Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata

Nuno Freire, INESC-ID
Valentine Charles, Europeana Foundation
Antoine Isaac, Europeana Foundation

In the World Wide Web, a very large number of resources is made available through digital libraries. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability, sharing and reuse of the resources. A widely-used approach is metadata aggregation, where centralized efforts like Europeana facilitate the discoverability and use of the resources by collecting their associated metadata. The cultural heritage domain embraced the aggregation approach while, at the same time, the technological landscape kept evolving. Nowadays, cultural heritage institutions are increasingly applying technologies designed for the wider interoperability on the Web. In this context, we have identified the Schema.org vocabulary as a potential technology for innovating metadata aggregation. We conducted two case studies that analysed Schema.org metadata from collections from cultural heritage institutions, and used, as evaluation criteria for this metadata, the specific requirements of the Europeana Network. These include the recommendations of the Europeana Data Model, which has been developed as a collaborative effort from all the domains represented in Europeana: libraries, museums, archives, and galleries. We concluded that Schema.org poses no obstacle that cannot be overcome to allow data providers to deliver metadata in full compliance with Europeana requirements and with the desired semantic quality. However, for specific requirements of Europeana or other aggregation networks, due to Schema.org's cross-domain applicability, its adoption must be accompanied by recommendations and/or specifications regarding how data providers should create their Schema.org metadata.

CCS Concepts: • **Information systems** → Information systems applications → Digital libraries and archives; • **Information systems** → World Wide Web → Web searching and information discovery;

General Terms: N/A

Additional Key Words and Phrases: Metadata · Cultural heritage · Metadata aggregation · Schema.org · Europeana Data Model.

ACM Reference format:
Nuno Freire, Valentine Charles, and Antoine Isaac. 2017. Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata. *ACM J. Comput. Cult. Herit.* ...., x pages.
DOI: ...

1 INTRODUCTION

In the World Wide Web, a very large number of resources is made available through digital libraries. The existence of many individual digital libraries, maintained by different organizations, brings challenges to the discoverability and usage of the resources by potential audiences.

An often-used approach is metadata aggregation, where a central organization takes the role of facilitating the discovery and use of the resources by collecting their associated metadata. Based on these aggregated datasets of metadata, the central organization (often called aggregator) is in a position to further promote the usage of the resources by means that cannot be efficiently undertaken by each digital library in isolation. This scenario is widely applied in the domain of cultural heritage (CH), where the number of organizations with their own digital libraries is very large. In Europe, Europeana has the role of facilitating the usage of CH resources from and about Europe [1], and although many European CH institutions do not yet have a presence in Europeana, it already holds metadata from over 3,700 providers [2].

In several contexts, the technological approach to metadata aggregation has been mostly based on the OAI-PMH protocol, a technology initially designed in 1999 [3]. OAI-PMH was meant to address shortcomings in scholarly communication by providing a technical interoperability solution for discovery of e-prints, via metadata aggregation. OAI-PMH allows metadata to be aggregated using any metadata schema, although its specification includes the use of the

Dublin Core Element Set [4] as the minimal metadata schema for aggregation, to enable the widest metadata interoperability across domains.

The cultural heritage domain embraced the solution offered by OAI-PMH, however, the technological landscape around our domain changed. Nowadays, cultural heritage organizations are increasingly applying technologies designed for the wider interoperability on the Web. Particularly relevant for our work are those related to the social web, the web of data, and internet search engine optimization. In this context, we have identified the Schema.org vocabulary and its associated web-based dissemination channels [5] as a potential technology for metadata aggregation in the CH domain. Our interest in Schema.org for metadata aggregation originates from our earlier work in reviewing the state of the art and emerging Web technologies for their applicability in the context of CH [6], where the relevance of Schema.org was identified from its relation to other technologies used by Internet search engines.

Europeana has recently evaluated Schema.org as a means to publish CH data [7]. This paper presents the results of our work on another application of Schema.org: to seek whether it could also bring usable data sources for CH aggregators, such as Europeana.

This paper starts by describing the motivation for evaluating Schema.org for applicability in metadata aggregation in the CH domain. It follows with a section about Schema.org, which provides a description of how it covers the representation of metadata about CH resources, the related technologies required for its processing, and the main requirements for its usage in metadata aggregation. We then present our case studies, the experimental setup and our observations and analysis on existing Schema.org metadata. The paper ends with our key conclusions regarding the impact of supporting Schema.org metadata in CH.

2 MOTIVATION AND CONTEXT

Schema.org is an activity for encouraging the publication and consumption of structured data in the Internet. Its main application is in web pages - for example, stating that a web page describes a culinary recipe, its ingredients and preparation method; or that it describes a movie, its actors, user reviews, etc… Web pages built according to the Schema.org principles (Schema.org data can be referenced or embedded in several different encodings, including RDFa[1], Microdata[2] and JSON-LD[3]) can be processed by search engines and other applications that use this structured data, in addition to text and links from the HTML body. The Schema.org website[4] reports usage in more than 10 million sites and Google, Microsoft, Pinterest, Yandex, among others, already provide services and applications that are based on the available Schema.org structured data.

Schema.org has applicability across a vast range of domains. Especially, it could allow CH institutions to reduce the overall effort on data conversion that they conduct for discovery purposes. From these institutions' point of view, Schema.org could indeed become a unified solution for allowing the discovery of their resources through *both* internet search engines and CH specific metadata aggregation networks, such as Europeana.

In the CH metadata aggregation approaches, a common practice has been to aggregate metadata using an agreed data model that allows to deal with the data heterogeneity between organizations and countries in a sustainable way. These data models typically aim to address two main requirements:

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

Under the guidance of these requirements, we have conducted two case studies to assess the suitability of the Schema.org vocabulary to support the metadata aggregation approach in CH. The case studies were also guided by the existing aggregation network of Europeana, from where we identify more detailed requirements for data modeling in real metadata aggregation scenarios.

