurrences: 13.5-32.7%), most of these methods remain largely unexplored. For instance, HDT and LDF are largely unknown to the participants, with only one implementation (by the same participant).

SPARQL is the means which participants implement the most in order to publish LOD (14 occurrences: 28%). HTTP Content Negotiation and providing RDF file dumps (both at 12: 23.1%) and publication of LOD inside HTML pages (11: 21.2%) follow. It should be noted that seven participants responded that they had implemented SPARQL, HTTP Content Negotiation as well as RDF file dumps, indicating that the degree of co-occurrence is almost two-thirds.

Figure 20: Awareness, use, and interest in publishing LOD[29]



	Within HTML pages	SPARQL	HTTP Content Negotiation	RDF file dumps	HDT	LDF
I have implemented it.	11	14	12	12	1	1
I have already tested it.	4	7	5	8	0	0
I am interested in it.	9	7	9	6	5	5
I am familiar with it.	17	17	11	16	7	11
I don't know this method of publishing LOD.	16	12	21	15	41	37

6 other ways of publishing LOD were raised by participants. A dedicated API was mentioned twice, and all other ways were referred once (cf. Table 8).

Table 8: Additional ways to publish LOD

Means to publish LOD	Occurrences
API	2
CSV, RDF generated on-the-fly though OAI-PMH, Vocabularies in SKOS RDF, RAW JSON dumps	1

Regarding the following question prompting for LOD examples and endpoints, 19 participants responded and 38 links pointing to LOD data were provided.

[29] Ibid.

5.2.2.6 IIIF

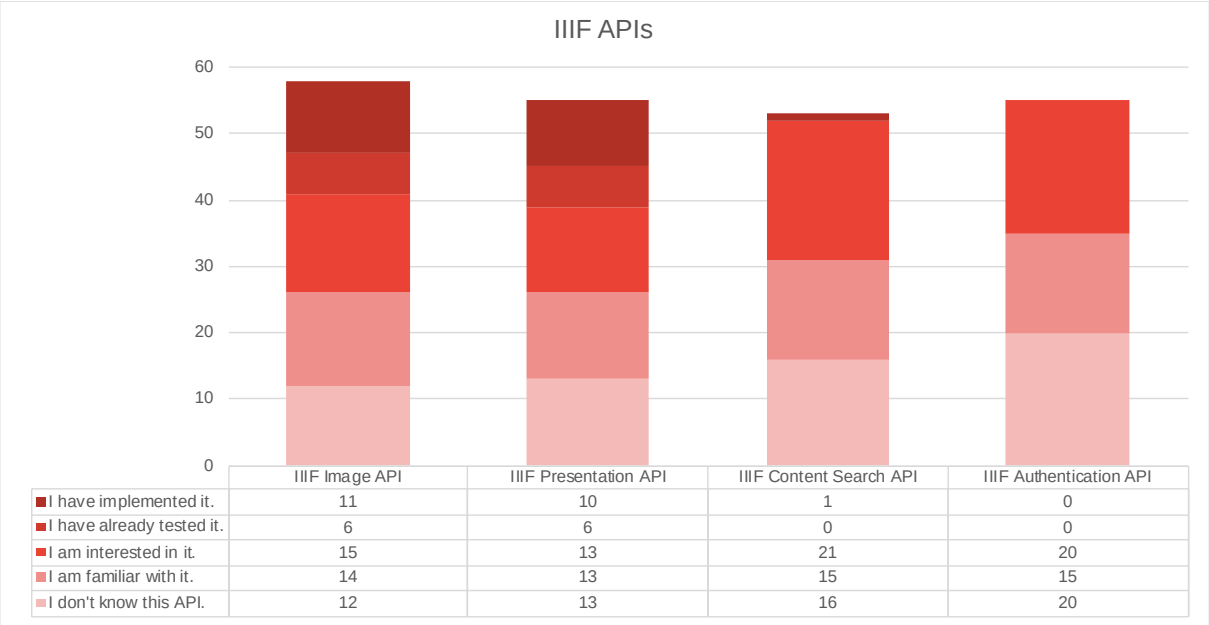
IIIF-based aggregation mechanisms are the ones where Europeana sees the most potential for innovating metadata aggregation in the shorter term. The survey asked participants about their awareness and experience in implementing the IIIF APIs, including the IIIF Presentation API which is key for accessing metadata via IIIF.

Regarding the awareness of the IIIF APIs, participants have a good understanding of the four specifications ranging from 60 to 75% (excluding those who answered "I don't know this API").

As for the implementation, the IIIF Image API is the most deployed specification among the participants (11 occurrences: 21.2%), followed by the IIIF Presentation API (10: 19.2%) and the IIIF Content Search API (once). All institutions that have deployed the IIIF Presentation API have also deployed the IIIF Image API (both APIs are often referred to as the "IIIF core APIs"), which is expected since the latter works in conjunction with the former.

The IIIF Authentication API has never been tested or implemented, which is in line with a survey conducted by IIIF in 2017 (Rabun 2017).

Figure 21: Awareness, use and interest in IIIF APIs[30]



	IIIF Image API	IIIF Presentation API	IIIF Content Search API	IIIF Authentication API
I have implemented it.	11	10	1	0
I have already tested it.	6	6	0	0
I am interested in it.	15	13	21	20
I am familiar with it.	14	13	15	15
I don't know this API.	12	13	16	20

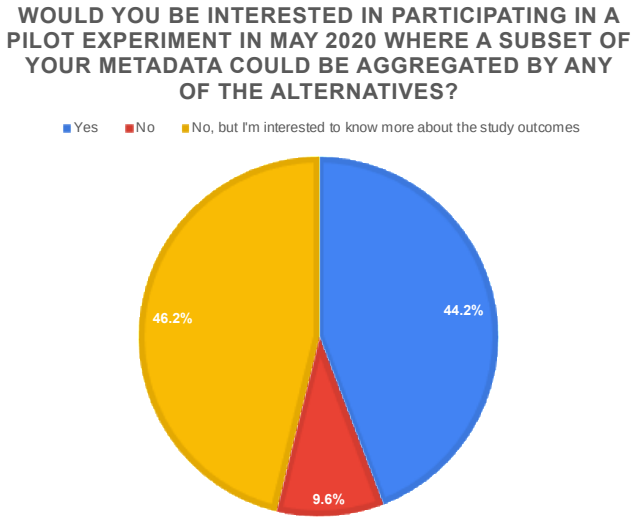
In addition, nine URLs (IIIF Manifests, Canvas and other landing pages related to IIIF) were given by six participants.

[30] Ibid.

5.2.2.7 Possibility of further experiments

Regarding the pilot experiment phase, of the 52 participants, 23 responded positively (44.2%) to have a subset of their metadata used to experiment with an alternative aggregation route. 24 (46.2%) were also interested in the results of the research without necessarily having to take part in an aggregation pilot (cf. Figure 22).

Figure 22: Interest in pilot participation



An investigation phase started in June (as opposed to May as hinted in the question) with interested organisations to understand what work would be required to best carry out aggregation pilots (cf. 5.3).

5.2.2.8 Feedback

17 people left comments in the last question and a few individual emails have been also received or sent by the author in response to certain requests, stemming either from the last question or the call for participation sent on EuropeanaTech's listserv. Below, listed in Table 9, is a summary of the written feedback as well as the conversation that took place during a break-out session of the EAF on standards and frameworks on May 6 that was led by Henning Scholz, Partner and Operations Manager at Europeana's Data Publishing Services.

Table 9: Survey participants' feedback

Area	Summary	Source
Aggregation methods	One of the solutions put forward by a participant for aggregation would be to deploy a SPARQL endpoint (it was not mentioned if the person was referring to the SPARQL that Europeana has already implemented). One respondent also indicated that the survey did not adequately explain the relevance of the methods advanced in the survey and the relationship between these methods and the overall ingestion into Europeana. With respect to the highlighted methods, one participant was perplexed that Signposting was not identified in the survey.	Survey
Cooperation between Europeana and Wikidata	There were a few comments on the use of Wikidata pre- or post-ingestion and that a better cooperation could be made between Wikidata and Europeana on data reconciliation/enrichment.	Survey

Area	Summary	Source
Monitoring and synchronisation	A point was made that what was missing was not necessarily the underlying technology, but the fact that error reports should be better monitored and that the people involved in metadata aggregation should be better instructed. It was also pointed out that it'd be interesting to have some kind of "*sync contracts*", which would provide a timestamp on the harvested data and that perhaps it'd be also useful if data providers could have direct access to Metis.	Survey
Metadata standards and serialisations	For some aggregators, it was difficult to give a factual answer on the use of specific standards and serialisations because the metadata delivered to them by data providers are not necessarily the ones they use.	Survey
IIIF	Some participants had difficulty in providing a IIIF endpoint or landing page. One participant also reacted to the fact that their organisation does not have a top-level collection because it would represent a very large number of IIIF Manifests. Finally, one person had questions about the use of IIIF in the context of publishing audio-visual resources and an email was sent to point them to the more appropriate entities.	Survey, email
Ease of use	Given that data providers and aggregators have already invested both time and resource to maintain OAI-PMH servers, a few respondents reported concerns should Europeana prefer other technologies for aggregation In addition, they hoped that alternative aggregation methods could be a low-barrier gateway for both aggregators and data providers.	EAF
Ground truth	It was also a concern that some data providers had not even implemented OAI-PMH and that aggregators sometimes have to do web scraping on institutions' websites to retrieve the necessary data. These CHIs do not have the technological or business skills and/or the necessary resources to provide their data seamlessly.	EAF, email

5.2.3 Survey biases

There are a fairly large number of aggregators compared to the other user groups, namely CHIs or individuals, who participated in the survey. This is likely due to the significant number of aggregators that completed the survey during the EAF window.

The data providers who responded to the survey are almost always large organisations, with potentially far more resources than small or medium-sized organisations. It is also difficult to determine whether the participants who responded were in fact the most appropriate people (within their own organisation) to fill out the survey.

Those who necessarily want to lead the way in terms of aggregation will respond more readily to this type of survey. The participation of well-informed users who are relatively familiar with the IIIF APIs or on how to publish LOD has certainly influenced the results.

Last but not least, the lockdown situation due to the COVID-19 crisis may have biased participation in the survey, both with individuals who had more time to complete the survey and with individuals who were technically unemployed or lacked access to the resources needed to complete the survey accurately.

5.3 Aggregation pilots

This section reports on the interest shown by online survey respondents in participating in an aggregation pilot, as well as the follow-up work that was carried out. It gives the overall parameters that first define the type of aggregation pilots envisaged and the subsequent analysis of whether and when they could take place.

5.3.1 Parameters for defining and assessing potential pilots

In order to provide an overall basis for the aggregation pilots, the following few parameters were determined:

- Identifying possible aggregation routes, with their different processes, prerequisites, and relevant resources.
- Selecting the aggregation routes for each interested organisation based upon the existing metadata, the interest on the proposed aggregation alternative, as well as on the use of LOD and IIIF.
- Avoiding using the same method with interested organisations who already conducted pilots with the Europeana R&D team (such as the University College Dublin, Wellcome, KB, National Library of Wales).
- Giving priority to data providers and aggregators of the Europeana Network.

5.3.2 Identifying aggregation routes for potential pilots

For any aggregation routes, the generic ingestion process can be summarised as follows (some of these steps may be optional):

- Importing/Crawling the data
- Mapping to EDM (or Schema.org)
- Validating the converted data
- Publishing

Table 10 lists thirteen aggregation routes based on the technologies identified in 3.3, identifying the relevant technologies for the steps above and the resources that support their deployment in the Europeana context.

The list includes some combinations of mechanisms that have never been experimented on by Europeana R&D. As this list is intended to be as exhaustive as possible, some of the deployment requirements are not yet known (which are labelled “TBD” – to be determined). The different aggregation routes are reviewed in terms of prerequisites (on the data provider’s side) and some of the necessary resources needed to carry them out.

Table 10: Alternative aggregation routes

ID	Aggregation route	Associated technologies and/or dependencies	Data provider prerequisites	Resources[31]
AR01	Aggregation via Sitemaps and Schema.org in HTML pages	Sitemaps Schema.org	Sitemaps, structured data in RDFa using the Schema.org vocabulary within each HTML pages containing CHOs	Guidelines for providing and handling Schema.org metadata in compliance with Europeana, DAL
AR02	Aggregation via LOD datasets based on dataset distribution	VoID DCAT Schema.org	RDF description of a dataset available via a downloadable distribution, using DCAT, VoID, or Schema.org	Specifying a linked data dataset for Europeana and aggregators, Guidelines for providing and handling Schema.org metadata in compliance with Europeana, DAL, LOD-aggregator
AR03	Aggregation via LOD datasets based on listing of URIs	VoID Schema.org	Using the `void:rootResource` properties that contain the URIs of the CHOs, which should point to RDF resources in EDM or in Schema.org	
AR04	Aggregation via Linked Open Data datasets	VoID DCAT SPARQL	SPARQL endpoint's URL must be specified with a property from VoID or DCAT.	
AR05	Aggregation of IIIF based on IIIF Collections	IIIF Image API IIIF Presentation API	IIIF Manifests with a link to structured metadata in `rdfs:seeAlso` (EDM or Schema.org) Collection of IIIF Manifests for aggregation	Awesome IIIF, DAL, Muzz.app (mapping from LIDO)
AR06	Aggregation of IIIF based on Sitemaps	Sitemaps ResourceSync IIIF Image API IIIF Presentation API	IIIF Manifests with a link to structured metadata in `rdfs:seeAlso` (EDM or Schema.org) Extended Sitemaps with elements from the namespaces of IIIF	Awesome IIIF, DAL

[31] For more details, i.e. the URL of the resources as well as a brief description of their features or purpose, see Table 24 in the Appendices.

