Elements for Constructing a Data Quality Policy to Aggregate Digital Cultural Collections: Cases of the Digital Public Library of America and Europeana Foundation

Joyce Siqueira[1], Danielle do Carmo[1], Dalton Lopes Martins[1](✉), Daniela Lucas da Silva Lemos[2], Vinicius Nunes Medeiros[3], and Luis Felipe Rosa de Oliveira[1]

[1] Universidade de Brasília, Brasilia, DF, Brazil
{joyce.siqueira,danielle.carmo,rosa.luis}@aluno.unb.br, daltonmartins@unb.br

[2] Universidade Federal do Espírito Santo, Vitoria, ES, Brazil
daniela.l.silva@ufes.br

[3] Tainacan.org, Brasilia, DF, Brazil
viniciusmedeiros@inf.ufg.br

Abstract. Institutions around the world have sought to aggregate different cultural heritage data sources in order to provide society with comprehensive and useful services. This qualitative exploratory study presents a comparative analysis of two paradigmatic institutional aggregators, namely the Europeana Foundation in the European Union and the Digital Public Library of America in the United States. To that end, strategic aggregation documents were identified and quality policy elements analyzed and compared. As a result, nine quality-oriented data aggregation elements were selected: data providers; application process; metadata model; data exchange agreement; copyright license; call for applications; metadata use; technical criteria for data quality and data validation and publication. The elements identified and described are important in formulating processes that make it possible to aggregate digital cultural heritage collections from different Brazilian institutions and provide support for the solution currently being developed in collaboration with the Brazilian Institute of Museums (Ibram).

Keywords: Cultural data aggregation · Data quality policy · Institutional aggregator

1 Introduction

The online availability of cultural collections in the form of digital catalogs and repositories as well as their digitization and socialization on the internet (Scopigno et al. 2017; Potenziani et al. 2018; Medeiros e Sá et al. 2019) are noteworthy phenomena in contemporary times, and have intensified with the increased use of the internet as tool for work, leisure and socialization.

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 106–122, 2021.
https://doi.org/10.1007/978-3-030-77417-2_8

Initiatives to integrate different data sources are aimed at making better use of these technologies and providing society with comprehensive and useful services. However, it is important to consider that in order to aggregate data from different sources, they must first be organized to ensure favorable conditions for diversified high-quality analyses, integrated searches and possibilities for reuse, with a view to constructing new informational services and products.

The term "organized" refers to the minimal organization of data structure. Data structure is the set of metadata used to describe an informational object, the cataloging rules used to populate these metadata and the standards and controlled vocabularies applied to index and classify objects based on a common framework (Gilliland 2016; Zeng and Qin 2016; Abbas 2010).

Important contemporary initiatives such as Findability, Accessibility, Interoperability and Reuse (FAIR) (Wilkinson et al. 2016) are aimed at good data management practices consistent with the concept of well-structured data and establishing quality standards. To some extent, digital repository data described based on these principles can be considered quality data.

Despite the different motivations behind the search for objective criteria to analyze data quality, it is directly linked to the intention to reuse data, generating value based on the possibility of connecting it with other databases and providing an opportunity to produce knowledge.

This raises the following questions: What is the minimum metadata structure needed for aggregation? How can data quality be measured? What are the criteria that define a quality digital object?

What determines whether data meet the minimum requirements for aggregation and reuse and what can be done to ensure they reach these objectives? The responses to these and other questions should form part of the data quality policies of institutions aimed at assessing datasets from different sources in order to aggregate them.

For the purposes of the present study, a data quality policy is a set of technical recommendations to guide institutions interested in aggregating their data (DPLA 2020c; Scholz 2019). Following the policy improves user experience and helps create comprehensive coherent collections (DPLA 2020c). It also encourages partners to not only submit a minimum of metadata and content quality, but to aim for rich metadata and the highest possible data quality (Scholz 2019).

It is important to underscore that different areas of knowledge have shown interest in data aggregation techniques and the possibilities they provide, with research confirming that the issue permeates the scientific world.

Solutions were found in the areas of bioinformatics, geography, medicine, music and culture (SIQUEIRA and MARTINS 2021), among others. Examples of integrated data in different fields include genetic expression data on the fruit fly (*Drosophila melanogaster*) to select genes and validate experimental results in bioinformatics (Miles et al. 2010); geographic location data to assess water license requests in geography (Ziébelin et al. 2017); input forms and information systems in epidemiological studies to determine the cause of diseases such as obesity, depression and dementia in medical sciences (Kirsten et al. 2017); data to provide a single web portal to access a collection of historical sources on the Mediterranean from around 15 different epigraphic archives (Mannocci et al. 2014). Although only these four areas are highlighted, each of them clearly needs to manage a wide range of data in a unique way.