Europeana provides access to digitized cultural resources from a wide range of cultural heritage institutions across Europe, mostly including libraries, museums and archives. It seeks to enable users to search and access knowledge in all the languages of Europe.

[1] https://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/

[2] https://www.w3.org/TR/microdata/

[3] https://www.w3.org/TR/json-ld/

[4] http://schema.org/docs/about.html

In the Europeana aggregation network, EDM (Europeana Data Model [8]) is the data model that supports data sharing efforts. It allows Europeana to be 'a big aggregation of digital representations of culture artefacts together with rich contextualization data and embedded in a linked Open Data architecture [9].

EDM also has a key role in many other parts of the Europeana Network[5]. EDM has been a collaborative effort from the very start, involving representatives from all the domains represented in Europeana: libraries, museums, archives, and galleries. It was initially defined in 2010 and has been under continuous improvement as a collaborative and community based activity, under the coordination and maintenance of Europeana. EDM supports several of the core processes of Europeana's operations, and contributes to the access layer of the Europeana Platform[6], where it supports the data reuse by third parties [10].

Another important aspect of EDM, is that it does not impose any constraint in the choice of Web technologies for metadata exchange. This comes from EDM following the principles of the Web of Data [11], and that it can be serialized in XML and RDF syntaxes (i.e., N-Triples[7], Turtle[8], JSON-LD, etc.). This aspect gives the Europeana Network much choice for technological innovation of the aggregation network, including the possibility to use other data models to aggregate metadata according to the requirements of its own data.

Other organizations using approaches for aggregation similar to that of Europeana also apply EDM. In our work, we have explored the Digital Public Library of America[9] (DPLA), which operates within the scope of the United States of America, and uses a model heavily based on EDM for its aggregation process [12].

Given our purpose of evaluating the suitability of Schema.org metadata for fulfilling the metadata aggregation in the specific domain of CH, we conducted case studies where we collected and analyzed Schema.org metadata from real collections and systems from CH institutions. We used cases of institutions that publish Schema.org metadata to make their resources better discoverable on the Web, and not for CH aggregation, therefore, the extent to which this data could fulfill the requirements of CH aggregation was unknown at the start of our work. As a platform for evaluation, we performed our analysis of Schema.org CH metadata according to the specific requirements of the Europeana Network [13].

The main premise behind our study is the following:

> *If Schema.org metadata can express the information requirements of the Europeana Data Model and the main factors for data quality defined by Europeana, then Schema.org may be used to fulfil the requirements of many cultural heritage services that are based on a metadata aggregation approach. Conversely, if using Schema.org data as source of data for Europeana is impossible or would require specific efforts, then the same obstacles will probably hold in other CH contexts.*

3 THE BASIC PRINCIPLES FOR APPLYING SCHEMA.ORG TO THE CULTURAL HERITAGE DOMAIN

Schema.org covers the data modeling needs of a wide range of domains. Its level of development varies across domains, however. This section provides a description of how it covers the representation of metadata about CH resources, the related technologies required for its processing, and the main requirements for its usage in metadata aggregation.

3.1 (Digital) Cultural heritage objects represented in Schema.org

Schema.org comes with a vocabulary that allows the description of entities of different types with subclasses, as well as attributes and relationships between entities, following the Semantic Web principles [14]. For CH digital libraries, Schema.org allows the description of books, maps, visual art, music recordings, and many other kinds of cultural resources.

[5] The Europeana Network is a community of experts working in the field of digital heritage, united by a shared mission to expand and improve access to Europe's digital cultural heritage. It has over 1,700 members who work for other organizations and share an interest in digital cultural heritage. Their contribution to improve Europeana sits within their work in their own organizations, and their experience and expertise help shape Europeana's work and services.

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

[7] https://www.w3.org/TR/n-triples/

[8] https://www.w3.org/TR/turtle/

[9] http://dp.la/

The most relevant Schema.org classes to Europeana are *schema:CreativeWork*[10] and several of its refining subclasses, which we detail here with their connection to the main modeling constructs of EDM:

- several types of *schema:CreativeWork*, such as *schema:VisualArtwork*, *schema:Book*, *schema:Painting*, *schema:Sculpture*, and *schema:Product*, can be matched to EDM's Provided Cultural Object (CHO) *edm:ProvidedCHO*, which represents the object that Cultural Heritage institutions are providing a metadata description about (and a digitized representation of). Each of these subclasses may be used with more specific properties than the ones available for *schema:CreativeWork* such as *schema:artMedium*[11] for *schema:VisualArtwork.*
- the subclass *schema:MediaObject* and its subclasses *schema:ImageObject*, *schema:VideoObject*, *schema:AudioObject* can be matched to what EDM defines as the *edm:WebResource*, which represents a digital version (representation) of the CHO.
- The schema:Person, schema:Place and schema:Organization classes match the semantics of EDM contextual classes edm:Agent, edm:Place and foaf:Organization.

Schema.org can also be extended to cover cases requiring properties or terms currently not available in the model. These extensions are either approved as part of the core Schema.org or are managed externally. Two of the existing extensions are of relevance to the CH domain:

- The Bibliographic Extension[12] provides additional properties and types to describe bibliographic resources. For example, terms such as 'atlas', 'newspaper', 'work and translation', or relationships such as *schema:exampleOfWork* and *schema:workExample*[13].
- The Architypes extension[14] currently works on identifying relevant types and properties to describe archives and their contents. The current proposal[15] defines three new classes: *Archive*, *ArchiveCollection* and *ArchiveItem.*

Schema.org is a collaborative and community based activity and its main platform of collaboration is the W3C Schema.org Community Group[16]. The Community Group also serves as a hub for discussion with other related communities, at W3C and elsewhere. E.g., other W3C Community Groups exist that are focused on specific domains, such as health, sports, bibliography, etc. Representatives of the Europeana community may be involved this way, should a need to 'improve' Schema.org for CH aggregation be raised.