ID	Aggregation route	Associated technologies and/or dependencies	Data provider prerequisites	Resources[31]
AR07	Aggregation of IIIF based on the IIIF Change Discovery API	IIIF Image API IIIF Presentation API IIIF Change Discovery API AS2	IIIF Manifests with a link to structured metadata in `rdfs:seeAlso` (EDM or Schema.org) ActivityStreams endpoint	
AR08	Aggregation of IIIF based on LDN	IIIF Image API IIIF Presentation API LDN	LDN implementation of a *Consumer*	TBD
AR09	Aggregation via ResourceSync in conjunction with WebSub	Sitemaps RS WebSub	Extended Sitemaps with RS namespaces	TBD
AR10	Aggregation via LDN and storage over LDP	LDN LDP	TBD	TBD
AR11	Aggregation via Webmention	Webmention	TBD	TBD
AR12	Aggregation via OPDS2	OPDS2	TBD	TBD
AR13	Aggregation via ActivityPub Delivery	AP AS2	Deployed AP endpoints (inbox and outbox) Messages compliant with ActivityStreams Vocabulary	TBD

5.3.3 Assessment of potential aggregation pilots

It was necessary to decide whether a pilot could be feasible or not for the 23 organisations that expressed interest.

5.3.3.1 Triage of potential pilots

The first assessment stage is based on an analysis of the information filled in by the survey respondents and (occasionally) a short investigation on their different web portals. The five following statuses were established:

- **Out of scope**: either there were no data available and it was impossible to track the survey participant, or it was suppliers providing services that allowed

aggregation, which would put them in a different position since they do not own any data.

- **Dismiss**: the survey participant had not implemented any mechanism allowing for (meta)data aggregation and/or expressed no interest in any technology.
- **Defer**: the survey participant had not implemented any mechanism allowing for (meta)data aggregation even though they had expressed interests in one or more technologies.
- **Investigate**: the survey participant had implemented one or several mechanisms allowing for (meta)data aggregation and expressed interest in one or more technologies.

As shown in Table 11, three participants were put in the "Out of scope" category, four in "Dismiss", two in "Defer" and 14 survey respondents were labelled as "Investigate".

Table 11: Triage on the conduct of potential aggregation pilots

Total	Outcome			
	Out of scope	Dismiss	Defer	Investigate
23	3	4	2	14

5.3.3.2 Aggregation route selection

The retained options in terms of aggregation routes were made on the basis of the survey results, as well as when the deployment requirements were known and ultimately if resources were available.

As a result, the following five aggregation routes were selected:

- **AR01**: Aggregation via Sitemaps and Schema.org in HTML pages
- **AR02**: Aggregation via LOD datasets based on dataset distribution
- **AR04**: Aggregation via LOD datasets based on SPARQL
- **AR05**: Aggregation of IIIF based on IIIF Collection
- **AR06**: Aggregation of IIIF based on Sitemaps

5.3.3.3 Follow-up emails

The next step was to write follow-up emails to contact the different respondents according to the preliminary decisions and also to inform the 24 people interested in the study findings (cf. 0) that a summary of the survey results was available online.

Five email templates were drafted as it was decided at that point to split in two the organisation types in the “Investigate” category, distinguishing “typical” aggregators and CHIs from entities affiliated with Wikimedia who play a particular role in terms of data enrichment. The type and number of follow-up emails are displayed in Table 12. Each template is also available in the Appendices on page 75.

Table 12: Type and number of follow-up email templates sent

Total	E-mail templates				
	Template 1 Interest in the study outcomes + out of scope	**Template 2** Dismiss	**Template 3** Defer	**Template 4** Investigate	**Template 5** Wikimedia-affiliated
47	27 (24+3)	4	2	12	2

It is also worth mentioning that the door to carry out an aggregation pilot was never closed and in each instance all email recipients were invited to contact a Europeana R&D team member at a later stage should they become interested in a mechanism or if they had managed to successfully implement a given technology.

For the respondents receiving the "Investigate" email template, it was required that they specify one or several of the selected aggregation routes. In addition to a response deadline set on 30 June 2020 (follow-up emails were sent by mid-June), several questions were asked in order to obtain additional information that was necessary to decide whether or not a pilot could be conducted.

Out of the twelve emails sent with template 4, 17 aggregation routes were proposed (an average of 1.4 per recipient). Table 13 lists the typical questions asked to respondents and the number of times the aggregation routes were suggested.

Table 13: Typical questions raised in the follow-up emails (template 4)

Aggregation routes	Questions[32]	Occurrences
Aggregation via Sitemaps and Schema.org in HTML pages (AR01)	• Does your Sitemap point to the pages containing cultural heritage objects? • Do these webpages containing CHOs have structured metadata in HTML?	3
Aggregation via LOD datasets based on dataset distribution (AR02)	• Are your Linked Data datasets available in EDM (or Schema.org) or should a mapping be carried out?	4
Aggregation via LOD datasets based on SPARQL (AR04)	• Are your Linked Data datasets available in EDM (or Schema.org) or should a mapping be carried out? • Could you provide a SPARQL query that lists the URIs of all (or a subset of) cultural objects' RDF resources in the dataset?	2

[32] All the possible questions are listed in the table. Naturally, each email was customised to reflect the interests, current implementations and aggregation route's requirements of each contacted entity.

Aggregation routes	Questions[32]	Occurrences
Aggregation of IIIF based on IIIF Collection (AR05)	• Could you give me a URL pointing to a IIIF (top-level) collection? • Is there a `rdfs:seeAlso` pointing to structured metadata within your IIIF Manifests?	5
Aggregation of IIIF based on Sitemaps (AR06)	• Does your Sitemaps contain any information regarding the IIIF Manifests?	3

5.3.3.4 Resolution on the immediate conduct of pilots

From the emails sent with templates 4 and 5, half of the recipients responded (seven out of fourteen). Based on the answers received, three types of resolutions were determined:

- **No pilot**: no aggregation pilots could be conducted because the feedback received indicated either that too much work needed to be done or that the relevant organisation did not have the time or resources to do so.
- **Hold**: the necessary information was obtained, but the time that was required to carry out the pilot went beyond the remaining timeframe of the master's thesis.
- **Try**: all the necessary information was obtained to conduct an aggregation pilot.

A single aggregation pilot with MuseuMap, a Hungarian aggregator in the museum field, was selected for realisation (cf. Table 14)[33]. With regard to the other six respondents, the main reason for not carrying out a pilot aggregation were the concerns surrounding metadata mapping (three occurrences), the lack of structured metadata within web pages or the absence of it in their IIIF Manifests via the `rdfs:seeAlso` property (two occurrences), the lack of budget (one occurrence), or that OAI-PMH was mentioned as a component in a LOD aggregation, which was out of scope for this study (one occurrence).

Table 14: Resolution on the conduct of aggregation pilots

Total	Outcome		
	Try	Hold	No pilot
7	1	4	2

The Europeana R&D team intends to keep in touch with all these organisations, however, in particular the Swedish Open Cultural Heritage (SOCH) and Wikimedia Sverige, which both expressed great interest in conducting pilots. The former has LOD metadata in DCAT-AP[34], an application profile of DCAT for data portals in Europe, which would require some mapping and figuring out how the Linked Data could be harvested as a complete dataset. These tasks could, for instance, be done with the LOD-aggregator and the NDE indicated their willingness to potentially integrate it as additional datasets to the ECC LOD pipeline. For the latter,

[33] While so far, all data has been anonymised, it was decided to disclose the names of the organisations most interested in the aggregation pilots, as details of their configuration, interests and needs have to be given regardless.

[34] https://riksantikvarieambetet.github.io/soch-dcat-ap/soch.rdf

Wikimedia Sverige, it would be a matter of using Wikidata as a “vocabulary data exchange” as well as developing a solution to upload individual datasets in Europeana to Wikimedia Commons[35].

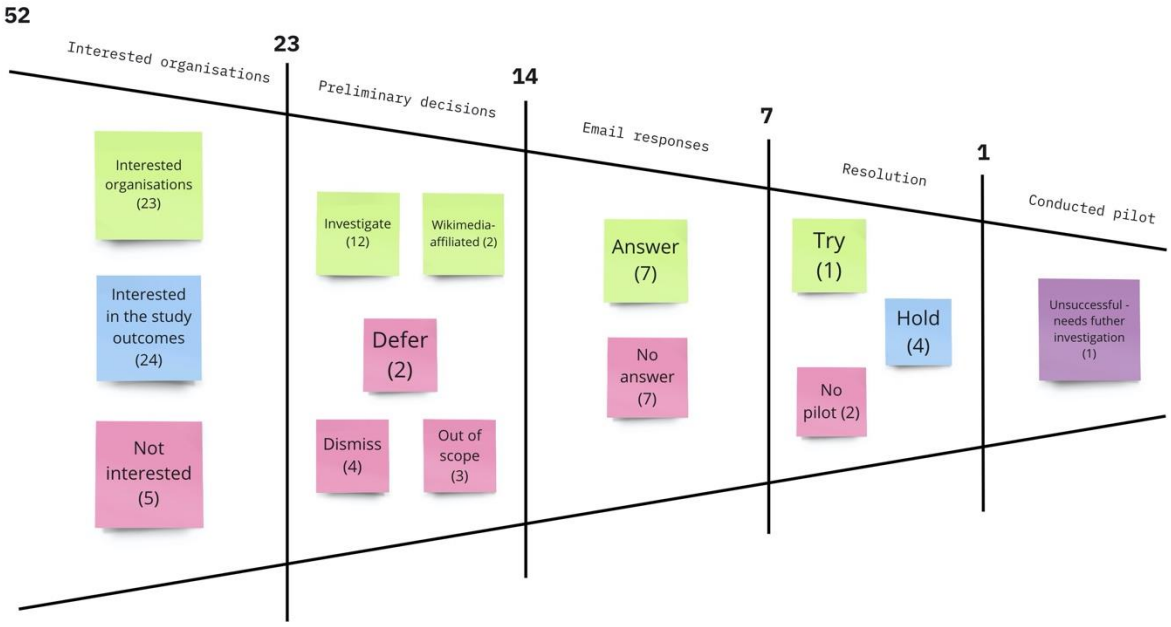
5.3.3.5 Attempts to carry out the MuseuMap pilot

The chosen aggregation route for the pilot with MuseuMap data was through Sitemaps and Schema.org in HTML pages (AR01). While the crawling of the sitemaps worked well with the DAL, it was unfortunately not possible to extract the Schema.org metadata from the web pages[36]. At the time of writing, further investigation is still underway to determine the origin of the error.

5.3.4 Conclusion

As an attempt to synthesize this phase of the research, a funnel-shaped visual representation illustrates the different steps that led from 52 survey participants to the (partial) execution of one aggregation pilot in the allotted time (cf. Figure 23).

Figure 23: Summarised representation of the assessment of aggregation pilots



In the end, the attempts to define and carry out pilots adding to the body of experiments already gathered by the Europeana R&D team have proved difficult to execute. However, as this final result derives from a principled assessment of the leads provided by quite a representative survey, one can regard this outcome as a useful lesson with regard to the interest and challenges related to state-of-the-art technology adoption as well as to the contribution of data via novel channels.

[35] This is not however related to any of the selected aggregation routes but is still mentioned here due to the interesting enrichment potential for both Wikimedia and Europeana.

[36] The dataset details can be accessible at the following URL: https://rnd-2.eanadev.org/data-aggregation-lab/harvester/dataset-detail?dataset=a7393aa5-625b-4ff8-874b-28c03c584fe2

6. Recommendations

This chapter aims to make recommendations on the choice and deployment of (meta)data aggregation mechanisms. It is divided into four sections, which employ visual representations from service design methods.

The first one considers the target levels to which solutions can be applied, the second focuses on the development of a representation based on the Opportunity Solution Tree template, the third associates the solutions to the three priorities defined in the Europeana Strategy 2020-2025, and, finally, the fourth discusses the process for implementing propositions that can enable better aggregation and discovery of CH resources. The recommendations consist of five solutions and a number of proposed detailed suggestions.

6.1 Target levels

We begin by identifying target levels as a non-prescriptive effort to prioritise certain aggregation routes.

Three target levels are identified: *the digital object level, the metadata level, and the providing institution level* (CHI and intermediary aggregator). We determined the latter under the influence of the Europeana Strategy 2020-2025 as well as the EPF metadata and content tiers.

6.2 Opportunity Solution Tree

A significant number of aggregation routes were outlined throughout this dissertation and especially in 5.3.2 where a list of thirteen different combinations of mechanisms (some of which are technically specializations of others, such as LOD datasets and IIIF aggregation routes numbering three and four respectively).

To better understand where Europeana and its partners should focus their energy in terms of operational approaches to aggregation, we chose to leverage the Opportunity Solution Tree. This method consists in separating components mapping out connections in this way: 1) desired outcome → 2) opportunities → 3) solutions → 4) experiments.

It is also worth pointing out that this design service technique, which visually resembles a mind map, is a way of mapping items and that it is a process of ideation rather than a rigid layout, since there are elements that could very well be placed under more than one branch. Moreover, within the Opportunity Tree Solution presented here[37], solutions and experiments can be approaches as well as technical means to achieve a given approach (such as the aggregation routes).