The aim of the present study is to find solutions in the field of culture and forms part of wider scientific initiative involving researchers from different academic institutions, who are collaborating with the Brazilian Institute of Museums (Ibram) in an important project to digitize and disseminate the collections of Brazilian museums.

Ibram is directly responsible for managing 30 museums[1] and, in 2016, adhered to the Tainacan platform, "a powerful and flexible repository platform for the WordPress tool that allows users to manage and publish digital collections" (GOV.BR 2020). It has since implemented the system in its museums as a strategy to provide online access to digital museum collections.

An important issue for the project is the ability to generate an integrated search service that allows users to access a single website in order to search for and retrieve information on all the museums that have provided their data in an open digital repository online.

As part of the intermediate phase in the scientific initiative to provide this service, the researchers involved have studied international facilities that perform similar tasks and can provide supporting frameworks and technical recommendations. The services selected were the Digital Public Library of America and Europeana Foundation, identifying the documents, procedures, criteria and fundamental stages of data aggregation, with a focus on measuring the quality of the data collected.

As such, the aim of this study is to propose essential elements for a data quality policy that contributes to aggregation procedures for online collections, with a view to: i) identifying and selecting strategic documents from the institutions investigated; and ii) analyzing and comparing the quality policy elements mapped here. In our view, these aspects are potential elements in formulating aggregation processes for collections.

The article is organized as follows: Sect. 2 describes the institutions and the methodological procedures used to collect and analyze data; Sect. 3 presents the results of the study, clarifying the elements identified and deemed relevant in a data quality policy aimed at aggregation; Sect. 4 discusses and compares how these elements are used in the context studied; and finally, Sect. 5 identifies important aspects for further research.

2 Methodology

This is a qualitative descriptive exploratory study involving two renowned institutions that aggregate cultural data, the Digital Public Library of America and Europeana Foundation, with a comparative analysis of their data quality policies in a case study design.

The Digital Public Library of America is a nonprofit organization aimed at amplifying the value of libraries and cultural organizations as Americans' most trusted sources of shared knowledge. Its integrated search portal DPLA - https://dp.la/ (DPLA 2020h) provides millions of materials from libraries, archives, museums and other cultural institutions across the country in a single point of access (DPLA 2020a).

[1] IBRAM, Ibram Museums, http://www.museus.gov.br/os-museus/museus-ibram/, last accessed on 10/05/2020.

The Europeana Foundation is an independent nonprofit organization that operates the Europeana Collections platform at https://www.europeana.eu/pt (EUROPEANA 2020b), with millions of items from European archives, libraries and museums, providing access to books, music, artworks and other materials via sophisticated research tools (EUROPEANA 2020a; EUROPEANA PRO 2020b).

These institutions were selected as paradigms in the field and for geopolitical reasons, since the DPLA covers the entire United States and the Europeana Foundation all of Europe. Also considered was the fact that the Europeana Foundation encompasses more than one country and several different languages while the DPLA deals with only one country and language; the relevance of the former as a data aggregation service, particularly because of its direct relationship with some of the world's most renowned cultural institutions such as the Louvre Museum and British Library; and the significant differences in the processes described for the two institutions, with the DPLA using simpler and more manual procedures than those of the Europeana Foundation.

Initial exploratory analysis also found that the DPLA model is far more compatible with the Brazilian reality than that of the Europeana Foundation.

The technical data collection and analysis methods used were a literature review and documentary analysis. The former was used to substantiate concepts and provide theoretical sustainability for the study and the latter to provide access (via the institutional websites) to documentary sources, including manuals, tutorials, video demonstrations, educational material and other content aimed specifically at supporting institutions that want to become data providers.

Documents from the two institutions selected were analyzed and interpreted using content analysis, a set of techniques aimed at obtaining indicators (quantitative or not) that enable inferences to be made regarding knowledge present in the material assessed.

The techniques involved include defining categories (shown in Table 1) to support content analysis of the selected materials. According to Bardin (2016), categories are established based on the following criteria: semantics (themes); syntax (verbs, adjectives, pronouns); lexicons (grouping synonyms or antonyms, or according to the meaning of words); and expression (grouping writing or linguistic deviations).

The criterion used in the present study was semantics, establishing themes related to the fields of information and technology. The following analysis categories (individual and comparative) were defined:

- Data providers: individual or institutional aggregators that submit data for aggregation.
- Application process: the means provided by the institutional aggregator to initiate the aggregation process.
- Metadata model: formal specification of the data model, listing the classes and properties of metadata.
- Data exchange agreement: Document governing the partnership between the aggregator and data provider.
- Copyright licenses: list of copyright licenses pre-selected by the aggregator.
- Metadata use: minimum metadata requirements for aggregation.
- Technical criteria for data quality: defining quality criteria for metadata content and/or digital media.

