information

MDPI

Article

Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network [†]

Nuno Freire [1,*], René Voorburg [2], Roland Cornelissen [3], Sjors de Valk [4], Enno Meijers [4] and Antoine Isaac [5,6]

[1] INESC-ID, Rua Alves Redol 9, 1000-029 Lisbon, Portugal
[2] National Library of the Netherlands, 2595 BE The Hague, The Netherlands
[3] Metamatter, 9832 TE Den Horn, The Netherlands
[4] Dutch Digital Heritage Network, 2595 BE The Hague, The Netherlands
[5] Europeana Foundation, 2595 BE The Hague, The Netherlands
[6] Department of Computer Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands
[*] Correspondence: nuno.freire@tecnico.ulisboa.pt; Tel.: +351-936-055-536

[†] This paper is an extended version of our presentation in the 2018 IEEE International Conference on Big Data, Seattle, WA, USA, 10–13 December 2018.

Received: 18 June 2019; Accepted: 27 July 2019; Published: 30 July 2019

Abstract: Online cultural heritage resources are widely available through digital libraries maintained by numerous organizations. In order to improve discoverability in cultural heritage, the typical approach is metadata aggregation, a method where centralized efforts such as Europeana improve the discoverability by collecting resource metadata. The redefinition of the traditional data models for cultural heritage resources into data models based on semantic technology has been a major activity of the cultural heritage community. Yet, linked data may bring new innovation opportunities for cultural heritage metadata aggregation. We present the outcomes of a case study that we conducted within the Europeana cultural heritage network. In this study, the National Library of The Netherlands contributed by providing the role of data provider, while the Dutch Digital Heritage Network contributed as an intermediary aggregator that aggregates datasets and provides them to Europeana, the central aggregator. We identified and analyzed the requirements for an aggregation solution for the linked data, guided by current aggregation practices of the Europeana network. These requirements guided the definition of a workflow that fulfils the same functional requirements as the existing one. The workflow was put into practice within this study and has led to the development of software applications for administrating datasets, crawling the web of data, harvesting linked data, data analysis and data integration. We present our analysis of the study outcomes and analyze the effort necessary, in terms of technology adoption, to establish a linked data approach, from the point of view of both data providers and aggregators. We also present the expertise requirements we identified for cultural heritage data analysts, as well as determining which supporting tools were required to be designed specifically for semantic data.

Keywords: data aggregation; data analysis; datasets; semantics; Big Data variety; RDF

1. Introduction

Nowadays, online cultural heritage resources are widely available on the web, through the digital libraries of heritage institutions. Several of these resources do not contain natural language texts (as, for example, the cases of pictures, videos, music and other sound recordings), and others that are textual often lack machine-readable text data that can be used for indexing by search engines. These latter resources consist of digitized images in which optical character recognition (OCR) was not applied, due to limited financial resources or the lack of adequate OCR technology (as, for example, in early printed materials containing hand-writings, or printed materials using old text fonts). Given the nonexistence of textual content, in order to enable the discoverability of these resources, cultural heritage institutions have traditionally turned to creating and exploiting metadata for the resources (that is, data records describing the resources).

The redefinition of the traditional data models for cultural heritage resources into novel data models based on semantic data technology has been a major activity of the cultural heritage community. Nowadays, many digital libraries and online catalogs make available linked data representation of the metadata about cultural heritage resources. Cultural heritage comprises a very diverse community, however, leading to the usage of several different data models. It constitutes several subcommunities (Libraries, Museums, Archives, and Galleries) which are of a transnational nature, leading to the application of many resource description practices among the different communities, often defined within a regional context, such as a country. For cultural heritage data, the most prevalent data management challenge is its variety [1], due to the lack of homogeneity in the application of data models, languages, and other data related practices.

Digital libraries have specific functionality for the retrieval of metadata descriptions. But given the large number of existing individual digital libraries which are hosted by different institutions, achieving wide, cross-institutional discovery is a challenging problem. Effective mechanisms designed for the discovery of resources by all potential users have to be put in practice. Discoverability is typically addressed by an organizational architecture, where a central organization runs services to improve the discoverability of cultural heritage resources by collecting their associated metadata. Typically, this central organization is an influential cultural heritage institution, but cases also exist of organizations being set up specifically for this purpose. The central organization is in a position to better enable the wide discovery and reuse of the resources, by application of practices and technologies that cannot be carried out sustainably by each single digital library on its own. Web portals are usually setup which deploy cultural heritage-adapted search engines that are tailored for the retrieval of metadata records.

The approach generally used to feed these portals is metadata aggregation. Within cultural heritage, specific data aggregation technologies are in use, which are different from the technologies used in other domains, such as by internet search engines or in the Web of Data. The Open Archives Initiative Protocol for Metadata Harvesting [2] (OAI-PMH) has been the dominant technology, since it is specialized for the aggregation of datasets of metadata, covering some specific requirements for the aggregation of this type of data. OAI-PMH has been defined by the academic preprints' community, and has seen significant participation by cultural heritage actors who have applied the technology since its early stages.

Models for the scalable and sustainable discoverability of resources have been prepared and put in practice in the domain of cultural heritage. Cooperation and new organizations have implemented these models in practice. Some well-known cases are Europeana in Europe, DPLA in the United States of America, Trove in Australia, Digital Library of India, and DigitalNZ in New Zealand. These organizations collect and facilitate public access to digitized cultural resources. The implementation of the technological infrastructures for data aggregation have high costs, however, and are particularly demanding on the data providing institutions. Reducing the effort and cost required for running these infrastructures would bring more data providing participants to these networks. A larger number of providing institutions is important in achieving a critical mass for increasing the sustainability of the whole networks [3]. In this scenario, if aggregators were able to make use of the available linked data, data providers could benefit from several advantages, providing them with further motivations, e.g.,

- For those already making available linked data through their digital libraries, the process of sharing metadata with cultural heritage aggregators would not require much additional effort, since they could simply provide the linked data that they already make available for other purposes.
- For those that are not yet publishing linked data, setting-up the technological infrastructure for cultural heritage aggregation based on linked data, would be more beneficial, because, in addition,

they would benefit from wider interoperability with other application cases than aggregation, and possibly with other domains besides cultural heritage.