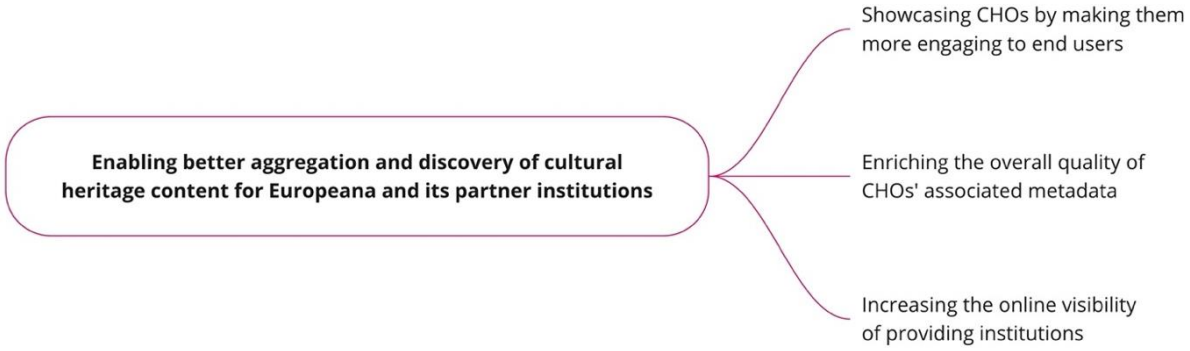
The desired outcome (*Enabling better aggregation and discovery of CH content for Europeana and its partner institutions*), which bears the eponymous title to this dissertation, divides into the following three sub-branches (opportunities), following the aforementioned target levels:

- **Digital object**: showcasing CHOs by making them more engaging to end users
- **Associated metadata**: enriching the overall quality of CHOs' associated metadata
- **Providing institution**: Increasing the online visibility of providing institutions

[37] The full visual representation can be found in the Appendices (cf. Figure 32). Within this section, the different steps are broken down into several illustrations.

The connections between the desired outcome and the opportunities, making up the first step of the Opportunity Solution Tree, are shown in Figure 24.

Figure 24: Desired outcome and opportunities with respect to the target levels



As illustrated in Figure 25, we selected the APIs that the IIIF community develops and maintains as the most appropriate solution to the first defined opportunity. The functionalities offered by IIIF-compliant resources, such as deep zoom or annotation capacities, provide a much more engaging user experience, whether on the providing institutions' platform or on Europeana's one. The experiments consist of the various aggregation routes related to IIIF[38], the IIIF and Europeana Working Group as well as the involvement of Europeana in the relevant IIIF committees and groups.

Figure 25: Proposed solution and experiments for the digital object level

We chose a solution based on LOD for the second opportunity concerning the overall metadata quality enhancement. The latter is broken down into two aggregation routes, namely those based on LOD datasets, which have already been tested within the ECC project, and an approach based on LDN and LDP, which is worth exploring over the next few years as these two specifications offer an interesting synchronisation versatility (cf. Figure 26).

Figure 26: Proposed solution and experiments for the metadata level

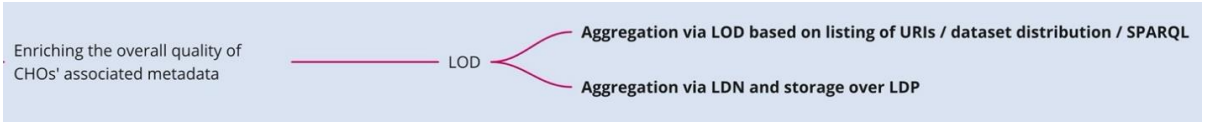


Figure 27 identifies the third opportunity focusing on the online visibility of providing institutions. Three solutions and their related experiments are proposed.

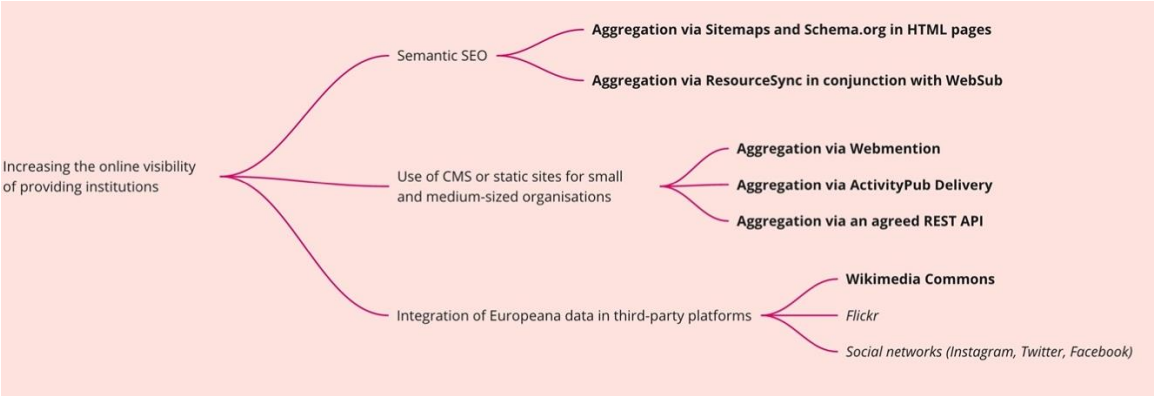
The first highlighted solution would be to improve the semantic SEO, notably though the aggregation mechanisms of Sitemaps and Schema.org as well as ResourceSync used in conjunction with WebSub.

[38] All experiments items related to an aggregation pilot are highlighted in **bold**.

A second solution would be for small and medium-sized institutions with little technical and financial means to leverage content management systems (CMS) or easily deployable static sites built on frameworks or plugins endorsed by Europeana and the aggregators. This would notably allow the use of Webmention, AP or a custom API enabling a relatively easy aggregation process between the systems.

The third solution involves the integration of Europeana's data on third-party platforms such as social networks, Flickr and Wikimedia. For the latter, initiatives such as the upload of Europeana data on Wikimedia Commons (cf. 5.3.3) could be perceived as a post-aggregation process that could improve both the metadata quality (target tier 2) and the global exposure of Europeana.

Figure 27: Proposed solutions and experiments for the providing institution level



6.3 Alignment with the Europeana Strategy

To determine the expected impact of deploying each of the aforementioned solutions, we tried to align them against the following three priorities outlined in the Europeana Strategy 2020-2025 (Europeana 2020):

- **Strengthen the infrastructure**: *[investment] in supporting innovation activities that keep the infrastructure aligned with state-of-the-art tech.*
- **Improve data quality**: *[investment of] resources in activities related to metadata and content improvement in collaboration with aggregators and data providers.*
- **Build capacity**: [*support] institutions in their digital transformation [by showcasing] the importance and added value of digitisation, adoption of standards, best practice and common solutions in making quality content that is useful for a global online audience and that fosters innovation*.

The last two branches of the Opportunity Solution Tree correspond to the priorities of the Europeana Strategy as the identified solutions and experiments enable us to respond to the issues regarding a more efficient way for data providers to share their collections as well as improving the reusability of their digital content. Furthermore, the range of solutions (*IIIF, LOD, Semantic SEO, CMS or static sites, third-party platforms*) addresses the different needs of the Europeana Network, which is made up of a diverse group of organisations operating with distinct mind-sets and within distinct technology environments.

As such, it can be seen in Table 15 below that most solutions fall under two of the Europeana Strategy's priorities (all boxes with light green shading and a tick).

Table 15: Alignment between identified solutions and Europeana Strategy priorities

	Strengthen the infrastructure	Improve data quality	Build capacity
IIIF	✓		✓
LOD	✓	✓	✓
Semantic SEO	✓	✓	
CMS / static sites	✓		
Third-party platforms		✓	✓

6.4 Suggestions for implementing the identified solutions

Table 16 contains detailed suggestions for implementing the five identified solutions to encourage the adoption of new aggregation mechanisms that, in the long term, have the potential to reduce non-automated labour, improve access to Europe's digital CH, and establish a series of measures to encourage partnerships. They are positioned against the five solutions of the Opportunity Solution Tree as well as align with the three priorities of the Europeana Strategy 2020-2025. The motivations behind these suggestions, with reference to insights obtained throughout the master's thesis, is also included. Finally, these suggestions are directed to Europeana [E], aggregators [A] as well as data providers [D].

Table 16: Proposed suggestions

Solution	Europeana Strategy	Suggestions	Motivations
IIIF	Strengthen the infrastructure	• Produce guidelines on best practices regarding how and which structured metadata should be linked within IIIF Manifests. [E] • Assist in the generation of IIIF Manifests and IIIF Collections for harvesting purposes as well as providing hosting capabilities if necessary. [AE] • Continue the integration and compliance of new versions of the IIIF APIs by upgrading the infrastructure and related R&D tools (DAL, Metadata Testing Tool). [E]	The online survey and the assessment of potential aggregation pilots showed that there is still a lot to be accomplished to facilitate IIIF-based aggregation. Among other things, this includes access to structured metadata, and the creation of IIIF Collections. More broadly, there is also a need to build knowledge and skills on IIIF.
	Build capacity	• Set up short workshops during the Europeana and IIIF Working Group videoconferences. [ADE] • Organise a "Train the trainer" course on IIIF at an upcoming EAF. [AE] • Facilitate the onboarding of Europeana stakeholders into the relevant IIIF groups and committees. [E]	

Solution	Europeana Strategy	Suggestions	Motivations
LOD	Strengthen the infrastructure	• Continuously improve the LOD-aggregator pipeline with new datasets and by wrapping the system within a user interface. [AE] • Investigate the capabilities offered by LDN in the synchronisation and update of CHOs. [E]	The discussions on sustainability that occurred during the ECC project indicated that improvements to the LOD-aggregator could be made. There was still interest from a significant part of the survey respondents on Social Web Protocols (including LDN) but almost no implementation. During the assessment of potential aggregation pilots, significant work was required on the quality of Schema.org metadata.
	Improve data quality	• Improve metadata quality by ingesting more EDM or Schema.org properties with relevant semantics. [ADE]	
	Build capacity	• Create cookbook recipes on the use of LOD in the context of aggregation. [E] • Monitor the development of the Schema.org vocabulary. [E] • Advocate for the further uptake of LOD in the CH domain, considering the specific requirements of each sub-domain. [A]	
Semantic SEO	Strengthen the infrastructure	• Integrate error reporting within DAL for the cases where the crawler is not able to extract structured metadata. [E] • Provide a tool for data providers and aggregators that is able to quickly build Sitemaps with relevant pages (for CHOs) and the integration of RS namespaces. [E]	The attempted aggregation pilot with MuseuMap showed the need of integrating new reporting features within DAL. During the assessment of potential pilots, it was noticed that some organisations didn't have up-to-date Sitemaps.
	Improve data quality	• Run web referencing tests both before and after deployment of Schema.org in the CHO's landing pages. [D]	
CMS / static site	Strengthen the infrastructure	• Set up Webmention and AP endpoints. [E] • Develop easy-to-deploy frameworks for data providers. [E]	At the EAF Spring 2020, solutions for small institutions deploying CMS to display their collections were discussed.
Third-party platform	Improve data quality	• Conduct regular assessment of the most appropriate vocabularies and authority files to leverage for indexing purposes. [A]	This comes from a response received from Wikimedia Sverige on the conduct of a potential enrichment pilot.
	Build capacity	• Carry out a pilot to integrate a maximum of Europeana's data into Wikimedia Commons. [ADE]	

Ultimately, two transversal ideas not captured in Table 16 are:

- the integration of technology support within the Metis Sandbox [E] – which can be applied to all propositions related to "Strengthen the infrastructure";
- the translation of the documentation into the various languages used within the ENA [ADE] – which can be applied to all propositions related to "Build capacity".

7. Conclusion

This chapter is divided into two sections that discuss the work that was accomplished as well as possible next actions to be undertaken.

7.1 Retrospective

This section draws on the issues outlined throughout this dissertation and provides a review of the study achievements and outcomes as well as addressing the two main research questions defined at the outset of the master's thesis.

7.1.1 Study achievements and outcomes

First, the literature review extended what had previously been explored, notably by Europeana R&D team members, and, for the first time, an extensive overview of various technologies was done in textual and tabular form. However, there are still many unresolved issues concerning the deployment requirements of certain mechanisms for harvesting purposes, namely the Social Web Protocols which are still little known or implemented in the CH domain or OPDS2 which is still a draft specification.

Then, participating in the ECC project with regard to the LOD Functional Application task, which was not planned at the beginning of the master's thesis, was an opportunity to try out the pipeline developed by the NDE, to discuss with the people who built the system and to understand the required process of a LOD dataset aggregation while improving the documentation on the dedicated GitHub repository.

Thirdly, the survey results indicated that interest in IIIF and LOD strongly supports the integration of these technologies within the Europeana Aggregation Strategy and, as a result, their potential deployment in the Metis Sandbox. Among the most positive aspects of the online survey was the involvement of more than 50 respondents from a wide range of institutions, of which almost half were interested in carrying out an aggregation pilot in collaboration with Europeana. It is also worth mentioning that the survey did not attempt to provide a comprehensive panorama of the different kinds of metadata used by CHIs and aggregators. Nevertheless, it was still important to have a sense of the metadata standards used by organisations as a basis on which mapping could be undertaken.

Fourth, the work concerning aggregation pilots ought to be more properly considered as an analysis phase, serving to highlight future opportunities for cooperation rather than as a proper test cycle. This assessment was nonetheless useful in identifying future pilots, building bridges with a number of organisations that had not previously participated in such pilots ,and identifying the types of problems that can be encountered when examining the use of alternative aggregation mechanisms. For instance, the only pilot that was conducted seemed straightforward at first glance but the inability for the DAL crawler to retrieve structured metadata from web pages demands further investigation.