- Call for applications: specifications on how collections are structured to keep aggregators informed.
- Data validation and publication: prototype of aggregated data that must be validated by the provider in order to be published.

It is important to note that categorization was essential in mapping and understanding the workflow stages involved in quality assessment and data aggregation by the two institutions.

Finally, data organization and management (the final stage of the method) made it possible to obtain a conceptual theoretical corpus based on the empirical objects investigated and draw theoretical and methodological conclusions regarding the underlying quality policy of both institutions.

Table 1 presents the supporting documents for quality assessment and data aggregation that guided this study. The list of documents and their purpose is an important tool for future research.

Table 1. Documentation on metadata and data quality.

Description	DPLA	Europeana
Data model	*Metadata Application Profile* v5.0 (Gueguen et al. 2017)	*Europeana Data Model v2.3* (Clayphan et al. 2016)
Data exchange agreement	*Data Exchange Agreement* (DPLA 2017a)	Europeana, o *Europeana Data Exchange Agreement* (EUROPEANA FOUNDATION 2020)
Metadata and content quality requirements	Partnering with DPLA Metadata (DPLA 2020g); DPLA Metadata Quality Guidelines (DPLA 2020i) and DPLA Geographic and Temporal Guidelines for MAP 3.1 (DPLA 2017b)	Europeana Publishing Guide v1.8 (Scholz 2019) and Europeana Publishing Framework 2.0 (EUROPEANA FOUNDATION 2019)
Copyright licenses	DPLA Standardized Rights Statements Implementation Guidelines (DPLA 2017c)	Europeana Licensing Framework (EUROPEANA THINK CULTURE 2011)
Guidelines for content development and management	–	Europeana Content Strategy (Scholz et al. 2017)
User scenarios and their metadata requirements	–	Discovery - User scenarios and their metadata requirements v3 (Charles et al. 2015)
Guidelines on the material to be collected	Collection Development Guidelines (DPLA 2020c)	–

The DPLA lists all the useful documents and other tools for data providers at https://pro.dp.la/hubs/documentation (DPLA 2020d) and the Europeana Foundation at https://pro.europeana.eu/share-your-data/process (EUROPEANA PRO 2020c).

3 Results

The nine stages obtained are illustrated in order in Fig. 1 for each institution.

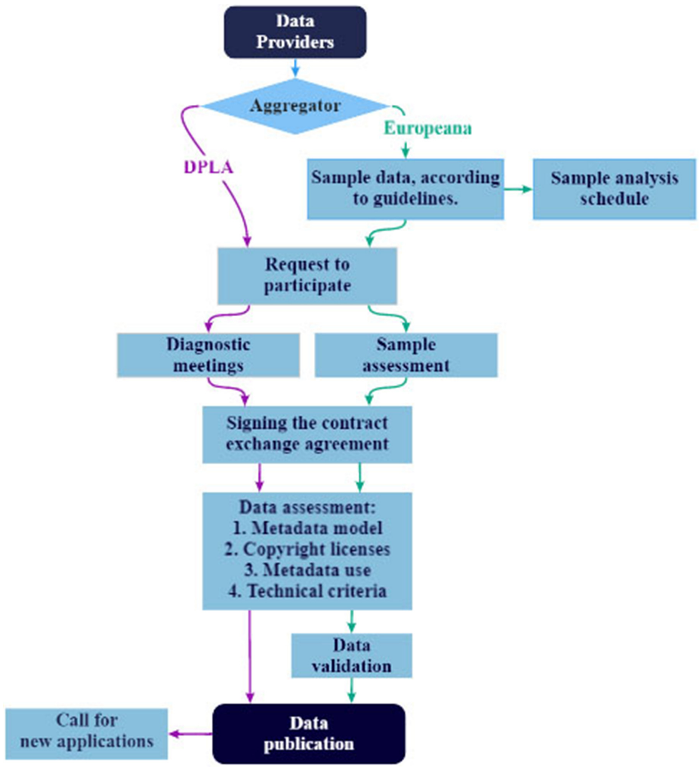

Fig. 1. Data aggregation stages, with a focus on data quality

The nine stages are described below: data providers; application process; metadata model; data exchange agreement; copyright licenses; call for applications; metadata use; technical criteria for data quality and; data validation and publication.

