Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata

Nuno Freire[1(✉)], Valentine Charles[2], and Antoine Isaac[2,3]

[1] INESC-ID, Lisbon, Portugal
nuno.freire@tecnico.ulisboa.pt
[2] Europeana Foundation, The Hague, The Netherlands
valentine.charles@europeana.eu
[3] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
aisaac@few.vu.nl

Abstract. 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. We used the requirements of the Europeana Network as evaluation criteria. These include the recommendations of the Europeana Data Model, which is 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, Schema.org's cross-domain applicability raises the need for accompanying its adoption by recommendations and/or specifications regarding how data providers should create their Schema.org metadata, so that they can meet the specific requirements of Europeana or other cultural aggregation networks.

Keywords: Metadata · Cultural heritage · Metadata aggregation
Schema.org · Europeana Data Model · Digital libraries

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.

A. Gangemi et al. (Eds.): ESWC 2018, LNCS 10843, pp. 225–239, 2018.
https://doi.org/10.1007/978-3-319-93417-4_15

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 digitized CH resources from and about Europe [1]. It seeks to enable users to search and access knowledge in all the languages of Europe via the Europeana Collections portal[1] and applications to use cultural data through open APIs. Although many European CH institutions do not yet have a presence in Europeana, it already holds metadata from over 3,700 providers, mostly libraries, museums and archives [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, CH 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 our work on another application of Schema.org: we seek to test whether it can also bring usable data sources for CH aggregators like Europeana. We therefore aim to investigate whether more incentive can be provided for the (still infrequent) use of Schema.org in the CH sector, beyond its original goal of publishing data for search engines.

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

[1] https://europeana.eu.

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[2], Microdata[3] and JSON-LD[4]) 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[5] 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 be a unified solution for allowing the discovery of their resources through both internet search engines and CH specific metadata aggregation efforts like 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.

In the Europeana aggregation process, the Europeana Data Model (EDM) [8] is the data model that 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 supports several of the core processes of Europeana's

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

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

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

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

operations and contributes to the access layer of the Europeana platform, supporting the sharing of data with third parties [10].

EDM has been a collaborative, community-based 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 since, under the coordination and maintenance of Europeana.

EDM also plays a key role in other parts of the Europeana Network[6] and elsewhere. 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[7] (DPLA), which operates within the USA and uses a model heavily based on EDM for its aggregation process [12].

An 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 various XML and RDF syntaxes (i.e., N-Triples[8], Turtle[9], JSON-LD, etc.). This flexibility gives the Europeana Network much choice for technological innovation of the aggregation network. The genericity of EDM's constructs (and of the systems built on top of them) also makes it easier to use other data models to aggregate metadata that, although not based on EDM, would match Europeana's information requirements.

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 Europeana [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 CH 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.

---

[6] The Europeana Network is a community of 1,700 experts with the shared mission to expand and improve access to Europe's digital cultural heritage, in the organization they work for and/or by contributing to shape Europeana's services.

[7] http://dp.la/.

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

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

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.

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 Heritage Object (CHO) *edm:ProvidedCHO*, which represents the object that CH institutions provide metadata about (and a digitized representation of). Each of these subclasses may be used with more specific properties than the ones available for *schema:CreativeWork,* e.g., *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 *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].

[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.

- 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. A process to aggregate Schema.org data can thus 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 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, some pages for CH objects may be 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.

---

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

We attempted to identify cases of usage of Schema.org metadata in CH institutions (libraries, archives, museums) from the Europeana Network, Digital Public Library of America, and other communities. In those cases, 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 Urbana-Champaign (UI) in the context of the Linked Data for Special Collections (LD4SC) project[19],
  - North Carolina State University Libraries (NCSU)[20]
  - OCLC's WorldCat[21],
  - data.bnf.fr[22] from the French National Library (BnF)
- Publication of linked data using Schema.org as the main vocabulary and HTTP content negotiation[23]:
  - OCLC's WorldCat,
  - BBC's Research and Education Space[24]
  - Schema.org referenced from IIIF[25] services, within the IIIF presentation layer (i.e. IIIF Manifests using a *seeAlso* property[26]):
  - North Carolina State University Libraries

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 selected, among the CH cases of Schema.org usage identified above, the most relevant ones to the metadata acquisition scenario of Europeana. I.e., the cases where:

- The Schema.org data directly derives from the source data at a CH institution (i.e. directly from a digital library catalogue).
- The Schema.org data is created by the institution owning the data, not by a third party, such as an aggregator.
- The same dataset is also available in EDM.