As for the recommendations, five main solutions and detailed proposed suggestions were listed and directed towards one or several of the key roles of the aggregation workflow (data provider, aggregator, Europeana). These propositions are intended to be pragmatic and have a focus on operational matters. Other approaches such as crowdsourcing or machine learning for the metadata enrichment could respond as well to the three types of opportunities identified, but they were not investigated so as not to fall outside the scope of this study. Moreover, it was

decided not to design a Service Blueprint emphasising the key differences between the current operating model and a model taking into account the integration of new ingestion features, as this work was already thought out and conceived within the Europeana Aggregation Strategy.

Overall, the integration of the author within the Europeana R&D team and the collaboration carried out within the ECC project made it possible to rapidly acquire the appropriate information as well as to explore various approaches regarding (meta)data aggregation mechanisms in the CH domain. Such interactions significantly influenced the overall direction of the master's thesis without undermining the scope nor the agreed objectives of the research. On the contrary, it provided an opportunity to gain an insight into the opinions of a variety of stakeholders, to establish contacts as well as to learn more about the activities being carried out within the Europeana Network. Although only one aggregation pilot was conducted, other concrete deliverables were produced throughout the course of this master's thesis, such as the lightning talk at the online EAF Spring 2020 event, the inclusion of a survey findings summary as part of a DSI milestone report, as well as the submission of a conference paper for MTSR2020.

7.1.2 Alternative mechanisms to OAI-PMH

The transition from OAI-PMH to alternative aggregation mechanisms is not only feasible but also is desirable within the CH domain. This shift in the operating model for Europeana's stakeholders is scalable and sustainable, even if the deployment of new technologies and standards requires extensive learning, human mediation and technical resources upstream.

The literature review and practical experiments proved that there are many possible combinations that allow for (meta)data aggregation in the CH domain and which could, in the long run, supersede OAI-PMH as the preferred protocol.

All the mechanisms where deployment requirements could be determined offer the same or higher capabilities as OAI-PMH. For instance, IIIF makes it easier for the providing institution to maintain control over the representation of their CHOs thanks to the interoperability of shared APIs. Likewise, Schema.org allows the organisations the implement it within their web pages to improve their web referencing, and other protocols such as RS, WebSub, LDN or AS2 offer the promise of easier metadata synchronisation.

It should be noted as well that technologies that have not yet been fully tested, such as those stemming from the Social Web Protocols, should not be overlooked, as they could offer significant advantages, especially for harvesting data from smaller institutions, which may require some sort of third-party hosting for the purpose of establishing compliant endpoints - an aspect where Europeana and aggregators could provide assistance.

Conversely, there is not necessarily one technology that is better than another to replace OAI-PMH over time, but rather different technologies based on the Web architecture that offer a variety of advantages. While this could indeed slow down the transition to other technologies, it should be recalled that many institutions use OAI-PMH only in the context of aggregation towards Europeana or other aggregators. In other words, it may well be more expensive to maintain such infrastructures over the long term than to undergo a changeover which could bring additional benefits.

7.1.3 Conditions to deploy alternative mechanisms for aggregation

The conditions for deploying alternative mechanisms are known to varying degrees, and when they are, it still requires a lot of human and technical resources on the part of all Europeana stakeholders to comply with the chosen technologies in the context of (meta)data aggregation. It must be said as well that the prerequisites on the data provider and aggregator side are not yet clear for all mechanisms, and that some technologies (IIIF, LOD datasets, Sitemaps and Schema.org) fare better in terms of their exploitability.

Support in deploying such mechanisms can be achieved through different means. Firstly, input could be provided by institutions that are pioneers in LOD publication and have a foothold in the IIIF community. Secondly, assistance can obviously come from Europeana's aggregators who are knowledgeable about the specific requirements of each domain or region as well as dealing on a continuous basis with data providers. Last but not least, Europeana should consider the knowledge and skills acquired by their R&D department on novel approaches to aggregation of CHOs and their associated metadata, notably through the pilots they carried out over the past few years and through the outcomes of this master's thesis conducted with their guidance. All of these constitute key factors that Europeana need to consider for the renewal of the ingestion service and the onboarding of CHIs on the use of alternative mechanisms.

The recommendations presented in this dissertation are meant to assist in the adoption of these technologies. To avoid too much investment by smaller institutions, it is essential that Europeana and the aggregators ensure that they carry out the necessary mediation and ad-hoc advocacy efforts to facilitate the transition from OAI-PMH to other technologies by providing resources such as guidelines in different languages and easy-to-use systems. An important concern that would need to be taken into account is to propose a scalable protocol that works for multiple large datasets, but also something that smaller CHIs with limited budgets and limited technical expertise can use.

One of the most promising outcomes from this work it that a significant number of the technologies discussed in this master's thesis are on the radar for incorporation into the Metis Sandbox. Once the system is usable and alternative aggregation routes are available to users, it will be easier to promote such mechanisms as well as to explain the additional benefits of each technology.

7.2 Future work and discussion

The new Europeana operating model will not instantaneously change the way data is harvested as OAI-PMH remains prevalent in the CH domain. However, the new functionalities within the Metis Sandbox will enable a little breakthrough as harvesting experiments will not only have to be carried out on a project basis but can also be explored by data providers directly.

Nevertheless, these providers will still have to be convinced to use such functionalities. One of the initial opportunities would be to rely on those organisations who expressed their interest in an aggregation pilot which could not be carried out during the course of this master's thesis. Should the Metis Sandbox not be ready, or if some functions are not integrated straight away, in the foreseeable future, the pilots could still be done first with the available resources developed by the R&D team, such as the DAL.

Another action (or step to be carried out in parallel with the conduct of aggregation pilots) would be to implement the dissertation's recommendations by prioritising solutions according to their identified target level.

A third and final approach would be to verify not only the quality of the data provided to Europeana but also the compliance with the alternative aggregation mechanisms, in line with what the EPF does for content and metadata (cf. 3.2.3), but, in this case, for the purposes of technically complying in the deployment of protocols and endpoints, or by making structured metadata available that can be easily crawled.

In the end, the shift in the operating model will constitute an advisory effort on the part of Europeana and the aggregators to facilitate the adoption of these alternative aggregation mechanisms within the network, since it will not only be a matter of thinking about the next standards, but also of how to address digital transformation with data providers.

While this research focused on the operational aspects of aggregation, the aim of providing better discoverability of content was always in the line of sight, for instance by addressing standards that could improve metadata enrichment or offer further potential. In this respect, it is appropriate to mention the Cultural Japan platform, which was opened to the public on 1 August 2020 and provides more than a million resources on Japanese culture. It has a user interface, a SPARQL endpoint, and more than half of the aggregated digital surrogates are IIIF Manifests. It is indeed really interesting to see how IIIF is leveraged apart from a generic search facet: the “Self-Museum” application allows end users to create their own three-dimensional gallery featuring IIIF-compliant resources that they can browse around while discovering the different objects.

Thus, alternative aggregation mechanisms of CHOs enable platforms that ingest these resources to showcase them in a different perspective, in some cases by means that are not or, as yet, rarely explored in the CH domain.

Bibliography

ALBERTONI, Riccardo, BROWNING, David, COX, Simon, GONZALEZ-BELTRAN, Alejandra, PEREGO, Andrea and WINSTANLEY, Peter, 2020. Data Catalog Vocabulary (DCAT) - Version 2. *W3C* [online]. 4 February 2020. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/vocab-dcat-2/

ALEXANDER, Martha Latika and GAUTAM, J N, 2004. Interoperability and Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). In: *2nd International CALIBER* [online]. New Delhi, India. February 2004. p. 8. [Accessed 1 August 2020]. Available from: https://pdfs.semanticscholar.org/ba2c/0ac2fea62801a866d8663a9488702ae54afb.pdf

AMERICAN NATIONAL STANDARDS INSTITUTE and NISO, 2017. ResourceSync Framework Specification (ANSI/NISO Z39.99-2017). *Open Archives Initiative* [online]. 2 February 2017. [Accessed 1 August 2020]. Available from: http://www.openarchives.org/rs/1.1/resourcesync

APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2016. IIIF Content Search API 1.0. *International Image Interoperability Framework* [online]. 12 May 2016. [Accessed 1 August 2020]. Available from: https://iiif.io/api/search/1.0/

APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2017. IIIF Authentication API 1.0. *International Image Interoperability Framework* [online]. 19 January 2017. [Accessed 1 August 2020]. Available from: https://iiif.io/api/auth/1.0/

APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2019. IIIF Content State API 0.2. *International Image Interoperability Framework* [online]. 4 February 2019. [Accessed 1 August 2020]. Available from: https://iiif.io/api/content-state/0.2/

APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2020a. IIIF Change Discovery API 0.9. *International Image Interoperability Framework* [online]. 4 June 2020. [Accessed 1 August 2020]. Available from: https://iiif.io/api/discovery/0.9/

APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2020b. IIIF Image API 3.0. *International Image Interoperability Framework* [online]. 3 June 2020. [Accessed 1 August 2020]. Available from: https://iiif.io/api/image/3.0/

APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2020c. IIIF Presentation API 3.0. *International Image Interoperability Framework* [online]. 3 June 2020. [Accessed 1 August 2020]. Available from: https://iiif.io/api/presentation/3.0/

BACA, Murtha, 2016a. Glossary. In: *Introduction to Metadata* [online]. Third edition. Los Angeles, CA, USA: Getty Research Institute. [Accessed 1 August 2020]. ISBN 978-1-60606-479-5. Available from: https://www.getty.edu/publications/intrometadata/glossary/

BACA, Murtha, 2016b. Practical Principles for Metadata Creation and Maintenance. In: *Introduction to Metadata* [online]. Third edition. Los Angeles, CA, USA: Getty Research Institute. [Accessed 1 August 2020]. ISBN 978-1-60606-479-5. Available from: https://www.getty.edu/publications/intrometadata/practical-principles/

BACA, Murtha (ed.), 2016c. *Introduction to metadata*. Third edition. Los Angeles, CA: Getty Research Institute. ISBN 978-1-60606-479-5.

BECKER, Jerry, 2020. Opportunity Solution Tree. *Open Practice Library* [online]. 16 April 2020. [Accessed 1 August 2020]. Available from: https://openpracticelibrary.com/practice/opportunity-solution-tree/

BERMÈS, Emmanuelle, ISAAC, Antoine and POUPEAU, Gautier, 2013. *Le Web sémantique en bibliothèque*. Paris: Electre/Cercle de la librairie. ISBN 978-2-7654-1417-9.

BERMÈS, Emmanuelle, 2011. Convergence and interoperability: a Linked Data perspective. In: *World Library and Information Congress: 77th IFLA General Conference and Assembly* [online]. San Juan, Puerto Rico. 19 December 2011. p. 1–12. [Accessed 1 August 2020]. Available from: https://www.ifla.org/past-wlic/2011/149-bermes-en.pdf

BERMÈS, Emmanuelle, 2020. *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)* [online]. PhD Thesis. Paris, Ecole nationale des chartes. [Accessed 1 August 2020]. Available from: https://tel.archives-ouvertes.fr/tel-02475991

BERNERS-LEE, Tim and FISCHETTI, Mark, 2001. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. DIANE Publishing Company. ISBN 978-0-7567-5231-6.

BERNERS-LEE, Tim, HENDLER, James and LASSILA, Ora, 2001. The Semantic Web. *Scientific American*. 2001. Vol. 284, no. 5, p. 34–43. JSTOR

BERNERS-LEE, Tim, 2006. Linked Data. *W3C* [online]. 27 July 2006. [Accessed 29 June 2020]. Available from: https://www.w3.org/DesignIssues/LinkedData.html

BERNERS-LEE, Tim, 2009. *The next web* [online]. February 2009. [Accessed 1 August 2020]. TED2009. Available from: https://www.ted.com/talks/tim_berners_lee_on_the_next_web

CAPADISLI, Sarven and GUY, Amy, 2017. Linked Data Notifications. *W3C* [online]. 2 May 2017. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/ldn/

CAPADISLI, Sarven, 2019. *Linked Research on the Decentralised Web* [online]. PhD Thesis. University of Bonn. [Accessed 1 August 2020]. Available from: https://csarven.ca/linked-research-decentralised-web

CHARLES, Valentine, ISAAC, Antoine, CLAYPHAN, Robina and MEGHINI, Carlo, 2017. Definition of the Europeana Data Model v5.2.8. *Europeana Pro* [online]. 6 October 2017. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf

CHARLES, Valentine and ISAAC, Antoine, 2015. *Enhancing the Europeana Data Model (EDM)* [online]. White paper. Available from: http://pro.europeana.eu/files/Europeana_Professional/Publications/EDM_WhitePaper_17062015.pdf

CLAYPHAN, Robina, DE HOOG, Kirsten and CHARLES, Valentine, 2017. Europeana Data Model – Mapping Guidelines v2.4. *Europeana Pro* [online]. 6 October 2017. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf

COBURN, Erin, LANZI, Elisa, O'KEEFE, Elizabeth, STEIN, Regine and WHITESIDE, Ann, 2010. The Cataloging Cultural Objects experience: Codifying practice for the cultural heritage community: *IFLA Journal*. 26 April 2010. DOI 10.1177/0340035209359561.