3.1 Data Providers

Data providers are institutions that submit their collections for aggregation. These consist of individual providers that submit their data directly to the institutional aggregator; and aggregators, who aggregate data from individual providers, typically grouped according to a specific niche such as region or theme, and submit the collections to the institutional aggregator. Both the DPLA and Europeana Foundation receive data from both types of providers based on their own specific criteria.

The DPLA currently has 43 data providers, 29 of which are aggregators, as well as around 41 million items from cultural institutions across the United States. Europeana has 38 data providers, all of which are aggregators, and over 50 million digitized items on European cultural heritage. The complete lists of data providers for both institutions are available at https://pro.dp.la/hubs/our-hubs (DPLA 2020f) and https://pro.europeana.eu/page/aggregators (EUROPEANA PRO 2020a).

3.2 Application Process

Prospective data providers for the DPLA must submit their request by email and will then be contacted by team members at the institutions for a series of mandatory conversations. Aggregators are then asked to fill out a form that diagnoses them as data partners. Once accepted, new partners will be notified and invited to review the next steps of the process: data exchange agreement, membership agreement and the metadata, tech and content information form (DPLA 2020b).

Europeana requires applicants to send an email to its team, who will then request a compacted data sample in eXtensible Markup Language format (XML)[2] or via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[3]. The data must be mapped in accordance with the Europeana Data Model (EDM) and are then automatically checked and validated against the Europeana Publishing Guide (Scholz 2019), using the Metis processing tool available from GitHub (GITHUB 2020).

Both institutions follow a metadata model, described in the next section. However, it should be noted that prospective Europeana data providers must have prior knowledge of the required documentation and data model, whereas the DPLA assists data providers in this process after they have been accepted, identifying which original metadata are compatible with its Metadata Application Profile (MAP) via a Google Drive spreadsheet (DPLA 2020e).

3.3 Metadata Model

Each institution has its own metadata model, which is vital in ensuring that an aggregation standard is in place. The Europeana Foundation developed the Europeana Data Model (EDM), a theoretical data model that allows data to be presented in accordance with the practices of different domains that contribute to the Europeana Collection (Clayphan et al. 2016).

The DPLA's Metadata Application Profile (MAP) is the basis for data structuring and validation and guides how they are stored, serialized and published. The MAP was designed based on the EDM (Gueguen et al. 2017).

---

[2] Markup language to facilitate online information sharing.
[3] Protocol for collecting metadata records in repositories.

3.4 Data Exchange Agreement

Once accepted, data providers must officially enter into a partnership with the institutional aggregator. An important part of this process is signing a data exchange agreement, which establishes the rules governing the partnership,

namely the DPLA's Data Exchange Agreement (DPLA 2017a) and Europeana Foundation's Europeana Data Exchange Agreement (EUROPEANA FOUNDATION 2020).

The most noteworthy element of these agreements is the need for the data provider to guarantee that their aggregated metadata and media are covered by a copyright license.

3.5 Copyright Licenses

As mentioned above, copyright licenses are a fundamental part of the data aggregation process because they govern how the end user will access and reuse the digital object. The institutions studied here adopt a similar stance in relation to copyrights.

In regard to metadata, both institutions require that these be submitted under a CC0 1.0 Universal Public Domain Dedication license[4]. For media, the DPLA lists 12 standardized rights statements and recommendations for copyrights (DPLA 2017a) and the Europeana Foundation 14 (EUROPEANA PRO 2015) .

The licenses are briefly analyzed to determine whether the metadata license used allows unrestricted reuse, while access and reuse of media s are graded on different levels.

3.6 Call for Applications

The DPLA does not have a well-defined schedule for including new data providers, but provides an online calendar (DPLA 2021) with new collection dates.

Europeana advises that initial data collection is performed on a first-come-first-serve basis; however, thematic collections, metadata and content quality, user demands and business priorities define which collections will receive more attention. The data are processed and published continuously throughout the year.

3.7 Metadata Use

The metadata schemas used by the two institutional aggregators cover a wide range of fields, but the present study focuses on the following categories: mandatory, mandatory when present, mandatory based on a predefined metadata group and; recommended (Table 2).

Metadata descriptions can be found in documents on the MAP (Gueguen et al. 2017) and EDM metadata models (Clayphan et al. 2016).

The DPLA lists 11 metadata elements, four **required**, four **required when present** and three **recommended**, while Europeana lists 14, six **required** and four **required**

[4] CREATIVE COMMONS. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication (2020), https://creativecommons.org/publicdomain/zero/1.0/, last accessed on 10/16/2020.