3.2 Aggregation mechanisms for Schema.org metadata

Originally, indexing of web pages is the main use case for the development of Schema.org, therefore, it is mostly found encoded within (or referred from) HTML pages. So, the mechanism to aggregate Schema.org data can start the same way as for crawling ordinary web pages. The remainder of the aggregation can also rely on a process comparable to the one for ordinary web pages, which is based on following the hyperlinks within the HTML. Schema.org has been developed according to the Semantic Web and Linked Data principles: whatever encoding used for Schema.org data (RDFa, Microdata or JSON-LD), Schema.org data always consists of an RDF graph. Therefore, applications only interested in the Schema.org data, and not on the (HTML) textual content, can crawl the web pages in the same way as search engines, but simply discard the textual content, extract hyperlinks from the HTML, links from the Schema.org RDF graph, and continue the crawling by following (a selection of) these links.

Webmasters may aid the web crawling process (both for "regular" HTML pages and Schema.org-enabled ones) by providing Sitemaps[17] of their website. These inform search engines about which of the website URLs are available for crawling and some additional information, such as update frequency and importance within the website, that will enable

[10] http://schema.org/CreativeWork
[11] http://schema.org/artMedium
[12] http://bib.0.3-2f.schemaorgae.appspot.com/
[13] http://blog.schema.org/2014/09/schemaorg-support-for-bibliographic_2.html
[14] https://www.w3.org/community/architypes/
[15] https://www.w3.org/community/architypes/wiki/Initial_model_proposal
[16] http://www.w3.org/community/schemaorg
[17] http://www.sitemaps.org/

the website to be crawled more effectively. In the case of digital library websites, Sitemaps help dealing with some typical discovery problems faced by CH institutions:

- They enable web crawlers to reach areas of the website that are not available through the browsable interface. For example, the CH objects' landing pages that are only reachable through searching via web forms.
- Often, CH digital libraries contain a very large number of objects, which varies in time as collections grow, and there are chances that the web crawlers will overlook some of the new or recently updated content.

The combination of Schema.org and Sitemaps is also used in customized indexing services provided by search engines, such as the Google's Custom Search Engine[18].

We attempted to identify cases of usage of Schema.org metadata in CH institutions from the Europeana Network, Digital Public Library of America, and other communities. In those we identified the use of the following encodings and mechanisms:

- Schema.org metadata encoded in HTML pages with JSON-LD and/or Microdata:
    - University of Illinois at Champaign in the context of the Linked Data for Special Collections (LD4SC) project[19],
    - North Carolina State University Libraries[20]
    - WorldCat[21],
    - data.bnf.fr[22]
- Publication of linked data using Schema.org as the main vocabulary and HTTP content negotiation[23]:
    - WorldCat,
    - Research and Education Space[24]
- Schema.org referenced from IIIF[25] services, within the presentation layer of the framework (i.e. IIIF Manifests[26] using a seeAlso property[27]):
    - NCSU

4 CASE STUDIES

To assess the suitability of Schema.org for carrying the necessary information for supporting the requirements of metadata aggregation in CH, we searched for existing cases of usage of Schema.org in CH contexts. Among those that were identified, we focused on the ones most relevant to the metadata acquisition scenario of Europeana, i.e. potential data sources that would allow us to do a comparison between Schema.org metadata and EDM metadata about the same CH objects and derived from the same source.

4.1 Experimental setup

We have analyzed data from two data providers from DPLA, which as mentioned earlier uses a model heavily based on EDM for its aggregation process: North Carolina State University Libraries (NCSU) and University of Illinois at Champaign (UI). Both institutions use digital library management systems based on other metadata standards than EDM or Schema.org. The representation of the CH objects from these providers in EDM and Schema.org is always derived from

[18] https://cse.google.com/cse/
[19] http://publish.illinois.edu/linkedspcollections/
[20] https://www.lib.ncsu.edu/
[21] http://www.worldcat.org/
[22] http://data.bnf.fr/
[23] https://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html
[24] https://bbcarchdev.github.io/res/
[25] The International Image Interoperability Framework, commonly known as IIIF [15], is a family of specifications that were conceived to facilitate systematic reuse of image resources in digital image repositories maintained by cultural heritage organizations. It specifies several HTTP based web services covering access to images, the presentation and structure of complex digital objects, composed of one or more images, and searching within their content.
[26] http://iiif.io/api/presentation/2.1/#manifest
[27] http://iiif.io/api/presentation/2.1/#seealso

the original metadata standard. The EDM metadata is created for the purposes of DPLA aggregation, and Schema.org is created for Internet discovery[28].

The activity of these data providers in working with both data models offered us a very suitable scenario to assess Schema.org data. Our basic idea is to combine the Schema.org available for these cases with a new iteration (actually, an inversion) of an available mapping from EDM to Schema.org [7]. This setting allows us to compare EDM metadata resulting from two different data conversion paths: the EDM metadata available in DPLA, and EDM metadata obtained after the application of the new mapping of Schema.org to EDM. This experimental setting is illustrated in the figure below.