---

[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 (IIIF) is a family of specifications that facilitate publishing and reuse of image resources [15]. It specifies HTTP based web services covering access to images, the presentation and structure of complex digital objects, and searching within their content.

[26] http://iiif.io/api/presentation/2.1/#seealso.

This setting allowed us to do an unbiased comparison between Schema.org metadata and EDM metadata about the same CH objects and derived from the same source. Two datasets fulfill these requirements: these of NCSU and UI.

4.1 Experimental Setup

We have analyzed data from NCSU and UI. Both institutions use digital library management systems based on other metadata standards than EDM or Schema.org, from which the representation in EDM and Schema.org is always derived. The EDM metadata is created for the purposes of DPLA aggregation, and Schema.org is created for Internet discovery.

The activity of these data providers in working with both data models offered us a very suitable scenario to assess Schema.org data. Our idea is to combine the Schema.org available for these cases with a new iteration (actually, an inversion) of a mapping from EDM to Schema.org we have created earlier [7]. This allows us to compare EDM metadata resulting from two different data conversion paths: the EDM data available in DPLA, and EDM data obtained after the application of the new mapping of Schema.org to EDM. This experimental setting is illustrated in Fig. 1.

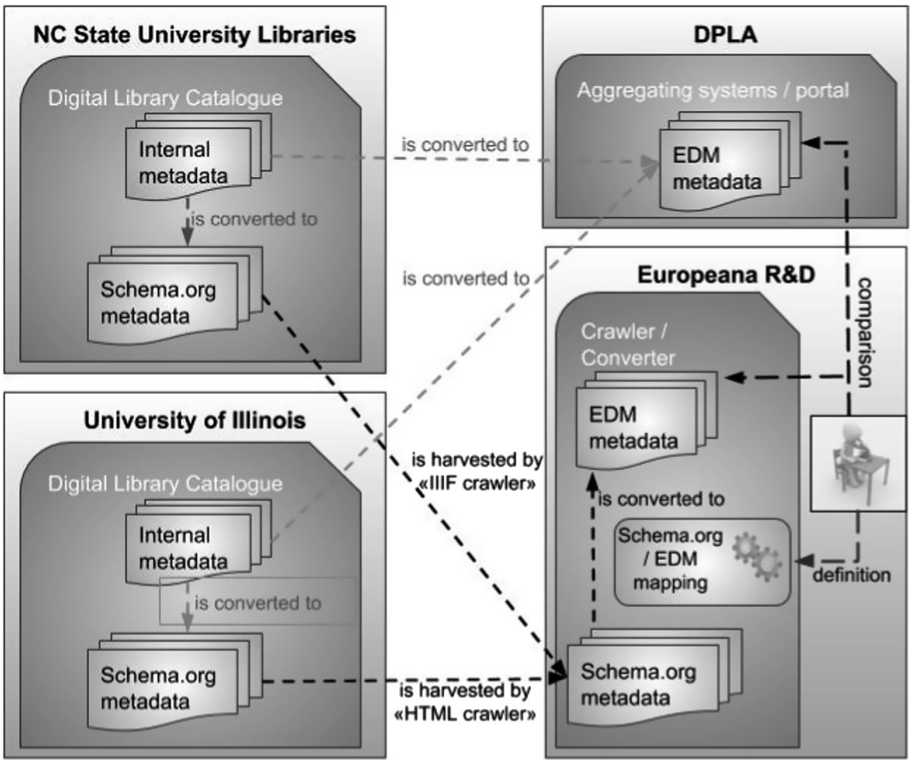

Fig. 1. The experimental process. Black lines indicate the steps performed for the experiment.

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

- being part of a collection, for which Schema.org metadata is available.
- being aggregated by DPLA already (so that EDM metadata is available).

For the selected objects, Schema.org metadata was obtained from both providers with the appropriate crawling mechanism for each case: 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[27].

The EDM metadata for the same set of objects was obtained from the DPLA downloadable data dump[28].