This article presents our long-term method for the innovation of cultural heritage data aggregation based on linked data and its underlying technologies. We analyzed the current data aggregation requirements of the Europeana Network for guiding our method in designing this novel linked data approach. We present the outcomes and discuss the conclusions of one case study that we performed by involving three members of the Europeana network, which include both aggregators and data providers. In this study, the National Library of The Netherlands contributed by assuming the role of a data provider, while the Dutch Digital Heritage Network contributed as an intermediary aggregator that aggregates datasets and provides them to Europeana, the central aggregator.

Section 1.1 will follow with a description of the technological approaches to metadata aggregation most commonly found in practice within cultural heritage. Section 1.2 presents related work on metadata aggregation in cultural heritage, covering both the current solutions and linked data aggregation. Section 2 presents the requirements that we identified for guiding the design of our method of linked data aggregation within the Europeana network. It continues with the presentation of an overview of our proposed method for an aggregation workflow for linked data. Section 3 presents the ways in which the workflow was realized within the case study, as well as the results of addressing the activities of that compose the overall workflow, such as the guidelines, specifications, adopted standards, procedures and software systems. In Section 4, the results are discussed and the resulting conclusions are presented. Section 4 also describes further studies that may be conducted in the future in order to progress towards the establishment of a linked data standard method for metadata aggregation in cultural heritage.

1.1. Characterization of Metadata Aggregation in Cultural Heritage, and the Specific Case of Europeana

Cultural heritage is a domain with its own characteristics which have guided the establishment of the current practices for metadata aggregation. We have identified these characteristics as the most influential:

- Cultural heritage is divided in subdomains: Libraries, Museums, Archives, and Galleries.
- All subdomains apply their own resource description data models, norms and practices.
- Not all subdomains have a significant definition of standards-based solutions for description of cultural heritage resources, or they are only adopted by a few institutions. Libraries have traditionally established cooperation within its domain to enable shared services for bibliographic data. For archives and museums, however, the existing norms and standards have resulted from more recent cooperation.
- Interoperable information systems and data models are not common across subdomains. Within each of the subdomain, however, interoperable systems are frequently found. The interoperability is sometimes established within a country, and also at a wider international level, particularly in the subdomains of libraries and archives.
- XML-Schema data models are the basis of most of the adopted standards. Interoperable data models defined for relational data are not commonly found.
- The adoption of new technologies is not very agile, and may not be systematic in the early stages (for example, we have observed this aspect in the application of structured data in the Web by cultural heritage institutions [4]). Organizations operate with limited financial resources to invest in information technology innovation.

In the metadata aggregation scenario, the usual practice has been to aggregate metadata represented according to data models that have been agreed upon within a specific domain/community. This approach enables the sharing of metadata across organizations and countries, sharing enough semantic detail for the target interoperability, and with some acceptable losses in the semantic details of the original metadata from the data providers. It has made possible the establishment of sustainable solutions for handling the high data heterogeneity of cultural heritage. These solutions usually adopt data models with low complexity, so that they simplify the understanding of the model by the different data providers (possibly across subdomains). Simpler data models also provide a lower barrier for the setup of data processing systems (such as data converters) by both providers and aggregators.

An important characteristic of metadata aggregation is the solution for sharing the sets of metadata from the providing institutions to the aggregator. The metadata is transferred to the aggregator, but it continues to evolve at the data provider side; thus, the aggregator needs to periodically update its copy of the data. Under this scenario, the need for data sharing can be described as a data synchronization problem across organizations.

OAI-PMH is the most well-established solution to the data synchronization problem within the cultural heritage domain. It is not restrictive in terms of the data model being used; therefore, it enables the sharing of metadata for aggregation according to the data model adopted within each aggregation network. OAI-PMH is only capable of transporting XML data, however. This imposes the representation of the metadata in XML.

In the case of Europeana, the solution for data modelling is the Europeana Data Model (EDM) [5]. EDM supports several of the core processes of Europeana's operations and contributes to the access layer of the Europeana platform, supporting the sharing of data with third parties [6] and allowing 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' [7]. EDM is also the underlying data model for other participants of the Europeana network, as well as outside the network. Other organizations using approaches for aggregation similar to that of Europeana also apply EDM, e.g., the Digital Public Library of America (DPLA), which operates within the United States of America and uses a model based on EDM within the aggregation process of its network [8]. EDM has been developed (and is still being improved) in collaboration with the wider cultural heritage community, involving representatives from all the domains represented in Europeana, i.e., libraries, museums, archives, and galleries. The classes defined by EDM are shown in Figure 1, and an illustrative example of a cultural heritage object represented in EDM is shown in Figure 2.

EDM can be represented in RDF by means of the standard serialization of RDF in XML, making it compatible with Europeana's choice to rely on OAI-PMH for metadata synchronization. In fact, EDM is based on the principles of the Web of Data [9] and uses fairly generic constructs. This relative genericity (which also applies to the systems built on top of EDM) makes it easier to explore technological alternatives for metadata aggregation, which could be based on other data models, as long as they are able to address the information needs of Europeana.

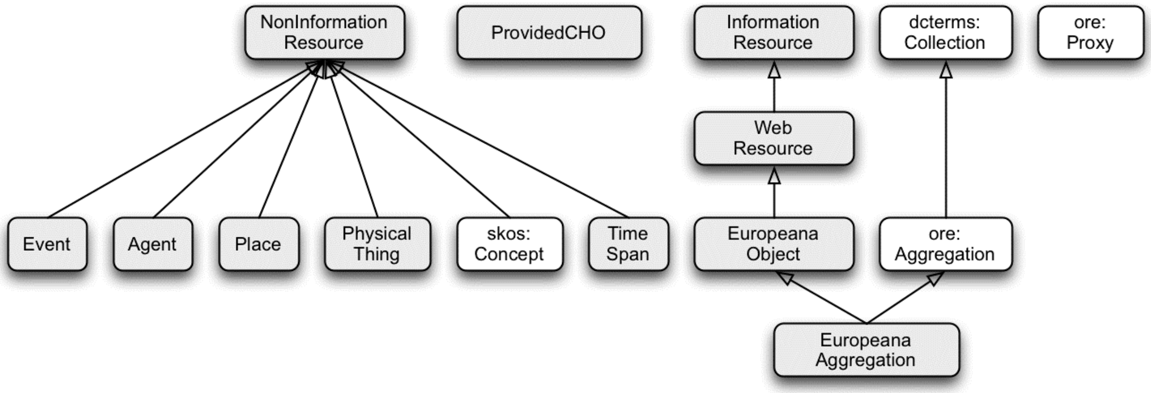

Figure 1. General overview of the classes defined in the Europeana Data Model. The 'ProvidedCHO', representing a cultural heritage object, is the central point of an EDM metadata record [5].