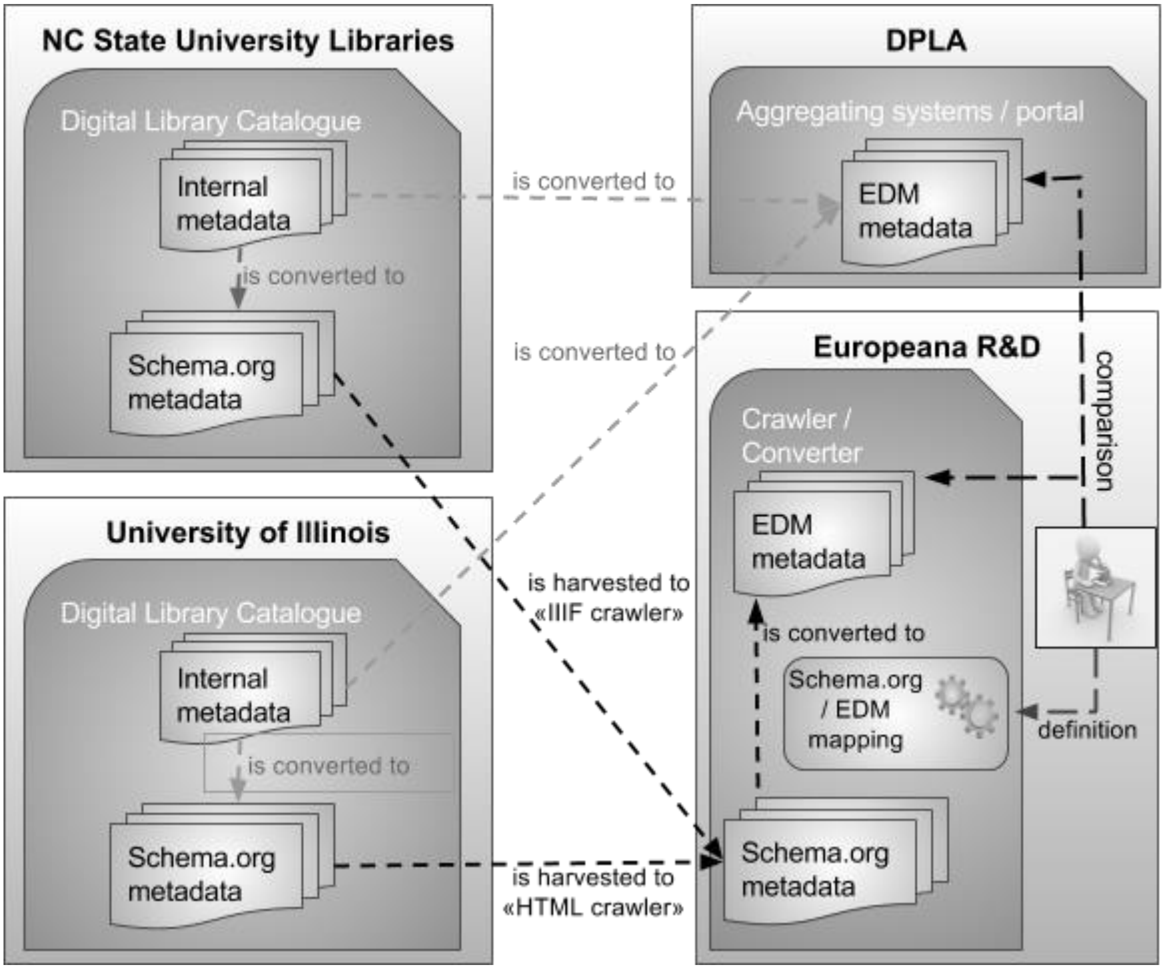

Fig. 1. **The experimental process. Black lines indicate the steps that were performed for the experiment.**

For each of the data providers, a subset of their digital library catalogue was selected. Selection of digital objects was made according to the following criteria:

- The object was part of a collection, for which Schema.org metadata was available.
- The object had been aggregated by DPLA already (and therefore EDM metadata was available).

[28] The University of Illinois at Champaign (UI) worked on a specific Schema.org mapping in the context of the Linked Data for Special Collections (LD4SC) project

For the selected digital objects, Schema.org metadata was obtained from both providers with the appropriate crawling mechanisms for their systems. Crawling of a IIIF service, for NCSU Libraries, and web page crawling via a Sitemap, for UI. Most of the crawling software required for the case studies was reused from our past work on data acquisition mechanisms [16]. We only needed to develop one software component to allow the parsing of data encoded within HTML pages[29].

The EDM metadata for the same set of objects was obtained from the DPLA downloadable data dump[30].

We initially collected a sample of 905 Schema.org objects from the UI and 1000 objects from NCSU Libraries, on which we did a preliminary analysis and data profiling to support the definition of the mapping. A listing was made of all Schema.org classes and properties used by both data providers (these listings can be consulted in appendices A and B).

For the following steps of the experiment, we focused our analysis on a maximum of 100 digital objects selected from each provider and resulting in a sample of metadata for 193 objects[31] (with each 193 Schema.org records and 193 DPLA EDM records).

The definition of the Schema.org to EDM mapping used, as a starting point, Europeana's earlier work [7] in mapping from the opposite direction, where EDM data was the input. The mapping was further refined based on the documentation of the Schema.org profile defined by the LD4SC project[32].

For this study, since our intent was to evaluate Schema.org created by data providers, the mappings to EDM did not include any classes or properties from Europeana's internal EDM profile[33] - we focused instead on the EDM expected from providers within Europeana's regular aggregation flow [17].

A software implementation of the mapping was done[34]. It was then applied to the sample dataset, the resulting EDM metadata was analysed, and a second revision of the mapping was done, before we performed the final analysis on the resulting EDM metadata (described in the next section). The full listing of the EDM classes and properties generated after the application of the mapping is available in Annex C.

5 ANALYSIS AND DISCUSSION

The metadata obtained after conversion from Schema.org to EDM was analyzed from two points of view:

- Analysis of the mapping from Schema.org to EDM that we defined and redefined, in terms of results achieved, potential for application in Europeana, and possibilities for improvement.
- Analysis of the obtained data using our experimental aggregation setup.

This section presents and discusses the outcomes of our analysis on these two aspects.

5.1 Results of the Schema.org to EDM mapping

We highlight three key aspects from our experience with mapping Schema.org data to EDM: requirements of the EDM model for aggregation [8, 17], overall data quality and comparative loss of information.

5.1.1 Basic EDM requirements:

Our first assessment consisted in making sure the different entities, defined in the Schema.org model, were matched with their corresponding EDM classes. *schema:CreativeWork* and its subclasses were mapped to a *edm:ProvidedCHO* and *schema:MediaObject* and its subclasses mapped to *edm:WebResource*. Only the *ore:Aggregation* class required in EDM – a rather artificial construct associating the Provided CHO to Web Resources and administrative metadata specific to the aggregation process, such as the data provider (see below) – had to be created as part of the conversion of the data.