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[29]).

For the following steps of the experiment, we focused our analysis on a maximum of 100 digital objects selected from each provider, resulting in a sample of metadata for 193 objects[30] (each with one Schema.org record and one DPLA EDM record).

The definition of the Schema.org to EDM mapping used, as a starting point, Europeana's earlier work in mapping from the opposite direction, where EDM data was the input [7]. The mapping was further refined based on the documentation of the Schema.org profile defined by UI's LD4SC project[31].

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[32]. We focused instead on the EDM expected from providers within Europeana's regular aggregation flow [13].

A software implementation of the mapping was done[33]. It was then applied to the sample dataset, the resulting EDM metadata was analyzed, and a second revision of the mapping was done, before we performed the final analysis on the resulting EDM

[27] For this we have used the open source tool Apache Any23, https://any23.apache.org/.

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

[29] Available at https://github.com/nfreire/Open-Data-Acquisition-Framework/tree/master/opaf-casestudies/eswc2018-paper.

[30] 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.

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

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

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

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 appendix C (See footnote 29).

5 Analysis and Discussion

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

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

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

5.1 Results of the Mapping from Schema.org to EDM

We highlight three key aspects from our experience with mapping Schema.org data to EDM: requirements of the EDM model for aggregation [8, 13], overall data quality and comparative loss of information.

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 *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 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 *edm:ProvidedCHO*. However, this won't be possible for the data simply 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 this property. Schema.org does contain properties (such as *schema:license*) that can be suitable vehicles for the values requested by EDM. However, while Schema.org does not restrict the values for schema:license, EDM expects rights statements from Creative Commons and RightsStatements.org[34]. In the future, providers expecting Europeana to harvest their

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

metadata via Schema.org would be expected to use these statements in an agreed Schema.org property. Nevertheless, since interoperable values like these are beneficial to other actors than Europeana, there should probably be more efforts to encourage their use in the wider Schema.org context.

Europeana requires also aggregation-specific EDM 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 is typically an organization fulfilling the role of data (sub-)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, 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, for their records to be valid, data providers would be expected to provide the information mandatory for Europeana aggregation process. This was not the case for this experiment, as seen. The missing elements were very closely related to usage of Europeana, yet, and one must note the institutions we selected had prepared their Schema.org metadata with Internet discovery in mind, and not specifically for a CH aggregation scenario such as Europeana's. It could be relatively easy for them to add the required information.

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. The semantics of *schema:CreativeWork* match well the semantics of *edm:ProvidedCHO*, but it would be better to use its subclasses when possible, such as *schema:Book* or *schema:Newspaper*. 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:WebResource* 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 their URIs be preserved in the process. A mapping in this direction does not cause any information loss during data conversion.

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 (http://ns.library.illinois.edu/scp/). Considering this information would have required an additional mapping; which we decided not to do. Our experiment has been carried out in the context of investigations

of alternative data acquisitions for Europeana. While mapping 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. 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. While this mapping would not retain the exact original semantics, it would retain all the data values.

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, and 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[35]). 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 – the Library of Congress Subjects Headings[36] (LCSH), the Virtual International Authority File[37]

[35] https://pro.europeana.eu/page/europeana-semantic-enrichment.

[36] http://id.loc.gov/authorities/subjects.html.

[37] https://viaf.org/.

(VIAF) and the Art and Architecture Thesaurus[38] (AAT) – and other related resources –Wikipedia[39], WorldCat, IMDB[40].

In a way and like the class mappings discussed in Sect. 5.1, 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, which is a key characteristic of their collection.

Note that EDM includes very generic properties such as *edm:hasType*[41] and *edm:isRelatedTo*, which can be used as fallback options 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 loss of semantic grain, it would help to progressively improve the granularity of data in Europeana, nonetheless.

6 Conclusions

The experiments we have reported in this paper show that Schema.org is suitable for describing CH objects. The data providers in the study prepared Schema.org metadata for Internet discovery, not specifically for CH aggregation. In spite of this, the EDM metadata derived from Schema.org has been found to be close to being fully suitable for aggregation by Europeana.

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 and requires the use of controlled vocabularies for which Schema.org provides no suitable solution by itself. In our case studies, we have 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. Yet, we can conclude that 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.