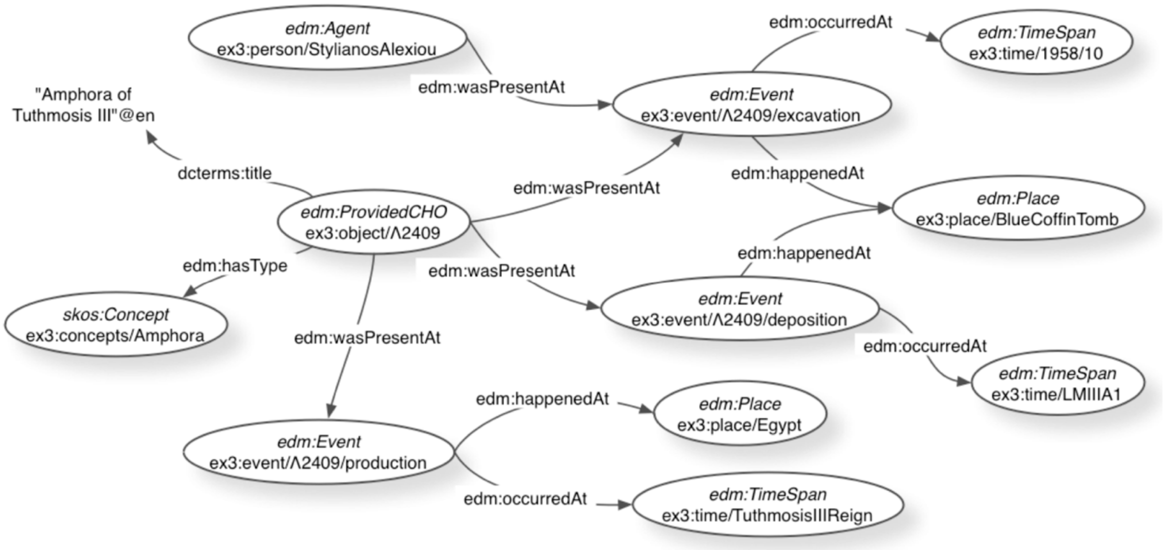

Figure 2. Example of metadata record using the Europeana Data Model, illustrated as a RDF graph.

1.2. Related Work

As stated, the metadata aggregation practice currently established within the cultural heritage domain is based on the OAI-PMH protocol. OAI-PMH was specified mainly by the academic community to address similar requirements for interoperability between digital libraries of e-print publications in scholarly communication. Like the e-prints community, the cultural heritage domain has extensively applied OAI-PMH. The use of OAI-PMH in Europe's cultural heritage institutions has been highly motivated by Europeana, that adopted it since its initial operation in order to aggregate metadata from its network of data providers and aggregators.

The solution for Europeana's data synchronization has not been reevaluated since its adoption by Europeana, however, even though the technological landscape has changed. The first specification of OAI-PMH date from 1999 [2], and its current version is still based on the same underlying technologies. Information technology continued to progress, however, and several new technologies are available based on progress in network communication, computational power, and Internet search engines. OAI-PMH was designed before the establishment of the major founding concepts of the Web of Data [9]. By being centered on the concept of a repository, instead of focusing on resources, the protocol is often poorly understood, resulting in failed implementations, or in deployments with incorrect implementation of the protocol that undermine the reliability of the aggregation process [10]. Another relevant aspect is that OAI-PMH was defined before the emergence of REST [11]. Therefore, it does not follow REST's underlying foundational concepts, which are nowadays widely adopted by software developers, yielding further resistance and difficulties to its understanding and implementation of OAI-PMH by cultural heritage institutions.

Currently, the discovery of e-prints is largely based on full-text processing, relying on newer technologies, such as ResourceSync [12], which are less specialized towards metadata and address more generic requirements of data synchronization.

Within the cultural heritage domain, metadata-based discovery remains the main information source for discovery of resources, since indexable full-text is still not available for most collections and types of content. Motivation for adopting OAI-PMH is lower for data providers, however, because it no longer presents itself as the only major viable technology. And in recent years, our experience with OAI-PMH is that many cultural heritage institutions need to implement it just for providing data within one single aggregation case, such as that of Europeana.

Linked data has received attention from many cultural heritage researchers and practitioners. However, we find in most cases that the main focus of the literature concerns the publication of linked data [13–17] and does not investigate in detail how the metadata aggregation in cultural heritage can be deployed on top of available linked data. Two noticeable exceptions are the Research and Education Space project (RES) and the Dutch Digital Heritage Network (NDE, which participated also in our case study).

NDE is a national program aiming to increase the social value of the collections maintained by libraries, archives and museums in the Netherlands. NDE is an ongoing project, and its initial proposals are based on specific APIs to enable data providers to centrally register the linked data URIs of their resources [18]. The initial results from NDE are thus based on its own defined API, and have not yet provided a solution purely based on linked data, which is the final goal of NDE.

The Research and Education Space project, finished in 2017, has successfully aggregated a considerable number of linked data resources from cultural heritage, education and academic data sources. The resulting aggregated dataset can be accessed online, but an evaluation of its aggregation procedures and results has not been published. The project's available technical documentation [7] addresses some of the required functionality that is relevant to our work. Some tasks, however, were not fully documented in the last specifications published by the project.

Solutions have been proposed by others for aggregation of linked data (for example [19]) that tried to tackle the issue with generic solutions. None of the work in this area resulted in a standardized approach, however, and we could not find any sustainable application within cultural heritage.

The work we present here was conducted in the overall context of a line of research within Europeana, which aimed at improving our Network's efficiency and sustainability. Our earlier studies on this topic identified linked data as a potential technical solution to perform better against these objectives in the area of metadata aggregation [20]. In particular, the study reported here is a follow-up to several experiments investigating various Web technologies for Europeana's aggregation purposes [21,22]. An important point is that linked data provides a technological basis for metadata aggregation, which brings new options in terms of both data synchronization and data modelling. We have studied both problems in the case study that we describe in the sections below.