EDM also requires the presence of some properties to ensure a basic level of data quality, as for example, at least one title, or alternatively a description. In the context of this analysis, all the mandatory properties required by EDM can be

[29] For this purpose, we have re-used an open source tool: Apache Any23, https://any23.apache.org/

[30] Digital Public Library of America website, bulk download section: https://dp.la/info/developers/download/. We used the dump from June 2017.

[31] 7 objects ended up missing in the University of Illinois dataset because the Schema.org data (on which the selection of 100 objects was made) and the DPLA one had been produced based on different versions of the dataset.

[32] http://publish.illinois.edu/linkedspcollections/methods-outcomes/

[33] https://github.com/europeana/corelib/wiki/EDMObjectTemplatesEuropeana

[34] https://github.com/nfreire/Open-Data-Acquisition-Framework

mapped from the Schema.org metadata, except for *edm:type* and *edm:rights,* which are defined by Europeana under controlled vocabularies for enabling functionalities of its portal or its licensing policies.

The case of *edm:type* may be overcome by inferencing the correct value from the specific Schema.org class mapped to the *edm:ProvidedCHO.* However this won't be possible for the data only defined as *schema:CreativeWork*.

Regarding *edm:rights*, in the metadata analyzed for this experiment we could not find data (literal or URI) that could be used for *edm:rights*. Schema.org does contain properties (such as *schema:license*) that can be suitable vehicles for the values requested by EDM. However, the values are not compatible with the ones that EDM expects (i.e., rights statements from Creative Commons and RightsStatements.org[35]). In the future, providers expecting Europeana to harvest their metadata via Schema.org would be expected to use these statements in an agreed Schema.org property. Since interoperable values like these are expected and designed to be beneficial to other actors than Europeana, there should probably be more efforts to encourage their use in the wider Schema.org context nevertheless.

EDM requires also aggregation specific metadata such as the name of the original data provider (*edm:dataProvider*) that is typically a CHI that holds the digital resources, and the direct provider (*edm:provider*) of the data to Europeana that typically is an organization fulfilling the role of data aggregator within the Europeana Network.. In the context of this experiment, we could find the relevant information to map a *schema:Organization* to *edm:dataProvider* and *edm:provider*. However, the absence of those elements would make metadata records invalid for Europeana aggregation. Also, for other datasets, the Schema.org elements we mapped from (*schema:provider* and *schema:copyrightHolder*) may be used in a slightly different way. In a Europeana aggregation scenario based on Schema.org data, data providers would be expected to provide the appropriate information required by Europeana aggregation process.

5.1.2 Overall data quality: getting rich data in EDM

The work on mapping from EDM to Schema.org that we started from had highlighted some limitations regarding the granularity of the mapping for the main CHO entity. *schema:CreativeWork* matches well the semantics of *edm:ProvidedCHO*, but it would be better to use its subclasses when possible, such as *schema:Book* or *schema:Newspapers.* Yet the initial mapping could not use these subclasses as the type would need to be deduced from the *dc:type* element in the EDM data, which is often not normalized. Unless Europeana data providers used controlled vocabularies in *dc:type* it is very difficult to define the precise type of a CHO in EDM. The same issue applies for the mapping between *edm:WebResources* and the subclasses of *schema:MediaObject.*

This issue does not happen in a mapping from Schema.org to EDM as the subclasses of *schema:CreativeWork* can directly be mapped onto *dc:type* and the URIs be preserved. A mapping in this direction does not cause any information loss during data conversion.

In general, obtaining valid EDM data after a mapping, requires all the mandatory elements in EDM to be present in the mapped Schema.org data. This was not the case for this experiment, where a few data elements were not present (see section 5.1.1). The missing elements were very closely related to usage of Europeana, however, and these data providers prepared their Schema.org metadata with Internet discovery in mind, and not specifically for a CH aggregation scenario such as Europeana.

5.1.3 Granularity mismatches between EDM and Schema.org and resulting information loss

Schema.org covers an extensive range of entity types that can be used in the description of all the "things" mentioned in the metadata as entities with their own URIs. It is much more comprehensive than EDM. This results in entity descriptions that are more granular than the ones in EDM, e.g., using *schema:Collection* for a collection level description, *schema:Distance* to specify the type of values available as height and width of the *schema:CreativeWork*. Some of those entities cannot be directly mapped in EDM since a corresponding EDM entity (or a relevant superclass) does not exist.

In addition, one of the data providers in our experiment extended its Schema.org profile with properties from its own namespace[36]. Considering this information would have required an additional mapping; which we decided not to do. This experiment has been done in the context of investigations of alternative data acquisitions for Europeana. While mapping

[35] http://pro.europeana.eu/available-rights-statements

[36] http://ns.library.illinois.edu/scp/

data to EDM from standard Schema.org data is an endeavor Europeana can afford, taking into account any extensions of it would not be a sustainable approach.

To address the absence of corresponding EDM classes for a given Schema.org or custom one, we could decide to map the data at a more generic level. While this mapping would not retain the exact original semantics, it would retain all the data values. For instance UI uses a custom class *scp:StageWork, which is a* subclass of *schema:CreativeWork*. It does not have an exact equivalent in EDM but could be mapped to *edm:ProvidedCHO* as explained earlier.

Another approach would be to describe these resources as "contextual works", for which the current approach in EDM would be to use *skos:Concept*. This is actually similar to another case we encountered, where the granularity mismatch was in the opposite direction: NCSU had used over one thousand instances of *schema:Thing* as objects for the *schema:about* property. We made the decision to map them to *skos:Concept* in a way that is quite collection-specific and rather bold: *schema:Thing* is not equivalent to *skos:Concept* in absolute.