Table 2. Metadata submission categories

No.	Metadata	DPLA	Europeana
1	Subject	SourceResource.subject[d]	dc:subject[c]
2	Location	SourceResource.spatial[d]	dcterms:spatial[c]
3	Type or genre	–	dc:type[c]
4	Time	–	dcterms:temporal[c]
5	Date created	SourceResource.date[d]	–
6	Copyright license	SourceResource.rights[a]	edm:rights e sua URI
7	Unique identifier	–	edm:ProvidedCHO, edm:WebResource, ore:Aggregation[a]
8	Thumbnail link	Aggregation.object[b]	–
9	Link to the original record	Aggregation.isShownAt[a]	edm:isShownBy ou edm:isShownAt[a]
10	Metadata populated	–	edm:ugc = true[b]
11	Name of the institution	Aggregation.dataProvider[a]	edm:dataProvider[a]
12	Name of the aggregator	Aggregation.provider[b]	edm:provider[b]
13	Textual objects	–	dc:language[b]
14	Type of resource	SourceResource.type[b]	edm:type[b]
15	Collection title	Collection.title[b]	–
16	Title	SourceResource.title[a]	dc:title[c]
17	Description	–	dc:description[c]

[a]Required
[b]Required when present
[c]Required, based on a predefined metadata group
[d]Recommended

when present. The difference in numbers is because the Europeana Foundation organizes metadata submissions by group, only one of which is mandatory, as is the case for *dc:title* and *dc:description.*

The next section addresses the quality of metadata and media content.

3.8 Technical Criteria for Data Quality

In additional to minimum metadata requirements, it is vital that the data submitted exhibit quality content, including digital media. Thus, in order to advise data providers on the quality of their own metadata and media and how to improve them when necessary, both the DPLA and Europeana provide technical guidelines for institutions interested in aggregating their data.

DPLA guidelines are developed around three tiers of recommendations for each property in the MAP, whereas the more complex Europeana framework involves two sets of tiers, one for metadata quality and the other for media quality.

The three tiers of the DPLA Metadata Quality Guidelines (DPLA 2020i) are minimal quality requirements, recommendations for improved quality and recommendations for best quality.

Guidelines under the "minimal quality requirements" tier are considered necessary for data aggregation and typically relate to correct use of the property or granularity issues. Although each property may or may not have minimal requirements, properties included in the record must follow the guidelines. For example, while there is no minimal quality requirement for the mandatory metadata property *title*, the optional *alternative title* property has a minimal requirement of not being a copy of the main title, although translated titles are acceptable.

The "recommendations for improved quality" tier is a set of guidelines aimed at improving data within the aggregation context, creating better overall consistency through actions such as the use of authorities or content standards. For example, for the optional metadata property *creator,* the use of name authority files such as LCNAF (Library of Congress Names[5]), VIAF (The Virtual International Authority File[6]) and ULAN (*Union List of Artist Names*[7]) is recommended. Another example is the optional *extent* property*,* with the recommended use of content standards such as CCO (Cataloging Cultural Objects[8]), RDA *(*Resource Description and Access[9]) and DCRM (Descriptive Cataloging of Rare Materials[10]).

Finally, the "recommendations for best quality" guidelines are related to the use of Uniform Resource Identifiers (URIs), a unique identifier that enables the use of Linked Open Data (LOD)[11]. Since linked data is still a relatively new concept in libraries, archives and museums, it is reserved for the highest quality tier (DPLA 2020b, p. 1).

Data quality and requirements for the Europeana Foundation are presented in two main documents, the Europeana Publishing Guide (Scholz 2019), which provides technical guidelines for data aggregation, and the Europeana Publishing Framework (EUROPEANA FOUNDATION 2019), whose primary objective is to inform data providers of the benefits of submitting quality metadata and media.

5 LIBRARY OFCONGRESS. LC Name Authority File (LCNAF), https://id.loc.gov/authorities/names.html, last accessed on 10/19/2020.

6 VIAF. The Virtual International Authority File, http://viaf.org/, last accessed on 10/19/2020.

7 GETTY. Union List of Artist Names, https://www.getty.edu/research/tools/vocabularies/ulan/index.html, last accessed on 10/19/2020.

8 VRA. Cataloging Cultural Objects (CCO), https://vraweb.org/resources/cataloging-cultural-objects/, last accessed on 10/19/2020.

9 RDA. Resource Description and Access, https://www.rdatoolkit.org/about, last accessed on 10/19/2020.

10 RBMS. Descriptive Cataloging of Rare Materials, http://rbms.info/dcrm/dcrmmss/, last accessed on 10/19/2020.

11 LOD is a set of best practices for publishing and interlinking structured data on the web, establishing links between items from different sources to form a global data space (HEATH and BIZER, 2011 apud Santarém Segundo 2015).