Our case study focused on a scenario of linked data aggregation using mostly the Schema.org vocabulary [23] to represent cultural heritage object metadata. Schema.org is an initiative which seeks to encourage the publication and consumption of structured data in the Internet. It is a cross domain vocabulary originally created by the major Internet search engines, but nowadays evolves as a community-based effort. Within the cultural heritage domain, the National Library of the Netherlands (Koninklijke Bibliotheek, KB) has created a Schema.org data representation for the Dutch National Bibliography (KB's linked data representation of the Dutch National Bibliography is hosted at http://data.bibliotheken.nl; it is mostly based on the Schema.org vocabulary, but it also uses a handful of properties from other vocabularies). It is available as linked data, as one of the openly published datasets by the KB. Europeana also makes available Schema.org metadata within the webpages of its portal. This publication of Schema.org by Europeana follows research about the best practices for publication of Schema.org cultural heritage metadata [24]. Directly related to the work of this particular paper is an evaluation of Schema.org usage in cultural heritage institutions for aggregation by Europeana [22], whose positive outcome has been support for also researching Schema.org for linked data aggregation, since Schema.org is extensively applied in linked data.

2. Materials and Methods

2.1. Requirements

The functionality to be provided by a linked data aggregation solution must comply with the currently existing requirements for aggregation in the Europeana network, which are based on the metadata harvesting functionality of OAI-PMH, and the EDM data model. In order to establish a linked data infrastructure for the Europeana network that is fully based on well-established standards, some aspects of standards and guidelines for linked data also need to be supported. This general context represents the following requirements that a technical solution must meet:

- Requirement R1—Data providers must be able to provide a linked data resource providing metadata about their dataset.
- Requirement R2—Data providers must be able to specify complete datasets or only some subset(s), because parts of the complete dataset may be out-of-scope, or not compliant, with aggregation for Europeana.
- Requirement R3—Data providers must be able to specify machine-actionable licensing statements regarding their metadata (we use the term 'machine-actionable' throughout this article because for automatic processing, a solution requires more than machine-readable metadata; it requires data that contains sufficient semantics for the functionality to be performed by software). Since it is not always the case that all resources in a dataset have the same associated license, data providers should be able to specify the license both for the complete dataset, and for each individual metadata resource.
- Requirement R4—Data providers must provide a machine-actionable specification of the available mechanisms, and their respective access information, to harvest the dataset or to download it in a standard distribution format.
- Requirement R5—The aggregator must be able to apply fully automatic methods to harvest or download the linked data datasets specified by its data providers.

Furthermore, considering the (operations) context of Europeana, the design of an innovative linked data aggregation solution should comply with the following constraints:

- Constraint C1—All data communications between data providers and the aggregator must be done with standard technologies of the Semantic Web and linked data.
- Constraint C2—The final data model of the aggregated metadata must be EDM, because EDM is the data model that all Europeana systems, and many of its aggregators, are built on (note that at Europeana, EDM supports more than the metadata aggregation infrastructure; for example, EDM is underlying its information retrieval engine, end-user portal, APIs and other systems for data distribution). Other data models, and vocabularies not used in EDM, may be used in the intermediate steps of the aggregation's workflow, however.
- Constraint C3—The solution must simplify the requirements for data providers to provide data to cultural heritage networks. The technical requirements should have low barriers for adoption by data providers and allow the re-use of existing linked data resources available at data providers, and which are compliant with the requirements of the cultural heritage network.
- Constraint C4—The manual effort required for the aggregators should not increase with a new solution. The data integration work required from linked data sources should be automated as much as possible in order to decrease, or at least maintain, the current operational costs of the cultural heritage networks.

Considering this context, we have defined an aggregation workflow proposal, of which we present an overview in the following section.

2.2. Workflow

We aim to provide the necessary methods and software tools that will assist data providers and aggregators to fulfill a role they already play. Our proposed aggregation workflow for linked data aggregation is therefore very similar to the current practices of cultural heritage aggregation networks. It is composed of 7 main activities, and involves 2 actors: data providers and aggregators (with Europeana as a specific case of an aggregator). It is executed within an information system that includes several software applications which support the different activities. The starting point of the workflow is when a data provider has already prepared and published the linked data describing the cultural heritage digital objects that constitute the dataset for aggregation.

The workflow is depicted in Figure 3, and its 7 activities are the following:

1. Publish dataset description—The data provider describes its dataset as a linked data RDF resource. It contains descriptive information about the cultural heritage content covered by the dataset and technical details about how the dataset can be collected by the aggregator.
2. Register the dataset's URI—The data provider notifies the aggregator of the availability of its dataset for aggregation. This notification consists of at least the URI of the dataset's RDF description that is created in the first activity, from which the aggregator may collect the information he requires to process the dataset.
3. Harvest dataset—The dataset is harvested by the aggregation system. The system follows the instructions for automatic dataset collection defined in the dataset's RDF description.
4. Profile dataset—The aggregator creates a data profile of the harvested dataset, supported by tools that analyze the aspects of the dataset relevant for performing its aggregation and integration into the central dataset. A dataset's profile allows the aggregator to detect deficiencies in the dataset and notify the data provider about them. The profile also supports the assessment of the effort (regarding specific data preparation tasks) required for the next steps of the workflow.
5. Align data model and vocabularies—The aggregator analyzes the data model and vocabularies used by the linked data of the data provider, and compares them with the reference data model it uses (typically EDM or a profile thereof). As a result of this analysis, the aggregator sets any alignments necessary between the two models. In cases where the data provider's linked data is based on EDM, no alignments are necessary, unless one of the existing extensions of EDM is used. The alignment task is supported by the report resulting from the profiling of the dataset and by software tools for data alignment and data integration.
6. Convert dataset—The system converts the harvested linked data to the data model used in the central aggregated dataset. The conversion follows the data model alignments defined in the previous activity. The resulting EDM data is inspected by the aggregator according to the relevant validation and data quality assurance practices.
7. Integrate dataset—The linked data aggregation system transfers the converted dataset to the regular ingestion systems of the aggregator. The dataset in EDM will then be processed by the regular data integration workflow of the aggregator, which is independent of the initial harvesting method applied for the dataset.