CORREA, Hector, 2015. Introduction to Linked Data Platform (LDP). *Hydra Connect 2015* [online]. 23 September 2015. [Accessed 1 August 2020]. Available from: https://www.slideshare.net/hectorwashere/introduction-to-linked-data-platform-ldp

COSSU, Stefano, 2020. IIIF at the Getty: Vision & Tactics. *CNI Spring 2020 Virtual Membership Meeting* [online]. 27 April 2020. [Accessed 1 August 2020]. Available from: https://www.cni.org/wp-content/uploads/2020/03/scossu_iiif_getty_vision_and_tactics_cni_spring_2020.pdf

D'ALTERIO, Emily, 2018. Who are we, Europeana, in this digital transformation?. *Europeana Pro* [online]. 30 July 2018. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/post/who-are-we-europeana-in-this-digital-transformation

DALEY, Beth, SCHOLZ, Henning and CHARLES, Valentine, 2019. Developing a metadata standard for digital culture: the story of the Europeana Publishing Framework. *Europeana Pro* [online]. 28 October 2019. [Accessed 4 November 2019]. Available from: https://pro.europeana.eu/post/developing-a-metadata-standard-for-digital-culture-the-story-of-the-europeana-publishing-framework

DELMAS-GLASS, Emmanuelle and SANDERSON, Robert, 2020. Fostering a community of PHAROS scholars through the adoption of open standards. *Art Libraries Journal*. January 2020. Vol. 45, no. 1, p. 19–23. DOI 10.1017/alj.2019.32.

DOERR, Martin, GRADMANN, Stefan, HENNICKE, Steffen, ISAAC, Antoine and MEGHINI, Carlo, 2010. The Europeana Data Model (EDM). In: *World Library and Information Congress: 76th IFLA general conference and assembly* [online]. Gothenburg, Sweden. August 2010. [Accessed 1 August 2020]. Available from: https://www.ifla.org/past-wlic/2010/149-doerr-en.pdf

ELINGS, Mary W. and WAIBEL, Günter, 2007. Metadata for all: Descriptive standards and metadata sharing across libraries, archives and museums. *First Monday*. 5 March 2007. DOI 10.5210/fm.v12i3.1628.

EUROPEANA, 2015. Glossary of Terms. *Europeana Pro* [online]. 15 January 2015. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/resources/standardization-tools/glossary

EUROPEANA, 2017. Strategy 2020 update. *Europeana* [online]. 28 February 2017. [Accessed 1 August 2020]. Available from: https://strategy2020.europeana.eu/update/

EUROPEANA, 2018. Europeana DSI-4. *Europeana Pro* [online]. 31 August 2018. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/project/europeana-dsi-4

EUROPEANA, 2019a. *Europeana DSI-4 Annual Report* [online]. Europeana Digital Service Infrastructure. The Hague, Netherlands: Europeana Foundation. [Accessed 1 August 2020]. Available from: https://doi.org/10.2759/07652

EUROPEANA, 2019b. Europeana Common Culture. *Europeana Pro* [online]. 14 March 2019. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/project/europeana-common-culture

EUROPEANA, 2020. Europeana Strategy 2020-2025: Empowering digital change. *Europeana Pro* [online]. May 2020. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/files/Europeana_Professional/Publications/EU2020StrategyDigital_May2020.pdf

FALLON, Julia, 2015. Available rights statements. *Europeana Pro* [online]. 6 February 2015. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/page/available-rights-statements

fat ping, 2015. *IndieWeb* [online]. [Accessed 1 August 2020]. Available from: https://indieweb.org/fat_ping

FREIRE, Nuno, CHARLES, Valentine and ISAAC, Antoine, 2018. Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata. In: *The Semantic Web*. Springer International Publishing. 2018. p. 225–239. Lecture Notes in Computer Science. ISBN 978-3-319-93417-4.

FREIRE, Nuno and CHARLES, Valentine, 2017. New approaches for data acquisition at Europeana: IIIF, Sitemaps and Schema.org. *Linked Data in Research and Cultural Heritage* [online]. The Hague, Netherlands. 1 May 2017. [Accessed 1 August 2020]. Available from: https://www.slideshare.net/NunoFreire2/new-approaches-for-data-acquisition-at-europeana-iiif-sitemaps-and-schemaorg-dans-seminar-2017

FREIRE, Nuno, ISAAC, Antoine and RAEMY, Julien Antoine, 2020. MS5 IIIF harvesting implemented, M22: *Metadata and content aggregation via linked data and IIIF: ingested datasets, data quality evaluation and other experiments*. Europeana DSI-4. The Hague, Netherlands: Europeana Foundation.

FREIRE, Nuno, ISAAC, Antoine, ROBSON, Glen, BROOKS, John and MANGUINHAS, Hugo, 2017. A survey of Web technology for metadata aggregation in cultural heritage. *Information Services & Use*. October 2017. Vol. 37, no. 4, p. 425–436. DOI 10.3233/ISU-170859.

FREIRE, Nuno, MEIJERS, Enno, DE VALK, Sjors, RAEMY, Julien Antoine and ISAAC, Antoine, 2020. Metadata Aggregation via Linked Data in Europeana: results of the Common Culture project. In: *Metadata and Semantic Research*. 2020. Unpublished.

FREIRE, Nuno, MEIJERS, Enno, VOORBURG, René and ISAAC, Antoine, 2018. Aggregation of cultural heritage datasets through the Web of Data. *Procedia Computer Science*. 1 January 2018. Vol. 137, p. 120–126. DOI 10.1016/j.procs.2018.09.012.

FREIRE, Nuno, ROBSON, Glen, HOWARD, John B., MANGUINHAS, Hugo and ISAAC, Antoine, 2018. Cultural heritage metadata aggregation using web technologies: IIIF, Sitemaps and Schema.org. *International Journal on Digital Libraries*. 26 October 2018. DOI 10.1007/s00799-018-0259-5.

FREIRE, Nuno, VERBRUGGEN, Erwin, MEIJERS, Enno, DE VALK, Sjors, GEORGIADIS, Haris, RÖNKÄ, Minna, PATRICIO, Helena and ISAAC, Antoine, 2019. Aggregation of Schema.org Linked Data for the Europeana Common Culture project. *Europeana Conference 2019* [online]. Lisbon, Portugal. 27 November 2019. [Accessed 25 June 2020]. Available from: https://www.slideshare.net/NunoFreire2/aggregation-of-schemaorg-linked-data-for-the-europeana-common-culture-project-236206695

FREIRE, Nuno, VOORBURG, René, CORNELISSEN, Roland, DE VALK, Sjors, MEIJERS, Enno and ISAAC, Antoine, 2019. Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network. *Information*. 30 July 2019. Vol. 10, no. 8, p. 252. DOI 10.3390/info10080252.

FREIRE, Nuno, 2020a. V0.1: *Guidelines for providing and handling Schema.org metadata in compliance with Europeana* [online]. The Hague, Netherlands: Europeana Foundation. [Accessed 1 August 2020]. Available from: https://zenodo.org/record/3817236

FREIRE, Nuno, 2020b. V0.2: *Specifying a linked data dataset for Europeana and aggregators* [online]. The Hague, Netherlands: Europeana Foundation. [Accessed 1 August 2020]. Available from: https://zenodo.org/record/3817314

GARDEUR, Hadrien, 2020. OPDS Catalog 2.0. *Open Publication Distribution System* [online]. March 2020. [Accessed 21 June 2020]. Available from: https://drafts.opds.io/opds-2.0.html

GAUDINAT, Arnaud, BEAUSIRE, Jonas, FUSS, Megan, BANFI, Elisa, GOBEILL, Julien and RUCH, Patrick, 2017. Global picture of OAI-PMH repositories through the analysis of 6 key open archive meta-catalogs. *arXiv:1708.08669 [cs]* [online]. 29 August 2017. [Accessed 1 August 2020]. Available from: http://arxiv.org/abs/1708.08669

GENESTOUX, Julien and PARECKI, Aaron, 2018. WebSub. *W3C* [online]. 23 January 2018. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/websub/

GREENBERG, Jane, 2005. Understanding Metadata and Metadata Schemes. *Cataloging & Classification Quarterly*. 9 September 2005. Vol. 40, no. 3–4, p. 17–36. DOI 10.1300/J104v40n03_02.

GUY, Amy, 2017. Social Web Protocols. *W3C* [online]. 25 December 2017. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/social-web-protocols/

HADRO, Josh, 2019. *Introduction to IIIF* [online]. Göttingen, Germany, 4 November 2019. [Accessed 1 August 2020]. 2019 IIIF Showcase. Available from: https://youtu.be/l8kc8nH5f8I

HASLHOFER, Bernhard, WARNER, Simeon, LAGOZE, Carl, KLEIN, Martin, SANDERSON, Robert, NELSON, Michael L. and VAN DE SOMPEL, Herbert, 2013. ResourceSync: leveraging sitemaps for resource synchronization. In: *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion* [online]. Rio de Janeiro, Brazil: ACM Press. 2013. p. 11–14. [Accessed 1 August 2020]. Available from: http://dl.acm.org/citation.cfm?doid=2487788.2487793

HILLMANN, Diane I., MARKER, Rhonda and BRADY, Chris, 2008. Metadata Standards and Applications. *The Serials Librarian*. 19 May 2008. Vol. 54, no. 1–2, p. 7–21. DOI 10.1080/03615260801973364.

HYLAND, Bernadette, ATEMEZING, Ghislain, PRENDLETON, Michael and SRIVASTAVA, Biplav, 2013. Linked Data Glossary. *W3C* [online]. 27 June 2013. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/ld-glossary/

ISAAC, Antoine and CHARLES, Valentine, 2016. Guidelines for submitting IIIF resources for objects in EDM. *Europeana Pro* [online]. 25 April 2016. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/IIIFtoEDM_profile_042016.pdf

ISAAC, Antoine and CLAYPHAN, Robina, 2013. Europeana Data Model Primer. *Europeana Pro* [online]. 14 July 2013. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf

ISAAC, Antoine, 2019. IIIF and the Europeana mission. *2019 IIIF Conference* [online]. Göttingen, Germany. 26 June 2019. [Accessed 1 August 2020]. Available from: https://www.slideshare.net/antoineisaac/iiif-and-the-europeana-mission

JACQUESSON, Alain, ROTEN, Gabrielle von and LEVRAT, Bernard, 2019. *Histoire d'une (r)évolution: l'informatisation des bibliothèques genevoises, 1963-2018*. ISBN 978-2-940587-11-7.

JAFFE, Rachel, 2017. Dublin Core Metadata Schema. *University of California Santa Cruz* [online]. 24 January 2017. [Accessed 1 August 2020]. Available from: https://guides.library.ucsc.edu/c.php?g=618773&p=4306386

KEITH, Alexander, CYGANIAK, Richard, HAUSENBLAS, Michael and ZHAO, Jun, 2011. Describing Linked Datasets with the VoID Vocabulary. *W3C* [online]. 3 March 2011. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/void/

KLEIN, Martin, SANDERSON, Robert, VAN DE SOMPEL, Herbert, WARNER, Simeon, HASLHOFER, Bernhard, LAGOZE, Carl and NELSON, Michael L., 2013. A Technical Framework for Resource Synchronization. *D-Lib Magazine*. January 2013. Vol. 19, no. 1/2. DOI 10.1045/january2013-klein.

LAGOZE, Carl, VAN DE SOMPEL, Herbert, NELSON, Michael and WARNER, Simeon, 2002. The Open Archives Initiative Protocol for Metadata Harvesting - v.2.0. *Open Archives Initiative* [online]. 14 June 2002. [Accessed 1 August 2020]. Available from: http://www.openarchives.org/OAI/openarchivesprotocol.html

LEMMER WEBBER, Christopher and TALLON, Jessica, 2018. ActivityPub. *W3C* [online]. 23 January 2018. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/activitypub/

LIM, Shirley and LI LIEW, Chern, 2011. Metadata quality and interoperability of GLAM digital images. *Aslib Proceedings*. 20 September 2011. Vol. 63, no. 5, p. 484–498. DOI 10.1108/00012531111164978.

LINDLAR, Michelle, 2020. A practical case study about metadata. *Building a Digital Future : Challenges & Solutions for Preserving 3D Models* [online]. Online DPC Briefing Day. 30 April 2020. [Accessed 1 August 2020]. Available from: https://www.dpconline.org/docs/miscellaneous/events/2020-events/2269-mickylindlar-metadata-3d/file

LOVREČIĆ, Katarina, 2010. InTech Supports the Open Archives Initiative Protocol. *InTechOpen Blog* [online]. 3 November 2010. [Accessed 1 August 2020]. Available from: https://intechweb.wordpress.com/2010/11/03/intech-supports-the-open-archives-initiative-protocol/

LYNCH, Clifford A., 1997. The Z39.50 Information Retrieval Standard. *D-Lib Magazine* [online]. April 1997. Vol. 3, no. 4. [Accessed 1 August 2020]. Available from: http://www.dlib.org/dlib/april97/04lynch.html

MCKENNA, Lucy, DEBRUYNE, Christophe and O'SULLIVAN, Declan, 2018. Understanding the Position of Information Professionals with regards to Linked Data: A Survey of Libraries, Archives and Museums. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* [online]. Fort Worth, Texas, USA: Association for Computing Machinery. 23 May 2018. p. 7–16. [Accessed 1 August 2020]. JCDL '18. ISBN 978-1-4503-5178-2. Available from: https://doi.org/10.1145/3197026.3197041

MCKENNA, Lucy, DEBRUYNE, Christophe and O'SULLIVAN, Declan, 2020. NAISC-L: An Authoritative Linked Data Interlinking Approach for the Library Domain. *Europeana Pro* [online]. 7 May 2020. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/page/issue-15-swib-2019#naisc-l-an-authoritative-linked-data-interlinking-approach-for-the-library-domain

MITCHELL, Erik T., 2013. Chapter 1: Metadata Developments in Libraries and Other Cultural Heritage Institutions. *Library Technology Reports*. 6 August 2013. Vol. 49, no. 5, p. 5–10.