We claim that these findings can be extended to CH aggregation services similar to Europeana, as EDM has been designed to meet very diverse CH cases and has indeed been re-used/extended in a number of CH data interchange scenarios[42]. In fact, Schema.org provides additional semantic granularity, which may allow the description

[38] http://www.getty.edu/research/tools/vocabularies/aat/.

[39] https://www.wikipedia.org/.

[40] http://www.imdb.com/.

[41] Note that edm:hasType is a different property from edm:type discussed earlier.

[42] https://pro.europeana.eu/page/edm-profiles.

of additional types and characteristics of CH objects found in specific metadata schemas but not (yet) implemented in EDM. For example, UI's theatre events can be represented using the class *schema:theaterEvent*. As it is backed – and consumed – by main web search engines, Schema.org brings further motivation to data owners for publishing richer data on the web.

With this combination of factors, we are convinced that Schema.org presents an opportunity for progressively improving the granularity of EDM data at Europeana and similar CH services, in a sustainable way. To ensure it, however, we must provide clear recommendations and/or specifications regarding how data providers should provide their Schema.org metadata.

Future work at Europeana will contribute to such support and try to motivate more CH institutions to publish Schema.org data. As a matter of fact, we plan to publish our complete dataset as Schema.org data in the second semester of 2018, based on our earlier work [7]. We will also investigate whether the handling of rights in Schema.org can be better aligned with CH practices. Europeana is involved in an initiative – RightsStatements.org – that provides a good framework for contributing to the Schema.org community-based extension process mentioned in Sect. 3, if appropriate.

Acknowledgements. We would like to acknowledge the support given by staff members of North Carolina State University Libraries, the University of Illinois at Urbana-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 agreement number CEF-TC-2015-1-01, and under contract number 30-CE-0885387/00-80.

References

1. Verwayen, H.: Business Plan 2017: Spreading the Word. Europeana Foundation (2017). https://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-business-plan-2017.pdf
2. Scholz, H., McCarthy, D., Gomez, P.U., Katrinaki, E., Herlt, K., Welter, J., Natale, M.T., Piccininno, M., Baumann, G., Fernie, K., Gavrilis, D., Rendina, M., Verbruggen, E., Ivacs, G., van Schaverbeke, N., Garvin, J.: Amount of Data Partners and Outeach to Major Institutions. Europeana Core Service Platform D1.2 (2017). 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. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S.: The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0 (2002). http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
4. Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1: Reference Description. DCMI Recommendation (2012). 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. Freire, N., Manguinhas, H., Isaac, A., Robson, G., Howard, J.B.: Web technologies: a survey of their applicability to metadata aggregation in cultural heritage. In: 21st International Conference on Electronic Publishing (2017)
7. Wallis, R., Isaac, A., Charles, V., Manguinhas, H.: 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 J. (36) (2017). http://journal.code4lib.org/articles/12330. ISSN 1940-5758
8. Definition of the Europeana Data Model v5.2.8. Europeana Foundation (2017). http://pro.europeana.eu/edm-documentation
9. Gradmann, S.: Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. Europeana Whitepaper (2010). http://pro.europeana.eu/publication/knowledgeinformation-in-context
10. Charles, V., Isaac, A.: Enhancing the Europeana Data Model (EDM). Europeana white paper (2015). http://pro.europeana.eu/files/Europeana_Professional/Publications/EDM_White Paper_17062015.pdf
11. Berners-Lee, T.: Linked Data. W3C Design Issues (2006). http://www.w3.org/DesignIssues/LinkedData.html
12. Digital Public Library of America. Metadata Application Profile, Version 4.0 (2015). https://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf
13. Europeana Data Model - Mapping Guidelines v2.4. Europeana Foundation (2017). http://pro.europeana.eu/edm-documentation
14. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 29–37 (2001)
15. Snydman, S., Sanderson, R., Cramer, T.: The International Image Interoperability Framework (IIIF): a community & technology approach for web-based images. Archiving (2015). http://purl.stanford.edu/df650pk4327
16. Freire, N., Robson, G., Howard, J.B., Manguinhas, H., Isaac, A.: Metadata aggregation: assessing the application of IIIF and sitemaps within cultural heritage. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 220–232. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_18