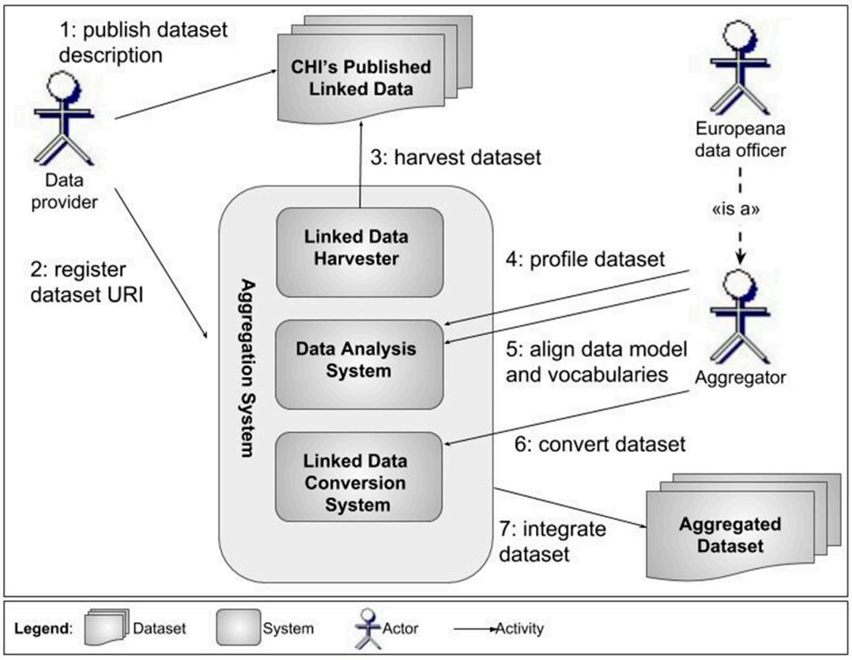

Figure 3. Overview of the linked data aggregation workflow for cultural heritage.

We have tested this workflow in a pilot with members of the Europeana network. The results of this initial practical experiment with the workflow are presented in the Section 3.

2.3. Pilot Setup

In pursuit of our objective for innovation in cultural heritage metadata aggregation, with a particular focus on sustainable solutions with low technical requirements, we have organized a group of institutions that represent the different points-of-view from partners that perform different roles in our workflow. Our pilot had three participants which were representative of each of the three roles in the organizational structure of the Europeana network: the National Library of The Netherlands contributed with the role of data provider, while the Dutch Digital Heritage Network contributed as an intermediary aggregator that aggregates datasets and provides them to Europeana, the central aggregator. All three have already been involved in discussions focused on finding new viable solutions that re-use existing know-how in cultural heritage, and reduce the overall resources and effort required for implementation and regular operation of an aggregation network.

In earlier work, the KB had created a Schema.org data representation of the Dutch National Bibliography, and had made initial steps toward the publication of linked data having Schema.org as the main vocabulary. This earlier work was applied to several but not all of the datasets in KB's digital library. In our pilot, the KB and NDE have worked on more collections, with particular focus on collections of interest for aggregation by Europeana. The KB has revised its earlier work with Schema.org modeling, focusing on the particular description of digital objects, which were not modeled in detail for the Dutch National Bibliography. KB and NDE also revised the modeling in order to ensure compliance with the informational requirements of Europeana. Finally, KB also implemented all the necessary mechanisms that are required for making the datasets harvestable (which we will describe in the following section). Europeana harvested the datasets, analyzed their compliance with its informational requirements, prepared the datasets and ingested them in its production environment. In order to have a running workflow, we have analyzed standards, published guidelines, implemented software systems and specified data conversions. Figure 4 shows the main systems and data resources that were involved in the pilot. In the following section, these components are discussed and presented in detail.

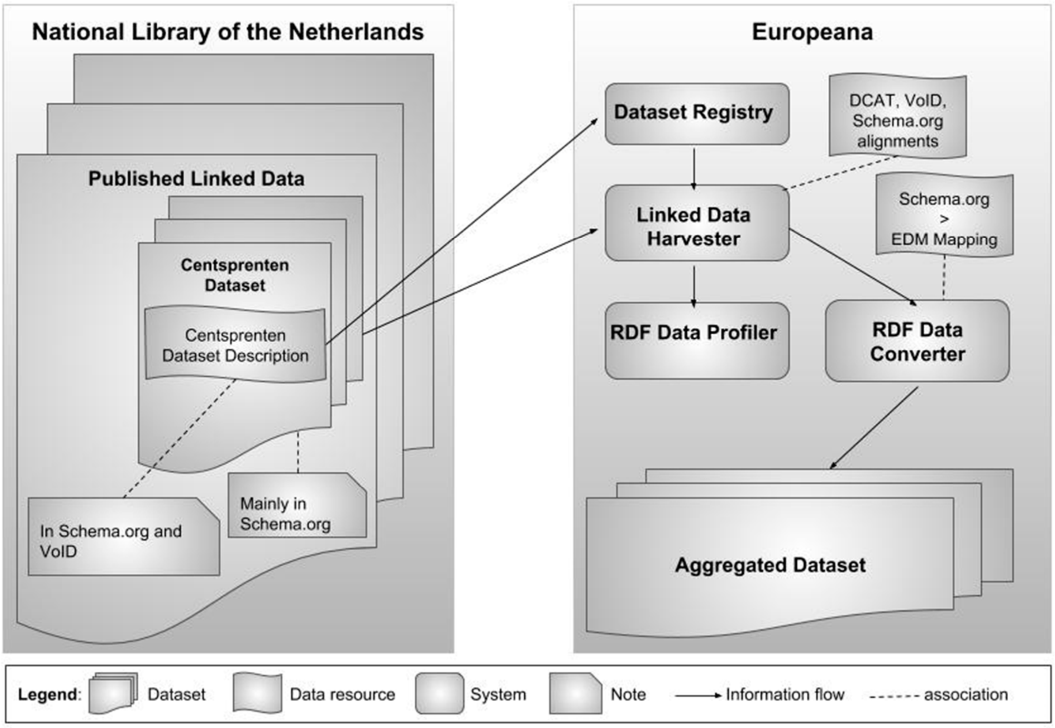

Figure 4. The main systems and data resources used in the pilot.