Europeana provides guidelines for metadata and media. How data are published by data providers is classified into tiers A, B and C, with A the lowest tier and C the highest. Metadata tier A represents Europeana Collections as a basic search platform; B as an exploration platform and C as a knowledge platform.

When a data provider's metadata fall under tier A, it can use Europeana Collections as a basic search platform, that is, users can find specific objects if they have the precise information they need, with the aid of search filters. In order to meet this requirement, at least 25% of the metadata elements provided and considered relevant must have at least one language tag. Additionally, the metadata should include at least one of the enabling elements taken from the discovery and user scenarios.

The discovery and user scenarios were developed by the Data Quality Committee (DQC) and reflect user needs when accessing information, listing problems and proposing actions. The goal of these scenarios is to demonstrate specific situations to users and encourage them to provide increasingly comprehensive metadata by including enabling elements (optional metadata elements) that support specific user scenarios, with a view to ensuring greater information retrieval and better services in the collections (Charles et al. 2015).

Metadata provided under tier B allow the providing institutions to use Europeana Collections as an exploration platform, whereby users can find objects by searching for both general and specific information. This allows collections to be contextualized and discovered in thematic collections, with findability enhanced by the possibility of multilingual searches, and presented on the resource pages of the entities featured on the platform. To that end, at least 50% of the relevant metadata elements provided must have at least one language tag. The metadata should also include at least three enabling elements covering at least two distinct discovery and user scenarios and one contextual class with all the minimum required elements or a link to LOD vocabulary.

Contextual classes give meaning to objects and make them easier to retrieve. The EDM includes four contextual classes, edm:Agent, edm:Place, edm:TimeSpan and skos:Concept, which can be used to capture distinct entities that are related to the cultural heritage objects. The use of references to multilingual vocabularies and LOD is recommended, including the Getty Art & Architecture Thesaurus (AAT), Wikidata or Geonames (Charles et al. 2015).

Tier C metadata allow the data provider to use Europeana Collections as a knowledge platform. This enables users to find the object they are looking for more accurately without needing specific information on it because the contextual metadata provided generate a linked network of knowledge so that connections can be discovered through the relationships between them and their identities. Metadata can therefore be used in partner projects such as Historiana[12] (for the educational sector) and CLARIN[13] (research infrastructure project) and others involving creative industries. This contextualized and enriched information can also be used in classes, studies and applications, ensuring greater reach for the collections within and beyond the Europeana Collections platform. To make this possible, at least 75% of the relevant metadata elements provided

[12] HISTORIANA. Historiana, https://historiana.eu/, last accessed on 10/19/2020.

[13] CLARIN. European Research Infrastructure for Language Resources and Technology, https://www.clarin.eu/, last accessed on 10/19/2020.

must have at least one language tag. The metadata should include at least three enabling elements covering a minimum of two distinct discovery and user scenarios and two different contextual classes with all the minimum required elements or links to LOD vocabularies.

The above information clearly shows that both aggregators (DPLA and Europeana) stipulate the lowest tiers as minimum network entry requirements, but also guide institutions on how to progressively manage their data to ensure better quality, thereby exploiting search resources more efficiently and discovering information provided by the platforms.

The publication of media submitted by data providers is classified into tier 1 for Europeana Collections as a search engine, 2 as a showcase; 3 as a distribution platform for reuse and 4 as a free reuse platform.

The tiers rate the possibility of reusing digital content, with tier 1 representing the lowest possible use and 4 the highest.

Digital objects compliant with tier 1 are sufficient to allow data providers to use Europeana Collections as a search engine so that users can find their collections. The digital object is visualized via the website of the original institution and not the Europeana Collections environment.

Tier 2 data means the Europeana Collections platform can be used as a showcase for the data provider's collections, enabling users to visualize high-quality versions of the collections. The objects can also be included in one of Europeana's thematic collections, providing greater exposure.

Data that fall under quality tier 3 allow providers to use Europeana Collections as a distribution platform for noncommercial reuse, meaning users can view high-quality versions of digital objects directly on the platform and incorporate their collections in existing projects such as Historiana or CLARIN. This is made possible by rights statements that authorize some forms of reuse, allowing collections to be used on websites, applications and other noncommercial services.

Tier 4 enables data providers to use Europeana Collections as a free reuse platform, whereby data related to collections can be incorporated in projects and partnerships with creative industries; shared on social media under the hashtag #OpenCollections; used on open platforms such as those of the Wikimedia Foundation and in hackathons. To that end, the collection must be published under rights statements that allow free reuse, enabling it to be used on websites, applications and commercial and noncommercial services.