Again, such mappings allow to keep data resources in, which the Schema.org data show to be very useful to contextualize digital objects. They however require harder work, and the creators of the data may find it cumbersome to handle the semantic gap between the original type and the notion of 'concept'. Both obstacles may explain why the data could not be found in the EDM for UI in DPLA, either.

5.2 Analysis of data obtained in the experimental Schema.org-based aggregation setup

Both datasets used in this experiment allowed us to explore the potential of Schema.org data for data aggregation into Europeana.

The mapping from source data to EDM done in the process of the DPLA aggregation enables the (re)structuring of source data according to the main information entities defined by the EDM model: each resource (CHO, digital object, collection, additional contextual resources) is identified by its own URI and gathers all the data referring to a particular entity. But the new efforts – by providers – for mapping this source data to Schema.org has shown that the data can be even better structured. Schema.org is indeed slightly more granular in terms of classes and properties, thus provides more motivation for data publishers to make a better mapping, and even enrich the data they have (a phenomenon that is already happening within the boundaries of EDM but can do with more encouragement[37]). In the particular cases we analyzed, we have especially noticed that the Schema.org data have been further enriched with links to external resources such as controlled vocabularies (Library of Congress Subjects Headings[38] (LCSH), VIAF[39] and Art and Architecture thesaurus[40] (AAT)) and related resources (Wikipedia[41], WorldCat, IMDB[42]).

In a way and like the class mappings discussed earlier in this section, a mapping from Schema.org to EDM can retain these enrichments as long as a suitable property for representing them is found. Most of these links are indeed a way to get richer, authoritative data from external sources - for example the UI dataset defines no name for the persons it uses, but all are in fact VIAF resources (URIs) that are provided with names via the OCLC service. UI's efforts can instead be focused on defining 'job titles' (e.g. roles of actors in plays) in the Schema.org data for these persons, a key characteristic of their collection.

Note that EDM includes very generic properties such as *edm:hasType*[43] and *edm:isRelatedTo*, which can be used as fallback option in case EDM has no property that would keep the semantics of the Schema.org one being used for the enrichment. Although this approach results in semantic loss of the data, it would help to progressively improve the granularity of data in Europeana, nonetheless.

[37] https://pro.europeana.eu/page/europeana-semantic-enrichment
[38] http://id.loc.gov/authorities/subjects.html
[39] https://viaf.org/
[40] http://www.getty.edu/research/tools/vocabularies/aat/
[41] https://en.wikipedia.org/wiki/Main_Page
[42] http://www.imdb.com/
[43] Note that edm:hasType is a different property from edm:type discussed earlier.

6 CONCLUSIONS

The experiments we have reported in this paper show that Schema.org metadata are suitable for metadata aggregation in CH contexts, considering that Europeana as a typical example of such aggregation. Mapping from Schema,org to the Europeana Data Model retains most of the semantics of the Schema.org classes and can take advantage of the enrichments promoted by the Schema,org model.

The data providers in the study prepared Schema.org metadata for Internet discovery, not specifically for CH applications. Despite this fact, the EDM metadata derived from Schema.org has been found to be very close to being fully suitable for aggregation by CH services such as Europeana.

In the concrete case of the Europeana Network, there are still some issues with employing Schema.org metadata acquisition as a direct replacement to the current Europeana metadata aggregation workflow. EDM defines some properties with specific semantics for Europeana aggregation, or with controlled vocabularies for which Schema.org provides no suitable solution by itself. In our case studies we identified several properties that would require particular attention at mapping time: *edm:rights*, *edm:type*, *edm:dataProvider* and *edm:provider*. Data providers expecting Europeana to harvest their metadata via Schema.org should provide the required information in agreed Schema.org properties.

To conclude, Schema.org poses no obstacle that cannot be overcome to allow data providers to deliver metadata in full compliance with EDM requirements and with the desired semantic quality. In fact, the semantic granularity of Schema.org, and the motivations Schema.org brings to data owners for publishing more data on the web, presents an opportunity for progressively improving the granularity of EDM data at Europeana, in a sustainable, low effort way. To ensure it however, Schema.org support in some CH services, such as Europeana, must be accompanied by recommendations and/or specifications regarding how data providers should provide their Schema.org metadata.

ACKNOWLEDGMENTS

We would like to acknowledge the support given by staff members of North Carolina State University Libraries, the University of Illinois at Champaign and the Digital Public Library of America, for their support in access and analysis of the data sources for the case studies: Jason Ronallo, Timothy Cole, Jacob Jett, Gretchen Gueguen and Michael Della Bitta.
This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by the European Commission under the Connecting Europe Facility, telecommunications sector, grant agree-ment number CEF-TC-2015-1-01.

A CLASS USAGE IN THE SCHEMA.ORG SOURCE DATA AND DERIVED EDM METADATA

This appendix contains the class usage profile of the Schema.org metadata from the initially collected sample of 1000 Schema.org records from NCSU Libraries and 905 records from the University of Illinois. The two tables below present all Schema.org classes used by either data provider (table on the left side) and the class usage profile of the EDM derived from the Schema.org metadata using our mapping (table on the right side).

Table 1. **Class usage in the Schema.org source data and derived EDM metadata**

Usage count in Schema.org			Usage count in EDM generated from Schema.org		
Class URI	NCSU	U. Illinois	Class URI	NCSU	U.Illinois
schema:CreativeWork	1000	1868	edm:ProvidedCHO	1000	3474
schema:Book	0	768			
schema:VisualArtwork	0	847			
			ore:Aggregation	1000	905
schema:AudioObject	7	0	edm:WebResource	7	1481
schema:ImageObject	0	1337			
schema:WebPage	0	160			
schema:VideoObject	25	0			
schema:Organization	388	1810	foaf:Organization	388	1804
schema:Person	324	2193	edm:Agent	324	2187
schema:Place	690	0	edm:Place	690	0
schema:PostalAddress	410	0			
schema:Distance	0	1298			
schema:GeoCoordinates	690	0			
schema:Thing	2733	0	skos:Concept	2733	0
rdfs:Resource	0	906			

B DATA PROFILE OF THE SCHEMA.ORG

This appendix contains the data profile of the Schema.org from the initially collected sample of 1000 Schema.org records from NCSU Libraries and 905 records from the University of Illinois.