3. Results

3.1. Dataset Description

The initial work of the pilot focused on the implementation options for the requirements presented in Section 2.1. We identified the need for the datasets to be described as linked data resources in order to create machine-actionable descriptions of the datasets, including technical details about how an aggregator can obtain them by automatic means. For this purpose, we have conducted a review of technologies that could be used for dataset descriptions. We focused on a review of several standard (Linked Data) vocabularies for describing datasets [25], and found vocabularies that are able to address requirements R1, R3 and R4. This work resulted in the creation of guidelines (Specifying a LOD dataset for aggregation by Europeana. Available online: https://github.com/nfreire/Open-Data-Acquisition-Framework/blob/master/opaf-documentation/SpecifyingLodDatasetForEuropeana.md#specifying-a-lod-dataset-for-aggregation-by-europeana) to support data providers in preparing their dataset descriptions for aggregation by Europeana.

We identified three suitable vocabularies to fulfill the requirements for linked data aggregation: VoID [26], DCAT [27], and Schema.org (note that this usage of Schema.org is separate from the usage for metadata of cultural heritage resources) [23]. These allow data providers to describe their datasets for humans, and also to express the licenses which are applicable to their data in a machine-actionable way (requirements R1 and R3). The main distinction between these vocabularies regards their ability to support all the options for allowing data providers to provide the technical details about how their datasets can be downloaded or harvested (requirement R4).

Any of the three vocabularies allow the specification of "distributions" of the datasets as downloadable RDF files (commonly known as "dumps"). VoID, however, is the only vocabulary that can be used to specify "root resources" for the dataset, which allows an entire dataset to be aggregated by linked data crawling methods. The VoID listing of root resources makes possible the application of a linked data crawler that resolves the root resources and follows the links to other URIs in the retrieved RDF response. This frees data providers from having to create and maintain downloadable distributions for their complete datasets, which practical experience in Europeana shows to be often a challenge for the under resourced information systems of cultural heritage institutions.

Since only VoID enables crawling, we considered that requirement R4 is only partially fulfilled by DCAT and Schema.org, but this is not an impediment to use these two vocabularies. We have brought this issue into the discussion within W3C's Data Exchange Working Group (DXWG) (W3C Data Exchange Working Group: https://www.w3.org/2017/dxwg/wiki/Main_Page), which maintains DCAT. The discussion on this issue may be followed in the issue management system used by the DXWG (the discussion of Europeana's requirement can be consulted at https://github.com/w3c/dxwg/issues/292; see also https://github.com/w3c/dxwg/issues/253 and https://github.com/w3c/dxwg/issues/482 for an overview of the solution that will be available in the coming version of DCAT). At the time of writing of this paper, a viable solution has been proposed and accepted by the DXWG; therefore, we expect that the next version of DCAT will fully support our requirement.

Given the availability of three possible solutions, all based on Web standards, we designed a flexible solution that privileges the ease of adoption by data providers. We allow data providers to use any expertise that they may have in-house, and apply any of the three vocabularies. Although supporting the use of all vocabularies may imply higher maintenance effort for Europeana in the future, we believe that this effort can be mitigated by leveraging on the active communities of users that these vocabularies have.

As we will describe in the next sections, the linked data harvester that we developed for the pilot is thus able to interpret the three vocabularies. The main aspects required by the harvester from the dataset resource descriptions are summarized in Table 1, which also indicates the key classes and properties of all three vocabularies.

Table 1. The most relevant classes and properties for describing datasets and their distributions.

Requirement	DCAT	Schema.org	VoID
Class for 'Dataset'	dcat:Dataset	Schema:Dataset	void:Dataset
Title of the dataset	dcterms:title	schema:name	dcterms:title
Class for Distribution	dcat:Distribution (associated with the dataset with dcat:distribution)	schema:DataDownload (associated with the dataset with schema:distribution)	A Distribution class is not defined by VoID. The distribution is represented as properties of the dataset
Dataset level license	dcterms:license	schema:license	dcterms:license
Downloadable distribution	dcat:download URL dcat:mediaType	schema:contentUrl schema:encodingFormat	void:dataDump void:feature void:TechnicalFeature
Listing of URIs	Not (yet) supported in DCAT	Not supported in Schema.org	void:rootResource

Note: Prefixes and namespaces used in this table—dcat="http://www.w3.org/ns/dcat"; schema="http://schema.org/"; void="http://rdfs.org/ns/void#"; dcterms="http://purl.org/dc/terms/".

3.2. Dataset Registry

A central dataset registry, where the aggregator manages the metadata regarding all the datasets it aggregates, is the first software component supporting the workflow. The registration of a dataset is done when the data provider notifies the aggregator of the availability of its dataset for aggregation. This notification shall include the URI of the dataset's resource description.

The dataset's RDF resource is then fetched by the registry. Periodically, the registry may poll the dataset's URIs to keep information about it up to date, and possibly to trigger a new iteration of its aggregation process for the dataset.

3.3. Linked Data Harvester

The aggregator's Linked Data Harvester supports the interpretation of the three vocabularies by implementing the alignments mentioned earlier, i.e., in Section 3.1 (see also Table 1).

By interpreting the dataset's RDF description, particularly the dataset's distribution information, it activates the appropriate harvesting mechanism:

- Based on crawling through root resources—in this option, data providers use the property void:rootResource pointing to entry points that allow the crawler to reach the resources that describe the cultural heritage objects. One or more root resources can be specified, and they may refer directly to the URIs of cultural heritage objects.
- Based on downloading a distribution—by using this option, data providers have the possibility to use the classes and properties of any of the vocabularies, given that all three support this mechanism. The downloadable distribution must be available as RDF files using one well known serialization for RDF.

The linked data harvester is under active development for supporting ongoing research and technology adoption in Europeana. The future outcomes and current source code can be accessed as open source online, as part of our "Data Aggregation Lab" (Data Aggregation Lab software: https://github.com/nfreire/data-aggregation-lab).

3.4. RDF Data Profiler

The RDF Data Profiler is the first tool of the workflow that Europeana, or other aggregators, would use for converting the dataset into their own data model (usually EDM) so that the data may be integrated into the central aggregated dataset. The profiler supports the aggregator to get an insight of the tasks that must be performed on the harvested data in order to successfully process and ingest the dataset. In particular, the profiler plays a key role in making the data model mapping task more efficient (effort-wise) and accurate.