The specific requirements for each type of media (image, audio, video and 3D) are described in detail in the Europeana Publishing Guide (Scholz 2019).

3.9 Data Validation and Publication

At the end of the process, Europeana provides applicants with a link to view their data beforehand and authorize its publication or determine whether changes are needed (Scholz 2019). The DPLA does not offer this type of service.

4 Discussion

According to the data obtained, the DPLA's aggregation processes are simpler and more manual than Europeana's complex procedures. An example is that Europeana requires sample data to be submitted in line with its available documentation and initially automatically validates the data via Metis software, whereas the DPLA maps metadata manually using an electronic spreadsheet after the institution has been accepted as a data provider.

Europeana requirements are stringent and their stages complex, which are positive points when considering the pedagogical effects for the field because they establish paths to follow, but also generate added difficulty for interested data providers. The process appears to require a certain level of technical and administrative maturity on the part of providers interested in publishing their data online.

The DPLA's data model (MAP) focuses on a common interface to describe media by reflecting information on content, whereas Europeana's EDM model differs in the description underlying the content (or describing knowledge about the resource) through semantic technologies, particularly semantic vocabularies available in the LOD environment, with a view to the formal and explicit publication of knowledge on the internet.

This allows data providers that use aggregation services such as Europeana Collections to expand search opportunities for their users on knowledge networks semantically linked to web data. This includes multimedia research laboratories, creative industries, the cultural and education sectors, and social media such as Wikidata, enabling information within and beyond the aggregation platform to be explored in greater depth.

Europeana operates in a scenario that assumes greater organizational maturity in its partners' information management processes. It is important to note that the institution essentially relies on national aggregators in the respective countries to collect data, as well as some specific thematic projects, such as in the field of cinema. As a result, the data structure and quality parameters governing its initial submission processes are more stringent, whereas the DPLA differs significantly in that it helps providers map their data according to the required model, providing technical and institutional support on how to manage and submit their data in a suitable format for aggregation. This pedagogical and supportive effect is particularly important in the present study.

Both institutions provide recommendations on metadata quality within an aggregation context and a general metadata framework for application in machine-readable media content, but their frameworks differ in terms of modelling multimedia information. Despite their similar vision on metadata, there are obvious differences in their semantic proposals, with Europeana clearly expressing its goal of collaborating and becoming a data provider in open linked data semantic networks, while the DPLA does not seem to explicitly share this objective.

It is also important to underscore that Europeana indicates possible uses for aggregated data in more complex scenarios, developing and establishing partnerships with thematic data reuse projects and relationships with knowledge bases such a Wikidata. It can undoubtedly be inferred that this greater complexity in managing and aggregating data can provide important benefits, resulting in higher quality data with better potential for reuse in future applications and services.

In regard to the minimum metadata requirements for aggregation in the different mandatory categories, the DPLA also exhibited a small dataset, required only 11 fields compared to Europeana's 14. One of the fundamental differences between the two institutions was the language requirement, heavily emphasized by Europeana in a number of points, but less so by the DPLA. The fact that Europeana covers different countries and official languages makes this factor more complex, generating demands that will translate into elements of the data model.

5 Final Considerations

The results obtained indicate advances in data aggregation research in terms of providing metadata models aimed at describing digital objects and multimedia online. As observed in the present study, the two aggregator institutions studied view metadata as the product of cataloging rules and consider it vital in describing informational resources in order to expand access points and improve the management, organization and retrieval of digital objects online.

The study shows the different stages and processes involved in compiling a data aggregation policy to develop a search and retrieval service in the area of culture. The different stages were analyzed, highlighting their technical characteristics and the policies involved, and clearly demonstrating that the process includes phases of dialogue, agreements, meetings, validation and necessary adjustments to achieve the aggregation result. It is important to note that this process could not be fully automated, with human intervention essential in certain stages when decisions and adjustments are needed. The important role of documentation experts such as librarians, museologists and archivists in projects of this type is also evident.

One of the most noteworthy points observed throughout the study is the pedagogical role of these initiatives in establishing levels of quality and data assessment. For institutions without the necessary training to produce data that meet these quality criteria, these initiatives establish objective parameters that serve as an inspiration and benchmark to improve the quality of their data. This is considered a significant shortcoming in the field of museology in Brazil and possibly even in other information-related fields.

It is hoped that this type of research will provide the necessary direction and initiatives to qualify documentation efforts and establish better criteria for institutions that want to publish their data on digital repositories for subsequent collection and participation in different information networks. These efforts not only improve the data quality of providing institutions, but the quality of cultural institution websites themselves.