The tables below present all Schema.org classes and properties used by either data provider. Each table presents the property usage within one class, indicating the number of times the properties where used having an instance the respective class as its subject and as its object.

Table 2. **Property usage in schema:AudioObject**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:contentUrl	7	0
schema:description	1	0
schema:encodingFormat	7	0
schema:name	7	0
schema:uploadDate	7	0
Usage as object		
schema:audio	7	0

Table 3. **Property usage in schema:Book**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:author	0	812
Usage as object		
schema:example	0	768

Table 4. **Property usage in schema:CreativeWork**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:about	3188	197
schema:additionalType	0	1810
schema:associatedMedia	0	179
schema:audio	7	0
schema:contentLocation	690	0
schema:contributor	32	1485
schema:copyrightHolder	0	58
schema:creator	231	58
schema:dateCreated	746	695
schema:description	206	40
schema:exampleOfWork	0	768

schema:genre	0	41
schema:height	0	22
schema:image	968	0
schema:inLanguage	591	0
schema:isPartOf	1200	116
schema:keywords	2710	0
schema:locationCreated	0	898
schema:mainEntityOfPage	0	160
schema:name	1000	962
schema:provider	0	58
schema:sameAs	0	1052
schema:thumbnailUrl	968	0
schema:url	1000	0
schema:video	25	0
schema:width	0	22
Usage as object		
schema:isPartOf	0	1810
schema:sameAs	0	8
schema:url	1000	0

Table 5. **Property usage in schema:Distance**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:name	0	1298
Usage as object		
schema:height	0	649
schema:width	0	649

Table 6. **Property usage in schema:GeoCoordinates**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:latitude	399	0
schema:longitude	399	0
Usage as object		
schema:geo	690	0

Table 7. **Property usage in schema:ImageObject**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:contentUrl	0	1337
schema:fileFormat	0	1337
schema:height	0	1309
schema:name	0	703
schema:width	0	1309
Usage as object		
schema:associatedMedia	0	1337

Table 8. **Property usage in schema:Organization**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:description	82	0
schema:name	388	1810
schema:sameAs	0	905
Usage as object		
schema:about	341	0
schema:contributor	31	0
schema:copyrightHolder	0	905
schema:creator	16	905
schema:provider	0	905

Table 9. **Property usage in schema:Person**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:birthDate	210	290
schema:deathDate	180	226
schema:description	83	0
schema:familyName	292	0
schema:gender	102	0
schema:givenName	286	0
schema:jobTitle	0	1486
schema:name	324	569
schema:sameAs	0	3234
Usage as object		
schema:about	108	0
schema:author	0	708

schema:contributor	1	1485
schema:creator	215	0
schema:sameAs	0	11

Table 10. **Property usage in schema:Place**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:address	410	0
schema:alternateName	410	0
schema:description	93	0
schema:geo	690	0
schema:name	690	0
Usage as object		
schema:contentLocation	690	0

Table 11. **Property usage in schema:PostalAddress**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:addressLocality	113	0
schema:addressRegion	406	0
schema:postalCode	98	0
Usage as object		
schema:address	410	0

Table 12. **Property usage in schema:Thing**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:name	2733	0
Usage as object		
schema:about	2733	0

Table 13. **Property usage in schema:VideoObject**

Property URI	*NCSU*	*U. Illinois*

Usage as subject		
schema:contentUrl	*25*	*0*
schema:description	*2*	*0*
schema:duration	*25*	*0*
schema:encodingFormat	*25*	*0*
schema:height	*25*	*0*
schema:name	*25*	*0*
schema:thumbnailUrl	*25*	*0*
schema:uploadDate	*25*	*0*
schema:width	*25*	*0*
Usage as object		
schema:video	*25*	*0*

Table 14. **Property usage in schema:VisualArtwork**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:about	0	4030
schema:artMedium	0	1445
schema:artform	0	840
schema:artworkSurface	0	222
schema:associatedMedia	0	1158
schema:copyrightHolder	0	847
schema:creator	0	847
schema:description	0	96
schema:genre	0	847
schema:height	0	627
schema:isPartOf	0	1694
schema:name	0	847
schema:provider	0	847
schema:sameAs	0	1693
schema:text	0	9
schema:width	0	627
Usage as object		
(never used as object)		

Table 15. **Property usage in schema:WebPage**

Property URI	NCSU	U. Illinois
Usage as subject		
(never used as subject)		

Usage as object		
schema:mainEntityOfPage	0	160

Table 16. **Property usage in rdf:Resource**

Property URI	NCSU	U. Illinois
Usage as subject		
schema:name	6	491
schema:sameAs	0	3336
Usage as object		
schema:about	6	0
schema:author	0	104
schema:locationCreated	0	802
schema:sameAs	0	294

C DATA PROFILE OF THE EDM METADATA DERIVED FROM SCHEMA.ORG

This appendix presents the data profile of the EDM derived from the Schema.org metadata (whose profile was presented in Appendix B) using the Schema.org to EDM mapping implemented for the case studies.

The tables below present all Europeana Data Model (EDM) classes and properties used by either data provider. EDM reuses classes and properties from other namespaces, therefore the listing contains information about data elements of other namespaces as well. Each table presents the property usage within one EDM class, indicating the number of times the properties where used having an instance the respective class as its subject and as its object.