In our work for the pilot, we have implemented a data profiler that is designed for RDF. This supports the aggregator in grasping a view of the data model being used in the harvested linked data by reporting on the usage of RDF classes and properties on the individual RDF triples and their usage as subject, object or predicate.

The data profiling capabilities developed for this pilot proved very helpful for the definition of the mapping from Schema.org to EDM, and also allowed us to identify other profiling operations and reports which were particularly useful for the definition of data conversions, as we will describe in our conclusion.

3.5. Schema.org to Europeana Data Model Mapping

The Europeana Data Model supports all data exchange within the Europeana aggregation process, enabling the expression of cultural heritage metadata with enough homogeneity. As stated in Section 1.1, an important aspect of Europeana's adoption of EDM, is that it paves the way for more flexibility on the choice of Web metadata technologies for exchange of metadata. Although until now, Europeana accepts only datasets that are prepared in EDM by data providers or aggregators, we believe that in a linked data metadata aggregation approach, supporting just EDM would not be feasible or would significantly reduce the benefits of linked data. Therefore, we consider the addition of a data model mapping step an essential aspect to be introduced in aggregation. In fact, linked data technology makes the data mapping task much more feasible across various contexts than it would be without the level of data standardization that it brings.

The usage of data models and vocabularies in linked data from cultural heritage is still very diverse, however, and an aggregator cannot be expected to accept any vocabulary or combinations thereof. In the longer term, we expect our work to address the conversion of several linked data models to EDM under the guidance of the community. In this pilot, we focused on Schema.org, a very detailed model that is gaining much interest in cultural heritage as a way to increase the wider interoperability of cultural heritage data on the Web.

In past experiments, we have worked on the definition of mappings between Schema.org and EDM [22,24]. In the pilot, we applied our earlier mapping from Schema.org and EDM and improved it based on the data profiling report that analyzes how the Schema.org data model is used in the KB dataset.

The definition and maintenance of data model mappings is a key aspect for achieving sustainable aggregation of linked data within the context of cultural heritage. In this pilot, we focused on how to make machine-actionable specifications of the mappings that cultural heritage data experts have prepared with us in past work, based on the characteristics of RDF data. We consider our work in this area to be still in its early stages: our representation of the mapping specifications is machine-actionable, but in a rather procedural way, as it is done in the Java programming language (Data Aggregation Lab software: https://github.com/nfreire/data-aggregation-lab). However, it allowed us to better identify and to begin implementing a sort of mapping repository functionality: aggregators can identify and maintain their mappings between data models, and can later reuse them for other linked data datasets.

Our mapping specifications also allow users to track the mappings and data manipulations affecting specific RDF properties and classes. When aggregators analyze the data model of a dataset, they can easily detect which classes and properties require attention for defining mappings. This task, when used in conjunction with the data profiling reports, helps the aggregators to perform the mapping tasks more efficiently.

3.6. RDF Data Converter

The RDF Data Converter implements the execution engine for the mapping specifications described in Section 3.5. This component transforms an RDF graph into another RDF graph by following the data conversion operations specified in a mapping specification.

The mapping specifications are managed persistently in the aggregation system, and they can be associated with datasets. Whenever a dataset needs to be processed for integration, either for the first integration of the dataset or for its future updates, the associated mapping specification is executed, resulting in the source data being converted into EDM.

The datasets, once converted to EDM, are then exported into files. In the final step of the workflow, they are provided to the core ingestion systems of Europeana, which will take the dataset through the regular data ingestion operations currently in operation at Europeana.

The pilot was one of the first applications of the RDF Data Converter on real datasets. It was possible to apply it without adding functionality that was not required in previous experiments. Therefore, we believe that the set of RDF data conversion operations which are currently supported are suitable for a high percentage of linked data, although further experimentation will be needed.

4. Discussion

Our pilot allowed us to identify and gain practical experience with several aspects of performing cultural heritage data aggregation based on linked data, mainly regarding technology adoption, aggregation processes and software applications.

In terms of technology adoption, a linked data solution has a different impact on the institutions that perform the different roles of data provider and aggregator:

- Adoption by data providers—During the pilot, the implementation of the requirements was always done without much effort from the KB. Also, in the process, the KB was able to reuse its previous linked data work. The example of the KB, however, cannot represent all cultural heritage data providers, since not all of them have previous experience with linked data. Institutions without in-house knowledge of linked data are likely to find the requirements more challenging, since they will need to employ many of the techniques underlying linked data technology. Another observation is that the freedom to apply another data model to represent the cultural heritage objects was a motivation for the KB. In this case Schema.org was a good motivation for KB since it is capable to reach wider audiences than EDM, particularly through internet search engines.
- Adoption by aggregators—We found that the implementation of the requirements was more demanding for the aggregator than for the data provider. Nowadays, the majority of aggregators operate based on XML solutions, while for a linked data workflow, aggregators need to deploy data processing software for RDF data. More knowledge in RDF may be necessary in organizations, especially for information technology departments and for data officers who need to align the linked data models in use by providers with the EDM required by Europeana. The impact on human resources would be especially important for the aggregators who do not yet need to align data models, because they require EDM from their providers and thus operate solely on EDM data.

The pilot has also confirmed the need for software applications to execute and support the operation of an aggregation of linked data. The alignment of the data models in use in cultural heritage linked data will demand additional effort from the aggregator. This higher demand would be unsustainable without the support of tools that automate and support data specialists in their alignment tasks. We observed in the pilot that alignment requires major efforts when facing a dataset with a data model encountered for the first time. However, the effort is greatly reduced when the same model is processed again for subsequent datasets. This case study was our third involving Schema.org, and the alignment effort was a fraction of what was necessary in the past. Tools are essential to allow data specialists to manage the accumulation and reuse of data model knowledge and alignments within the organization.

Another relevant aspect of linked data aggregation is the need to manage the semantics represented in RDF (e.g., relations between specialized and more general properties), a key concept of the Web of Data. The functionalities provided by aggregation software applications to data analysts need to be specialized in order to work with this sort of semantic context, which implies also that expertise in semantic data is required from data analysts.

This study and practical experience have clearly identified future work for the software applications that support the aggregator. Additional case studies should also be conducted, in particular with data providers of other subdomains of cultural heritage (museum, archives, etc.), and of smaller dimensions. Technology adoption by smaller data providers is a greater challenge, and they have a strong presence in Europeana that counts some 3700 data providers in its network of partners.