NEALE, Andy and CHARLES, Valentine, 2020. MS68 Metis strategic recommendations M18: *Aggregation Strategy*. Europeana DSI-4. The Hague, Netherlands: Europeana Foundation.

PARECKI, Aaron, 2017. Webmention. *W3C* [online]. 12 January 2017. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/webmention/

RABUN, Sheila, 2016. *Scoping the "IIIF universe": First Steps to Discovery*. Research Proposal. Seattle, WA: Information School, University of Washington.

RABUN, Sheila, 2017. IIIF Community Newsletter, Volume 1 Issue 3. *International Image Interoperability Framework* [online]. 25 May 2017. [Accessed 1 August 2020]. Available from: https://iiif.io/news/2017/05/25/newsletter/#community-snapshot

RAEMY, Julien Antoine and FREIRE, Nuno, 2020. *Overview of alternative technologies for (meta)data aggregation* [online]. 6 May 2020. [Accessed 1 August 2020]. Europeana Aggregators Forum (EAF) Spring 2020. Available from: https://vimeo.com/410919947/3023e31d7c

RAEMY, Julien Antoine and SCHNEIDER, René, 2019. *Suggested Measures for Deploying IIIF within Swiss Organisations* [online]. White paper. Geneva, Switzerland: HES-SO University of Applied Sciences and Arts, Haute école de gestion de Genève. [Accessed 1 August 2020]. Available from: https://doi.org/10.5281/zenodo.2640415

RAEMY, Julien Antoine, 2017. *The International Image Interoperability Framework (IIIF): raising awareness of the user benefits for scholarly editions* [online]. Bachelor's thesis. Geneva, Switzerland: Haute école de gestion de Genève. [Accessed 1 August 2020]. Available from: https://doc.rero.ch/record/306498

RAEMY, Julien Antoine, 2020a. *Survey results on alternative aggregation mechanisms (anonymised version)*. Data set. 15 June 2020 [Accessed 1 August 2020]. Available from: https://doi.org/10.5281/zenodo.3966692

RAEMY, Julien Antoine, 2020b. *Enabling better aggregation and discovery of cultural heritage content for Europeana and its partner institutions. Master's thesis oral defence*. 28 August 2020 [Accessed 30 August 2020]. Available from: https://doi.org/10.5281/zenodo.3975522

REISS, Kevin, 2007. SRU, Open Data and the Future of Metasearch. *Internet Reference Services Quarterly*. 20 September 2007. Vol. 12, no. 3–4, p. 369–386. DOI 10.1300/J136v12n03_09.

RILEY, Jenn and BECKER, Devin, 2010. Seeing Standards: a visualization of the metadata universe. *Information Package: Metadata* [online]. 2010. [Accessed 1 August 2020]. Available from: https://lis4206metadata.files.wordpress.com/2012/04/seeingstandards1.jpg

ROBSON, Glen, APPLEBY, Michael, CRANE, Tom, SANDERSON, Robert, STROOP, Jon and WARNER, Simeon, 2020. *Technical Roadmap Session* [online]. 4 June 2020. [Accessed 1 August 2020]. IIIF Week 2020. Available from: https://www.youtube.com/watch?v=5Vx0W5XQcWQ

SANDERSON, Robert, 2018. IIIF Discovery Walkthrough. *2018 IIIF Conference* [online]. Library of Congress, Washington DC, USA. 24 May 2018. [Accessed 1 August 2020]. Available from: https://www.slideshare.net/azaroth42/iiif-discovery-walkthrough

SANDERSON, Robert, 2020a. Tiers of Abstraction and Audience in Cultural Heritage Data Modeling. *Information Access Seminar* [online]. Berkeley CA, USA. 13 March 2020. [Accessed 1 August 2020]. Available from: https://www.slideshare.net/azaroth42/tiers-of-abstraction-and-audience-in-cultural-heritage-data-modeling-230217697

SANDERSON, Robert, 2020b. The Importance of being LOUD. *LODLAM 2020* [online]. Los Angeles, CA. 5 February 2020. [Accessed 1 August 2020]. Available from: https://www.slideshare.net/azaroth42/the-importance-of-being-loud

SCHOLZ, Henning, 2019a. Europeana Publishing Guide v1.8. *Europeana Pro* [online]. 31 July 2019. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Publishing%20Guide%20v1.8.pdf

SCHOLZ, Henning, 2019b. Publishing Framework. *Europeana Pro* [online]. 1 November 2019. [Accessed 1 August 2020]. Available from: https://pro.europeana.eu/post/publishing-framework

SCHONFELD, Uri and SHIVAKUMAR, Narayanan, 2009. Sitemaps: above and beyond the crawl of duty. In: *Proceedings of the 18th international conference on World wide web* [online]. Madrid, Spain: Association for Computing Machinery. 20 April 2009. p. 991–1000. [Accessed 1 August 2020]. WWW '09. ISBN 978-1-60558-487-4. Available from: https://doi.org/10.1145/1526709.1526842

SMITH-YOSHIMURA, Karen, 2018. Analysis of 2018 International Linked Data Survey for Implementers. *The Code4Lib Journal* [online]. 8 November 2018. No. 42. [Accessed 1 August 2020]. Available from: https://journal.code4lib.org/articles/13867

SNELL, James M. and PRODROMOU, Evan, 2017a. Activity Streams 2.0. *W3C* [online]. 23 May 2017. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/activitystreams-core/

SNELL, James M. and PRODROMOU, Evan, 2017b. Activity Vocabulary. *W3C* [online]. 23 May 2017. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/activitystreams-vocabulary/

SNYDMAN, Stuart, SANDERSON, Robert and CRAMER, Tom, 2015. The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. In: *Archiving Conference* [online]. Los Angeles, CA: IS&T. May 2015. p. 16–21. [Accessed 1 August 2020]. Available from: https://purl.stanford.edu/df650pk4327

SPEICHER, Steve, ARWE, John and MALHOTRA, Ashok, 2015. Linked Data Platform 1.0. *W3C* [online]. 26 February 2015. [Accessed 1 August 2020]. Available from: https://www.w3.org/TR/ldp/

VAN DE SOMPEL, Herbert and NELSON, Michael L., 2015. Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Magazine*. November 2015. Vol. 21, no. 11/12. DOI 10.1045/november2015-vandesompel.

VANDER SANDE, Miel, VERBORGH, Ruben, HOCHSTENBACH, Patrick and VAN DE SOMPEL, Herbert, 2018. Toward sustainable publishing and querying of distributed Linked Data archives. *Journal of Documentation*. 1 January 2018. Vol. 74, no. 1, p. 195–222. DOI 10.1108/JD-03-2017-0040.

WALLIS, Richard, ISAAC, Antoine, CHARLES, Valentine and MANGUINHAS, Hugo, 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. *The Code4Lib Journal* [online]. 14 June 2017. No. 36. [Accessed 1 August 2020]. Available from: https://journal.code4lib.org/articles/12330

WARNER, Simeon, 2017. Discovery of IIIF resources. *2017 IIIF Conference* [online]. Augustinianum, The Vatican City. 6 June 2017. [Accessed 1 August 2020]. Available from: https://docs.google.com/presentation/d/12M_oOwwXtOZLfgAeqkJDwFWaOux_0mdmF6doyjYPSzM/edit

WHALEN, Maureen, 2016. Rights Metadata Made Simple. In: *Introduction to Metadata* [online]. Third edition. Los Angeles, CA, USA: Getty Research Institute. [Accessed 1 August 2020]. ISBN 978-1-60606-479-5. Available from: https://www.getty.edu/publications/intrometadata/rights-metadata/

WITT, Jeffrey, 2017a. IIIF and Linked Data Notifications - Thoughts and Reflections. *LombardPress* [online]. 28 February 2017. [Accessed 1 August 2020]. Available from: https://lombardpress.org/2017/02/28/datasharing-iiif-and-ldn/

WITT, Jeffrey, 2017b. Linking Research, the SCTA, LombardPress, and LinkedData Notifications. *LombardPress* [online]. 24 January 2017. [Accessed 1 August 2020]. Available from: https://lombardpress.org/2017/01/24/linking-research/

WOLF, Mischa, 1998. Data Model Working Draft. *DCMI* [online]. 7 October 1998. [Accessed 1 August 2020]. Available from: https://www.dublincore.org/specifications/dublin-core/datamodel/

ZENG, Marcia Lei and QIN, Jian, 2016. *Metadata*. 2nd edition. Chicago, IL, USA: Neal-Schuman. ISBN 978-1-55570-965-5.

Appendix 1: Research Stakeholders

Table 17: Research Stakeholders

Entity *(main contact people)*	**Short description**
Aggregators	Intermediaries that gather data from a specific country, region or domain
Data providers	CHIs that provide metadata and links to digitised surrogates onto the Europeana platform
Europeana Aggregators Forum (EAF)	A biannual event aimed at fostering deeper working relationships between aggregators.
Europeana Aggregation Services *Valentine Charles*	A service who is, among other things, responsible for Metis
Europeana Common Culture (ECC) *Cosmina Berta, Nuno Freire, Enno Meijers, Erwin Verbruggen*	A two-year project (2019-2020) aimed at improving the quality of content and metadata.
Europeana Foundation (EF)	A non-profit foundation based in The Hague
Europeana Network Association (ENA)	A community of experts working in the field of digital cultural heritage
Europeana Platform Services *Hugo Manguinhas*	A range of services covering the Europeana's application programming interfaces (APIs) and other supporting services
Europeana Research and Development (R&D) *Antoine Isaac, Nuno Freire, Albin Larsson*	A department of Europeana dealing with data exchange, data quality, multilingualism and search
EuropeanaTech *Gregory Markus*	A community of experts, developers, and researchers from the R&D sector within the broader ENA.
Poznan Supercomputing and Networking Center (PSNC)	A partner who is responsible for hosting and developing Europeana's data infrastructure

(Europeana 2015; 2017; 2019b)

Appendix 2: Europeana's current ingestion process

Table 18: Ingestion process in Metis[39]

Step	Description
1) Import	Once a dataset with a unique identifier has been created, it can be imported either via a HTTP request (ZIP file hosted on a server) or via OAI-PMH.
2) **Validate** (external)	Validation phase of an external EDM, i.e. a simplified version of EDM as described in the Guidelines.
3) Transform	The metadata goes to a conversion phase from external EDM external to internal EDM.
4) **Validate** (internal)	The metadata transformed into internal EDM, which is the data model published by Europeana and documented on their application programming interface (API) is then validated.
5) Normalise	This phase cleans up the data (for example, by removing white spaces or certain characters or even by checking that language tags use the three-letter ISO standard for languages).
6) Enrich	Enrich data to point them to Linked Data vocabularies if links are present in the metadata. This phase is carried out using an algorithm.
7) Process Media	This phase is used to look at the links pointing to media resources so that a thumbnail is generated from the image or the video. This phase is also used to extract the technical metadata to create facets (e.g. by colour).
8) Preview	This is a final verification phase to preview the data.
9) Publish	The aggregated CHOs and their associated metadata are published on Europeana.

(Clayphan, de Hoog, Charles 2017; Scholz 2019a; Neale, Charles 2020)

[39] The specific details of each ingestion step were discussed during a videoconference with Valentine Charles on 5 March 2020 where she conducted a demo of Metis.

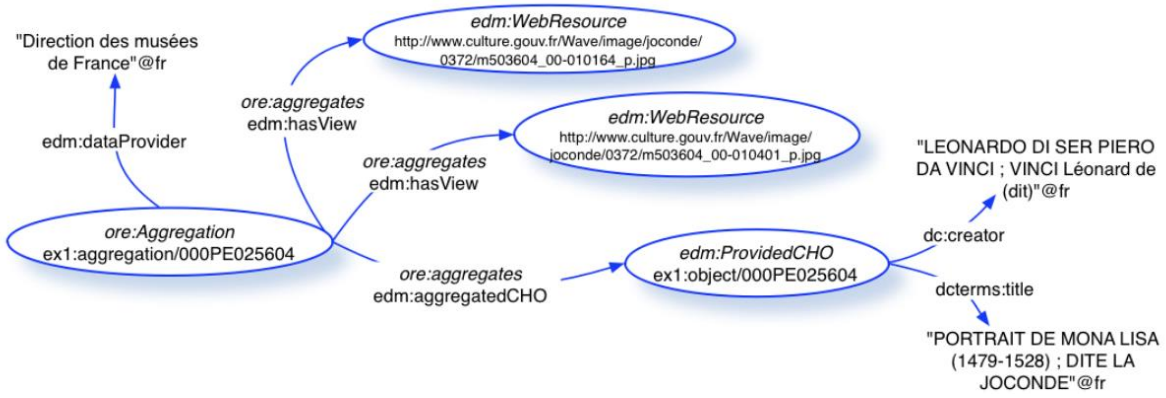
Appendix 3: Mapping examples of the Mona Lisa in EDM

The following three figures are examples[40] of representations of an aggregated CHO ("*Portrait of Mona Lisa (1479-1528); Dite La Joconde*") in EDM available on the Europeana platform at the following URLs:

- User Interface: https://www.europeana.eu/en/item/03919/public_mistral_joconde_fr_ACTION_CHERCHER_FIELD_1_REF_VALUE_1_000PE025604
- API (JSON-LD): https://www.europeana.eu/en/item/03919/public_mistral_joconde_fr_ACTION_CHERCHER_FIELD_1_REF_VALUE_1_000PE025604.json

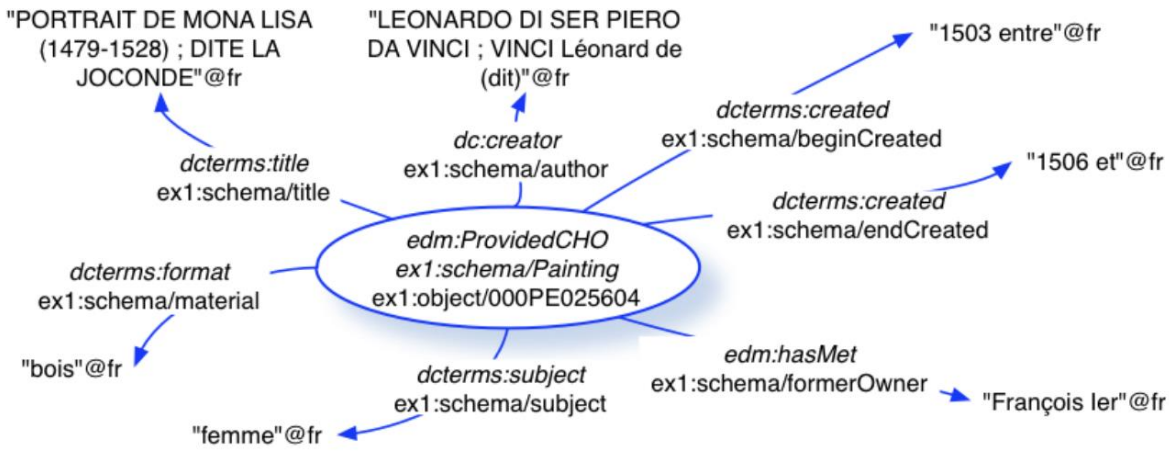
Figure 28 is a fairly basic representation of the provider's aggregation with descriptive metadata. Figure 29 and Figure 30 are examples having more precise descriptions, the first with an object-centric perspective and the second with an event-centric perspective.