Table 17. **Property usage in edm:Agent**

Property URI	NCSU	U. Illinois
Usage as subject		
rdagr2:birthDate	210	0
rdagr2:deathDate	180	0
rdagr2:professionOrOccupation	0	1482
owl:sameAs	0	3225
skos:prefLabel	610	568
Usage as object		
dc:contributor	1	1481
dc:creator	215	706
dc:subject	108	0
owl:sameAs	0	11

Table 18. **Property usage in edm:Place**

Property URI	NCSU	U. Illinois
Usage as subject		
wgs84pos:lat	399	0
wgs84pos:long	399	0
skos:altLabel	410	0
skos:prefLabel	690	0
Usage as object		
dc:subject	690	0

Table 19. **Property usage in edm:ProvidedCHO**

Property URI	NCSU	U. Illinois
Usage as subject		
dc:contributor	32	1481
dc:creator	231	1608
dc:description	206	2390

dc:language	591	0
dc:subject	6582	4215
dc:title	1000	1803
dc:type	0	1723
dcterms:created	746	692
dcterms:extent	0	1291
dcterms:isPartOf	1200	1804
dcterms:medium	0	221
dcterms:rights	0	902
dcterms:tableOfContents	0	159
edm:hasType	1000	5281
edm:realizes	0	765
owl:sameAs	0	2737
Usage as object		
dcterms:isPartOf	0	1804
edm:aggregatedCHO	1000	905
edm:isShownAt	1000[44]	0
edm:realizes	0	765
owl:sameAs	0	8

Table 20. **Property usage in edm:WebResource**

Property URI	NCSU	U. Illinois
Usage as subject		
dc:description	8	692
dc:format	7	1322
dc:type	7	1481
dcterms:extent	0	2588
Usage as object		
dcterms:tableOfContents	0	159
edm:isShownBy	7	1322

Table 21. **Property usage in ore:Aggregation**

Property URI	NCSU	U. Illinois
Usage as subject		

[44] In the Schema.org data from NCSU Libraries, the same URI is used to identify the schema:MediaObject instance (or WebResource, in EDM terms) and the CreativeWork instance (or ProvidedCHO, in EDM terms), as a result after conversion, the edm:isShownAt property refers to both the ProvidedCHO and the WebResource. In the context of Europeana, we follow the recommendation in [6]: 'This URI (of the ProvidedCHO) should resolve to a description of the resource'.

edm:aggregatedCHO	1000	905
edm:dataProvider	1000	905
edm:isShownAt	1000	0
edm:isShownBy	975	1322
edm:object	968	0
edm:provider	1000	905
Usage as object		
(never used as object)		

Table 22. **Property usage in skos:Concept**

Property URI	NCSU	U. Illinois
Usage as subject		
skos:prefLabel	2733	0
Usage as object		
dc:subject	2733	0

Table 23. **Property usage in foaf:Organization**

Property URI	NCSU	U. Illinois
Usage as subject		
owl:sameAs	0	902
skos:note	82	0
skos:prefLabel	388	1804
Usage as object		
dc:contributor	31	0
dc:creator	16	902
dc:subject	341	0
dcterms:rights	0	902
edm:provider	0	902

REFERENCES

[1] Harry Verwayen. 2017. Business Plan 2017: 'Spreading the Word'. Europeana Foundation. <https://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-business-plan-2017.pdf>

[2] Henning Scholz, Douglas McCarthy, Pablo Uceda Gomez, Evangelia Katrinaki, Kerstin Herlt, Julia Welter, Maria Teresa Natale, Marzia Piccininno, Gisela Baumann, Kate Fernie, Dimitris Gavrilis, Marco Rendina, Erwin Verbruggen, Gariella Ivacs, Nienke van Schaverbeke, Joseph Garvin. 2017. 'Amount of Data Partners and Outeach to Major Institutions', Deliverable 1.2 of the project Europeana Core Service Platform. <https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.2-amount-of-data-partners-and-outreach-to-major-institutions.pdf>

[3] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0 <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>.

[4] Dublin Core Metadata Initiative. 2012. Dublin Core Metadata Element Set, Version 1.1: Reference Description. DCMI Recommendation. <http://www.dublincore.org/documents/dces/>

[5] Google Inc., Yahoo Inc., Microsoft Corporation and Yandex. About Schema.org. <http://schema.org/docs/about.html>

[6] Nuno Freire, Hugo Manguinhas, Antoine Isaac, Glen Robson, John B. Howard. 2017. Web technologies: a survey of their applicability to metadata aggregation in cultural heritage. 21st International Conference on Electronic Publishing.

[7] Richard Wallis, Antoine Isaac, Valentine Charles, and Hugo Manguinhas. 2017. Recommendations for the application of Schema.org to aggregated Cultural Heritage metadata to increase relevance and visibility to search engines: the case of Europeana. Code4Lib Journal, Issue 36. ISSN 1940-5758. <http://journal.code4lib.org/articles/12330>

[8] Europeana Foundation. 2017. The EDM Definition V5.2.8. Europeana Foundation. <http://pro.europeana.eu/edm-documentation>

[9] Stefan Gradmann. 2015. Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. <http://pro.europeana.eu/publication/knowledgeinformation-in-context>

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

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

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

[13] Antoine Isaac, Robina Clayphan (eds.). 2013. Europeana Data Model Primer. Europeana Foundation. <https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf>

[14] Timothy Berners-Lee, James Hendler and Ora Lassila, 2001. The Semantic Web. Scientific American, Volume 284, Issue 5, p. 29-37.

[15] Stuart Snydman, Robert Sanderson and Tom Cramer. 2015. The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. Archiving 2015. <http://purl.stanford.edu/df650pk4327>

[16] Nuno Freire, Glen Robson, John B. Howard, Hugo Manguinhas, Antoine Isaac. 2017. Metadata Aggregation: Assessing the Application of IIIF and Sitemaps within Cultural Heritage. 21st International Conference on Theory and Practice in Digital Libraries.

[17] Europeana Foundation. 2017. Europeana Data Model - Mapping Guidelines V2.4. Europeana Foundation. <https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines _v2.4_102017.pdf>