Author Contributions: This article was written based on the following contributions by the co-authors. Conceptualization of the case study by N.F., E.M. and A.I.; Design of the methodology by N.F.; Software development by N.F., R.V. and R.C.; Data curation of the source linked dataset, and validation of the resulting data by R.V. and R.C.; The writing—original draft preparation, was done by N.F., writing—review and editing, was contributed by S.d.V., E.M. and A.I.

Funding: 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 contract number 30-CE-0885387/00-80.e, and grant agreement number INEA/CEF/ICT/A2018/1633581.

Acknowledgments: We would like to acknowledge the supporting work by Valentine Charles and Marjolein de Vos, from the Europeana Foundation, in the analysis of the EDM and Schema.org data models, as well as their analysis of samples collected during the case study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Laney, D. 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research. Available online: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed on 29 May 2019).
2. Lagoze, C.; van de Sompel, H.; Nelson, M.L.; Warner, S. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. Open Archives Initiative. Available online: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm (accessed on 29 May 2019).
3. Niggermann, E.; Cousins, J.; Sanderhoff, M. Europeana Business Plan 2018 'Democratizing Culture'. Europeana Foundation. Available online: https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana_Business_Plan_2018.pdf (accessed on 29 May 2019).
4. Freire, N.; Calado, P.; Martins, B. Availability of Cultural Heritage Structured Metadata in the World Wide Web. In *Connecting the Knowledge Commons—From Projects to Sustainable Infrastructure*; Chan, L., Mounier, P., Eds.; OpenEdition Press: Marseille, France, 2019.
5. Europeana Foundation. Definition of the Europeana Data Model v5.2.8. Available online: http://pro.europeana.eu/edm-documentation (accessed on 29 May 2019).
6. Gradmann, S. Knowledge = Information in Context: On the Importance of Semantic Contextualisation in Europeana. Europeana Foundation. Available online: http://pro.europeana.eu/publication/knowledgeinformation-in-context (accessed on 29 May 2019).
7. BBC. A Guide to the Research & Education Space for Contributors and Developers. Available online: https://bbcarchdev.github.io/inside-acropolis/ (accessed on 29 May 2019).
8. DPLA. Metadata Application Profile, version 5.0. Digital Public Library of America. Available online: https://drive.google.com/file/d/1fJEWhnYy5Ch7_ef_-V48-FAViA72OieG/view (accessed on 29 May 2019).
9. Berners-Lee, T. Linked Data Design Issues. W3C-Internal Document. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 29 May 2019).
10. Van de Sompel, H.; Michael, L.N. Reminiscing About 15 Years of Interoperability Efforts. *D-Lib Mag.* **2015**, *21*. [CrossRef]
11. Richardson, L.; Ruby, S. *Restful Web Services*; O'Reilly: Boston, MA, USA, 2007.

12. NISO. ResourceSync Framework Specification. National Information Standards Organization. Available online: http://www.niso.org/apps/group_public/download.php/12904/z39-99-2014_resourcesync.pdf (accessed on 29 May 2019).
13. Simou, N.; Chortaras, A.; Stamou, G.; Kollias, S. Enriching and Publishing Cultural Heritage as Linked Open Data. In *Mixed Reality and Gamification for Cultural Heritage*; Springer: Cham, Switzerland, 2017; pp. 201–223.
14. Hyvönen, E. Publishing and Using Cultural Heritage Linked Data on the Semantic Web. *Synth. Lect. Semantic Web Theory Technol.* **2012**, *2*. [CrossRef]
15. Jones, E.; Seikel, M. *Linked Data for Cultural Heritage*; Facet Publishing: Cambridge, UK, 2016.
16. Szekely, P.; Knoblock, C.A.; Yang, F.; Zhu, X.; Fink, E.E.; Allen, R.; Goodlander, G. Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In Proceedings of the Semantic Web: Semantics and Big Data, Montpellier, France, 26–30 May 2013; pp. 593–607.
17. Dragoni, M.; Tonelli, S.; Moretti, G. A Knowledge Management Architecture for Digital Cultural Heritage. *J. Comput. Cult. Herit.* **2017**, *10*, 3. [CrossRef]
18. Meijer, E.; de Valk, S. A Distributed Network of Heritage Information. Available online: https://github.com/netwerk-digitaal-erfgoed/general-documentation/blob/master/Whitepaper%20A%20distributed%20network%20of%20heritage%20information.md (accessed on 15 June 2019).
19. Vander Sande, M.; Verborgh, R.; Hochstenbach, P.; Van de Sompel, H. Towards sustainable publishing and querying of distributed Linked Data archives. *J. Documentation* **2018**, *74*, 195–222. [CrossRef]
20. Freire, N.; Manguinhas, H.; Isaac, A.; Robson, G.; Howard, J.B. Web technologies: A survey of their applicability to metadata aggregation in cultural heritage. *Inf. Serv. Use J.* **2018**, *37*, 4.
21. Freire, N.; Robson, G.; Howard, J.B.; Manguinhas, H.; Isaac, A. Metadata Aggregation: Assessing the Application of IIIF and Sitemaps within Cultural Heritage. In Proceedings of the Research and Advanced Technology for Digital Libraries, Thessaloniki, Greece, 18–21 September 2017.
22. Freire, N.; Charles, V.; Isaac, A. Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata. In Proceedings of the Semantic Web (ESWC 2018), Heraklion, Crete, Greece, 3–7 June 2018.
23. Google Inc.; Yahoo Inc. Microsoft Corporation and Yandex, "About Schema.org", n.d. Available online: http://schema.org/docs/about.html (accessed on 29 May 2019).
24. 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.* **2017**, *36*, 12330.
25. Freire, N.; Meijers, E.; Voorburg, R.; Isaac, A. Aggregation of cultural heritage datasets through the Web of Data. *Procedia Comput. Sci.* **2018**, *137*, 120–126. [CrossRef]
26. Alexander, K.; Cyganiak, R.; Hausenblas, M.; Zhao, J. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note. Available online: https://www.w3.org/TR/void/ (accessed on 29 May 2019).
27. Maali, F.; Reikson, J. Data Catalog Vocabulary (DCAT). W3C Recommendation. Available online: https://www.w3.org/TR/vocab-dcat/ (accessed on 29 May 2019).