Figure 28: Simple representation of the Mona Lisa in EDM



(Isaac, Clayphan 2013, p. 12)

Figure 29: Object-centric representation of the Mona Lisa in EDM



(Isaac, Clayphan 2013, p. 19)

[40] For more information, please consult the EDM Primer: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf

Figure 30: Event-centric representation of the Mona Lisa in EDM

(Isaac, Clayphan 2013, p. 19)

Appendix 4: Survey invitation and reminder

Survey invitation on EuropeanaTech's listserv (20 April)[41]

Dear all,

In the context of a Master's thesis in Information Science that I am carrying out in cooperation with Europeana R&D team, I would like to invite you to take part in **a survey to gauge the use and interest of alternative technologies other than OAI-PMH for (meta)data aggregation** (such as methods based on Linked Open Data, IIIF, ResourceSync, etc.) within the Europeana Network.

This survey (https://forms.gle/iq2fZ8wCgBMGTrDq6) should take you between 10 to 15 minutes to complete. **Thank you for filling it out by Friday, May 8.**

Please do not hesitate to contact me if you have any questions or comments.

Kind regards,

Julien A. Raemy ▪ Master's student in Information Science

HES-SO University of Applied Sciences and Arts Western Switzerland

Rue de la Tambourine 17

CH-1227 Carouge

@julsraemy ▪ https://julsraemy.github.io/

Survey reminder on EuropeanaTech's listserv (4 May)[42]

Hi everyone,

This is a reminder that the survey on *alternative aggregation methods* is still open until **Friday, May 8**. All data providers and aggregators within the Europeana Network are welcome to participate. If you haven't had a chance to fill it out yet, you can do it via this Google Form: https://forms.gle/iq2fZ8wCgBMGTrDq6

I would also like to thank the participants who have already responded to the survey and I shall be happy to answer any questions or comments you may have.

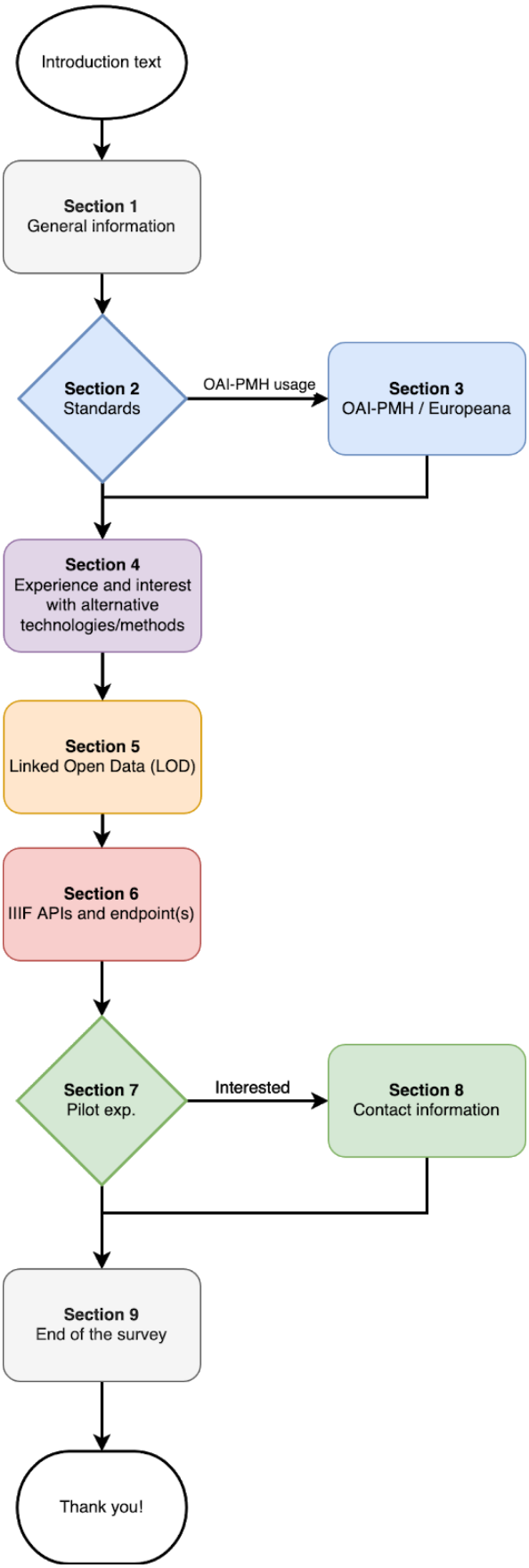
Stay healthy and all the best,

Julien A. Raemy

[41] https://list.ecompass.nl/listserv/cgi-bin/wa?A2=EUROPEANA-TECH;3eb9d145.2004
[42] https://list.ecompass.nl/listserv/cgi-bin/wa?A2=EUROPEANA-TECH;c4d154a7.2005

Appendix 5: Survey structure

Figure 31: General structure of the survey on alternative aggregation mechanisms



Appendix 6: Survey questions

Greeting message

This is a survey on the interest and possible implementations of alternative methods to OAI-PMH (such as mechanisms based on Linked Open Data, the International Image Interoperability Framework - IIIF, ResourceSync, etc.) that organisations could adopt to have their data ingested on Europeana or other aggregator platforms. It will also allow pilot experiments to be set up with interested organisations.

This survey, which should take you 10 to 15 minutes, is carried out in the context of a master's thesis in Information Science with the partnership of Europeana R&D team.

Section 1 – General information

1) Name of the organisation: _____ ***[43]

2) Institution type or institutional domain (*Please the most suitable option, several choices are allowed if your organisation belongs to or represents various domain*) ***

   - ☐ Gallery
   - ☐ Library
   - ☐ Archive
   - ☐ Museum
   - ☐ Research institute
   - ☐ Industry
   - ☐ Aggregator
   - ☐ Other: _____

Section 2 – Standards for sharing metadata over OAI-PMH and other exchange mechanisms

3) Metadata for publishing and exchanging purposes - please tick the most appropriate answer(s) ***

In the following question you will find a list of different metadata schemes or models to publish and exchange cultural heritage objects.

Table 19: Options for survey question 3

Rows	Columns (checkbox grid)
1. Bibliographic Framework (BIBFRAME) 2. CIDOC-CRM 3. Dublin Core (DC) 4. Encoded Archival Description (EAD) 5. Europeana Data Model (EDM) 6. Lightweight Information Describing Objects (LIDO) 7. Linked.art	☐ I don't know this scheme/model. ☐ I am familiar with it. ☐ I don't use it. ☐ I am interested in using it. ☐ I use it.

[43] *** denotes a required question

Rows	Columns (checkbox grid)
8. Machine-readable cataloguing (MARC) Standards 9. Metadata Object Description Schema (MODS) 10. Resource Description and Access (RDA) 11. Records in Contexts (RiC) 12. Schema.org	

4) Do you use other metadata standards for publication and exchange purposes than those outlined in the previous question? If so, which one(s)?

5) Metadata serialisations - please tick the most appropriate answer(s) ***

In the following question you will find a list of different metadata serialisations.

Table 20: Options for survey question 5

Rows	Columns (checkbox grid)
1. Comma-separated values (CSV) 2. JavaScript Object Notation (JSON) 3. MARCXML or MarcXchange 4. Resource Description Framework in Attributes (RDFa) 5. RDF serialisations (RDF/XML, Turtle, Notation3, N-Triples, JSON-LD) 6. Extensible Markup Language (XML)	☐ I don't know this serialisation. ☐ I am familiar with it. ☐ I am interested in using it. ☐ I use it.

6) Do you use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)? ***

- Yes → section 3
- No → section 4
- I don't know → section 4

Section 3 – OAI-PMH / Europeana

7) Is your OAI-PMH server used for anything other than aggregation towards the Europeana platform? ***

- Yes
- No

Section 4 – Experience and interest with alternative mechanisms for harvesting (meta)data

8) Aggregation mechanisms – please select the most appropriate answer ***

In the following question you will find a list of ten possible ways to harvest (meta)data. This question is meant to gauge awareness and interest in these different mechanisms.

Table 21: Options for survey question 8

Rows	Columns (checkbox grid)
1. Aggregation via Sitemaps and Schema.org in HTML pages (as for Internet Search Engines) 2. Aggregation via Linked Open Data (LOD) datasets 3. Aggregation of International Image Interoperability Framework (IIIF) based on IIIF Collections 4. Aggregation of IIIF based on Sitemaps 5. Aggregation of IIIF based on IIIF Change Discovery API (ActivityStreams) 6. Linked Data Notifications (LDN) 7. Linked Data Platform (LDP) 8. Open Publication Distribution System (OPDS) 9. ResourceSync (RS) in conjunction with WebSub 10. Webmention	☐ I don't know this technology/method. ☐ I am familiar with it. ☐ I am interested in it. ☐ I have already tested it. ☐ I have implemented it.

Section 5 – Linked Open Data (LOD)

9) Publishing LOD – please tick the most appropriate answer(s) ***

Table 22: Options for survey question 9

Rows	Columns (checkbox grid)
1. Within HTML pages (RDFa, embedded JSON-LD) 2. SPARQL endpoint 3. HTTP Content Negotiation 4. RDF file dumps 5. Header Dictionary Triples (HDT) 6. Linked Data Fragments (LDF)	☐ I don't know this technology/method. ☐ I am familiar with it. ☐ I am interested in it. ☐ I have already tested it. ☐ I have implemented it.

10) Do you publish LOD in any other way than those outlined in the previous question? If so, which one(s)?

11) If you have published LOD, could you provide examples? (Please provide URLs)

Section 6 – IIIF APIs and endpoint(s)

12) IIIF APIs – please tick the most appropriate answer(s) ***

Table 23: Options for survey question 12

Rows	Columns (checkbox grid)
1. IIIF Image API 2. IIIF Presentation API 3. IIIF Content Search API 4. IIIF Authentication API	☐ I don't know this API. ☐ I am familiar with it. ☐ I am interested in deploying it. ☐ I have already tested it. ☐ I have implemented it.

13) If you've already deployed a IIIF-compliant solution, what is/are your IIIF endpoint(s)? (Please provide URLs)

Section 7 – Aggregation pilots

14) Would you be interested in participating in a pilot experiment in May 2020 where a subset of your metadata could be aggregated by any of the mentioned alternatives? ***

- Yes → section 8
- No, but I'm interested to know more about the study outcomes → section 8
- No → section 9

Section 8 – Contact information

15) Please provide your email address if you are interested in participating in a pilot experiment or if you're interested in the study outcomes. Email: _____ ***

Section 9 – End of the survey

Thank you very much for participating in this survey!

16) Do you have any further comments, for example are you interested in another technology for metadata aggregation?

Appendix 7: Identified resources to support aggregation

Table 24: Resources for facilitating alternative (meta)data aggregation

Resource	Short description
Specifying a linked data dataset for Europeana and aggregators	Report which identifies the technical requirements for LOD harvesting (through a downloadable dataset distribution, a listing of URIs, or by specifying a SPARQL service).
Guidelines for providing and handling Schema.org metadata in compliance with Europeana	Guidelines which provides a general level of guidance for usage of Schema.org metadata that, after conversion to EDM, will results in metadata that is suitable for aggregation by Europeana.
Europeana Metadata Testing Tool	This prototype aims to support data providers of Europeana in checking the compliance of their implementation of solutions for delivering data to Europeana. It includes the following: • Wikidata entity about a cultural heritage object • Schema.org included in a webpage (as used for Internet search engines) • IIIF manifest with structured metadata in Schema.org or EDM • sitemap.xml and robots.txt of a website • Evaluate data quality of a record at Europeana
Awesome IIIF	A list of lists of IIIF resources (compliant servers and viewers, IIIF Manifests Tools, API Libraries, etc.)
NDE's LOD-aggregator	The LOD-aggregator, which was created for the ECC project, harvests the published Linked Data and converts the Schema.org information to EDM to make ingestion onto the Europeana harvesting platform possible. This tool can do the following process: crawling service to harvest the data described by a dataset description, mapping to EDM, and validation.
Data Aggregation Lab (DAL)	DAL is a workbench system which can currently do the following: • Register a LOD dataset for aggregation by Europeana • Register a IIIF dataset for aggregation by Europeana (IIIF Collections, Sitemaps, IIIF Change Discovery API 0.3) • Register a WWW dataset for aggregation by Europeana
Targl	Command-line tool that convert CSV files to RDF using SPARQL 1.1 (CONSTRUCT queries)
Metadata Ingestion Services (MINT) Login/Register	Web-based platform that was designed and developed to facilitate aggregation initiatives for CH content and metadata in Europe where registered organisations can upload (http, ftp, oai-pmh) their metadata records in XML or CSV in order to manage, aggregate and publish their collections.
Muzz.app	Data aggregation platform with built-in exports, such as a converter into EDM/XML (from LIDO) and IIIF Collections creation. Pricing varies from €20 to €100 per month, depending on the number of users and storage capacity.
Metis Sandbox	TBD

Appendix 8: Survey findings and pilot follow-up email templates

Template 1: Interest in the study outcomes + out of scope

Hello,

I hope this email finds you well.

I am writing to you as a follow-up to the survey I conducted in May on alternative aggregation mechanisms in the context of a master's thesis I am carrying out in collaboration with Europeana.

You had indicated that you were interested in obtaining the survey findings and I am thus giving you a link to an anonymised summary in which you also have the possibility to provide comments until Tuesday, June 30th: https://docs.google.com/document/d/10TgY8OolEgCn-9TxZv_RhAma4tDPxqMj1jFWw557hiQ/edit?usp=sharing

If you would like to take part in aggregation pilots for testing alternative aggregation routes towards the Europeana platform at a later stage, you can always contact Antoine Isaac and Nuno Freire (both at Europeana R&D) who are in CC of this email. Meanwhile, we are still investigating aggregation pilots that can be completed in the short term, i.e. before the end of my master's thesis in August. Naturally all the important outcomes will be included in my dissertation, which will be published in a few months.

Thank you again for taking part in the survey and please do not hesitate to contact me if you have any questions or comments.

Kind regards,

Julien A. Raemy

Template 2: Dismiss

Hello,

I hope this email finds you well.

I am writing to you because you took part in the survey I conducted in May on alternative aggregation mechanisms in the context of a master's thesis that I am carrying out in collaboration with Europeana and because you expressed your interest in participating in an aggregation pilot.

First, you had indicated that you were interested in obtaining the survey findings and I am thus giving you a link to an anonymised summary in which you also have the possibility to provide comments until Tuesday, June 30th: https://docs.google.com/document/d/10TgY8OolEgCn-9TxZv_RhAma4tDPxqMj1jFWw557hiQ/edit?usp=sharing

Secondly, after assessment, it appears that none of the alternative aggregation routes highlighted in the survey are deployable at this time, but if you think that there is a possibility,

I encourage you contact us by Tuesday June 30th (myself as well as Nuno Freire and Antoine Isaac in CC) to determine what kind of pilots could be feasible.

Meanwhile, we are still investigating pilots that can be completed in the short term, i.e. before the end of my master's thesis in August. Naturally all the important outcomes will be included in my dissertation, which will be published in a few months.

Also, if you would like to set up an aggregation pilot for testing alternative aggregation routes towards the Europeana platform at a later stage, I also encourage you to contact Antoine Isaac and Nuno Freire (both at Europeana R&D).

Thank you again for taking part in the survey and please do not hesitate to contact me if you have any further questions or comments.

Kind regards,

Julien A. Raemy

Template 3: Defer

Dear <name>,

I hope this email finds you well.

I am writing to you because you took part in the survey I conducted in May on alternative aggregation mechanisms in the context of a master's thesis that I am carrying out in collaboration with Europeana and because you expressed your interest in participating in an aggregation pilot.

First, you had indicated that you were interested in obtaining the survey findings and I am thus giving you a link to an anonymised summary in which you also have the possibility to provide comments until Tuesday, June 30th: https://docs.google.com/document/d/10TgY8OolEgCn-9TxZv_RhAma4tDPxqMj1jFWw557hiQ/edit?usp=sharing

Secondly, after assessment, it appears that the interests you mentioned for the different aggregation alternatives (<mechanism(s)>) are not possible at the moment but we would be happy to see what is feasible to assist you in this effort and perhaps to carry out a pilot at a later stage.

Meanwhile, we are still investigating pilots that can be completed in the short term, i.e. before the end of my master's thesis in August. Naturally, all the important outcomes will be included in my dissertation, which will be published in a few months., If indeed, you would like to set up a future aggregation pilot for testing alternative aggregation routes towards the Europeana platform, I would encourage you to contact Antoine Isaac and Nuno Freire (both at Europeana R&D and in CC of this email).

Thank you again for taking part in the survey and please do not hesitate to contact me if you have any further questions or comments.

Kind regards,

Julien A. Raemy

Template 4: Investigate

Dear <name>,

I hope this email finds you well.

I am writing to you because you took part in the survey I conducted in May on alternative aggregation mechanisms in the context of a master's thesis that I am carrying out in collaboration with Europeana and because you expressed your interest in participating in an aggregation pilot.

First, you had indicated that you were interested in obtaining the survey findings and I am thus giving you a link to an anonymised summary in which you also have the possibility to provide comments until Tuesday, June 30th: https://docs.google.com/document/d/10TgY8OolEgCn-9TxZv_RhAma4tDPxgMj1jFWw557hiQ/edit?usp=sharing

Secondly, after assessment, it appears that an alternative aggregation route could be possible (it should be noted that is not necessarily an aggregation route that you mention in the survey and your organisation may not be interested in).

However, I would like to have the following information to be sure that the following aggregation pilot(s) is(are) feasible:

<aggregation route(s)>	<question(s)>

Would you please provide us with more information by **June 30th** to assess whether such a pilot could be considered, if it is feasible in the short term and how it could be conducted?

If you would like to set up an aggregation pilot for testing alternative aggregation routes towards the Europeana platform at a later stage, i.e. after the end of master's thesis in August 2020, I would encourage you to contact Antoine Isaac and Nuno Freire (both at Europeana R&D and in CC of this email).

Thank you again for taking part in the survey and please do not hesitate to contact me if you have any further questions or comments. Naturally, all the important outcomes of the survey findings and aggregation pilots will be included in my dissertation, which should be published in a few months.

Kind regards,

Julien A. Raemy

Template 5: Wikimedia-affiliated

Dear <name>,

I hope this email finds you well.

I am writing to you because you took part in the survey I conducted in May on alternative aggregation mechanisms in the context of a master's thesis that I am carrying out in collaboration with Europeana and because you expressed your interest in participating in an aggregation pilot.

First, you had indicated that you were interested in obtaining the survey findings and I am thus giving you a link to an anonymised summary in which you also have the possibility to provide comments until Tuesday, June 30th: https://docs.google.com/document/d/10TgY8OolEgCn-9TxZv_RhAma4tDPxgMj1jFWw557hiQ/edit?usp=sharing

After assessment, since there is a rather special case between Europeana and Wikimedia, I wondered how a pilot to aggregate content onto the Europeana platform could make sense, seeing rather the collaboration between the two entities in a different way. Indeed, data from Wikidata can enrich those from Europeana and vice versa. But perhaps I may be wrong, and I was wondering if a special case of pilot could occur. If that's the case, can you tell me until June 30th and then we'll try to organize some experiments in the coming weeks.

Finally, if you would like to set up an aggregation pilot for testing alternative aggregation routes towards the Europeana platform at a later stage, i.e. after the end of master's thesis in August 2020, I would encourage you to contact Antoine Isaac and Nuno Freire (both at Europeana R&D and in CC of this email).

Thank you again for taking part in the survey and please do not hesitate to contact me if you have any further questions or comments. Naturally, all the important outcomes of the survey findings and aggregation pilots will be included in my dissertation, which should be published in a few months.

Kind regards,

Julien Raemy

Appendix 9: Overview of aggregation mechanisms

Table 25: High-level overview of aggregation mechanisms

Technology	URL	Version	Date	Aggregation component	Short description	Governance	HTTP Requests	Serialisations	Notification
ActivityStreams 2.0 (AS2)	https://www.w3.org/TR/activitystreams-core/	2	2017-05-23	Data transfer and synchronisation	Syntax and vocabulary for representing potential and completed activities	World Wide Web Consortium (W3C) Part of the Social Web Protocols	GET HEAD POST	JSON-LD	*N/A*
ActivityPub (AP)	https://www.w3.org/TR/activitypub/	1	2018-01-23	Data transfer and synchronisation	Client to server API for creating, updating and deleting content, as well as a federated server to server API for delivering notifications and content	World Wide Web Consortium (W3C) Part of the Social Web Protocols	GET HEAD POST	JSON-LD	*N/A*
IIIF Change Discovery API	https://iiif.io/api/discovery/0.9/	0.9	2020-06-04	Data transfer and synchronisation	Machine to machine API that provides the information needed to discover and subsequently make use of IIIF resources	International Image Interoperability Framework (IIIF)	GET HEAD	JSON-LD	*N/A*
IIIF Content State API	https://iiif.io/api/content-state/0.2/	0.2	2018-10-31	Data transfer and synchronisation	Describes the current or desired state of the content that a IIIF-compliant client is rendering to a user	International Image Interoperability Framework (IIIF)	GET HEAD	JSON-LD	*N/A*
IIIF Image API	https://iiif.io/api/image/3.0/	3	2020-06-03	Data transfer and synchronisation	Web service that returns an image in response to a standard HTTP(S) request	International Image Interoperability Framework (IIIF)	GET HEAD	JSON-LD	*N/A*
IIIF Presentation API	https://iiif.io/api/presentation/3.0/	3	2020-06-03	Data transfer and synchronisation	Provides the necessary information about the object structure and layout of IIIF resources	International Image Interoperability Framework (IIIF)	GET HEAD	JSON-LD	*N/A*
Linked Data Notifications (LDN)	https://www.w3.org/TR/ldn/	1	2017-05-02	Data transfer and synchronisation	Describing how servers can have messages pushed to them by applications.	World Wide Web Consortium (W3C) Part of the Social Web Protocols	GET HEAD POST	JSON-LD Other RDF serialisations are allowed through HTTP Content Negotiation	Pull
Linked Data Platform (LDP)	https://www.w3.org/TR/ldp/	1	2015-02-26	Data transfer and synchronisation	Architecture for read-write Linked Data	World Wide Web Consortium (W3C)	GET HEAD POST PUT DELETE PATCH OPTIONS	HTTP Headers Turtle JSON-LD	*N/A*
Open Publication Distribution System Catalog 2.0 (OPDS2)	https://drafts.opds.io/opds-2.0	2	*2020-03-03 (Draft)*	Data transfer and synchronisation	Aggregation and distribution of electronic publications	ODPS Working Group Feedbooks	GET HEAD	JSON-LD	Pull
ResourceSync Framework Specification (RS)	http://www.openarchives.org/rs/1.1/resourcesync	1.1	2017-02-02	Data transfer and synchronisation	Data Harvesting mechanism that allows third-party stems to remain synchronized	Open Archives Initiative (OAI) National Information Standards Organization (NISO)	GET HEAD POST	XML	Pull and Push
Sitemaps	https://www.sitemaps.org/protocol.html	0.9	2006-11-16	Data transfer and synchronisation	Agreed protocol for web crawling	Google Yahoo Microsoft	GET	XML	*N/A*
The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)	http://www.openarchives.org/OAI/openarchivesprotocol.html	2	2002-06-14	Data transfer and synchronisation	Metadata harvesting of records stored in archives/repositories	Open Archives Initiative (OAI)	GET HEAD POST	XML	Pull
Webmention	https://www.w3.org/TR/webmention/	1	2017-01-12	Data transfer and synchronisation	Notifying URL when mentioned on a given site	World Wide Web Consortium (W3C) Part of the Social Web Protocols	GET HEAD POST	URL Encoded (x-www-urlencoded content)	Pull

WebSub	https://www.w3.org/TR/websub/	1	2018-01-23	Data transfer and synchronisation	Mechanism for communication between publishers and their subscribers	World Wide Web Consortium (W3C) Part of the Social Web Protocols	GET HEAD POST	URL Encoded (x-www-urlencoded content)	Push
Data Catalog Vocabulary (DCAT)	https://www.w3.org/ns/dcat#	2	2020-02-04	Data modelling and representation	RDF vocabulary facilitating interoperability between different data catalogues published on the Web	World Wide Web Consortium (W3C)	*N/A*	RDF serialisations	*N/A*
Europeana Data Model (EDM)	https://pro.europeana.eu/page/edm-documentation	5.2.8	2017-10-06	Data modelling and representation	Common-top level ontology within the Europeana Network	Europeana	*N/A*	RDF serialisations	*N/A*
Schema.org	https://schema.org/docs/schemas.html	9	2020-07-21	Data modelling and representation	RDF Vocabulary that enables better structured data on the Web. It can also describe CHOs.	Google Microsoft Yahoo Yandex World Wide Web Consortium (W3C)	*N/A*	RDF serialisations CSV	*N/A*
Vocabulary of Interlinked Datasets (VoID)	https://www.w3.org/TR/void/	1	2011-03-03	Data modelling and representation	RDF vocabulary for discovering and leveraging Linked Data sets	World Wide Web Consortium (W3C)	*N/A*	RDF serialisations	*N/A*

Appendix 10: Opportunity Solution Tree (full)

Figure 32: Opportunity Solution Tree to enable better aggregation and discovery of cultural heritage content