The objectives of the study were achieved, identifying the relevant documents and referencing those that are essential in explaining the rules, guidelines and standards adopted by the institutions in their efforts at aggregation. It is important to note that these documents provide significant guidance for researchers and students interested in the issue as well as a framework that presents them systematically in order of their importance and role in the services established. The secondary objective of analyzing and comparing the elements of the two institutions' quality policies was also met, with the difference and similarities clearly identified. In general, the DPLA's process is simpler and more collaborative, providing more direct support for data providers in adjusting and preparing their data for aggregation. The Europeana Foundation gives providers greater autonomy in the submission process and has stricter, more complex requirements, but also offers broader possibilities for data reuse and greater integration with knowledge bases generated based on semantic web principles.

In terms of gathering subsidies and information that contribute to compiling a data quality policy for the aggregation of digital repositories from museums managed by the Brazilian Institute of Museums, the study conducted a detailed analysis of the processes of both institutions and identified elements to serve as inspiration and a basis for reflection in the Brazilian initiative. It is hoped that part of the results can be used in future research in the field of Brazilian museums.

References

Abbas, J.: Structures for Organizing Knowledge: Exploring Taxonomies, Ontologies, and Other Schema. Neal-Schuman Publishers, New York (2010)

Bardin, L.: Análise de Conteúdo. 3ª Reimpressão da 1, vol. 70. Edições, São Paulo (2016)

Charles, V., Isaac, A., Hill, T. (ed.): Discovery - User scenarios and their metadata requirements v3. Europeana Foundation (2015). https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee/DQC_DiscoveryUserScenarios_v3.pdf. Accessed 23 Sept 2020

Clayphan, R., Charles, V., Isaac, A.: Europeana Data Model – Mapping Guidelines v2.3. Europeana Think Culture (2016). https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.3_112016.pdf. Accessed 23 Sept 2020

DPLA: about us (2020a). https://dp.la/about. Accessed 28 Sept 2020

DPLA: Becoming a Service Hub (2020b). https://pro.dp.la/prospective-hubs/becoming-a-service-hub. Accessed 05 Oct 2020.

DPLA: Collection Development Guidelines (2020c). https://pro.dp.la/hubs/collection-development-guidelines. Accessed 23 Sept 2020

DPLA: Documentation and Tools (2020d). https://pro.dp.la/hubs/documentation. Accessed 27 Sept 2020

DPLA: DPLA Data Exchange Agreement (2017a). https://digitalpubliclibraryofamerica.atlassian.net/wiki/download/attachments/85406202/DataExchangeAgreement2017March.docx?version=1&modificationDate=1494427927703&cacheVersion=1&api=v2. Accessed 16 Oct 2020

DPLA: DPLA Geographic and Temporal Guidelinesfor MAP 3.1 (2017b). https://docs.google.com/document/d/1b2iJI90I24hUp-8kCfnZhAQcefBt0vPUjMzhDn9HOZ0/edit#heading=h.s8e27gcgg4cl. Accessed 23 Sept 2020

DPLA: DPLA Metadata Quality Guidelines. Digital Public Library of America (2020i). https://docs.google.com/document/d/1dITqEYEWsMX1a2pLPmkL78k1LN2b4im03spn8_QFscY/edit#heading=h.8lontuvk72d8. Accessed 23 Sept 2020

DPLA: DPLA Partners Crosswalk, MAP v4.0 (2020e). https://docs.google.com/spreadsheets/d/1BzZvDOf4fgas3TD21xF40lu2pk2XW0k2pTGJKIt6438/edit#gid=1453046017. Accessed 12 Oct 2020

DPLA: DPLA Standardized Rights Statements Implementation Guidelines (2017c). https://docs.google.com/document/d/1aInokOIIsgf-B4iMTXU33qYN5B2jA3s91KgWoh7DZ7Q/edit#heading=h.ma4e226diad0. Accessed 23 Sept 2020

DPLA: Our Hubs (2020f). https://pro.dp.la/hubs/our-hubs. Accessed 16 Oct 2020

DPLA: Partnering with DPLA Metadata (2020g). https://docs.google.com/document/d/1gshwQ0Oj84l5q-_JxHo7wjyLeXln9lCg6SUtIMkPu_g/edit. Accessed 23 Sept 2020

DPLA: Portal Digital Public Library of America (2020h). https://dp.la/. Accessed 16 Oct 2020

DPLA. Hub Re-ingest Schedule (2021). https://digitalpubliclibraryofamerica.atlassian.net/wiki/spaces/CT/pages/84969744/Hub+Re-ingest+Schedule. Accessed 27 May 2021

Europeana: about us (2020a). https://www.europeana.eu/en/about-us. Accessed 28 Oct 2020

Europeana: Europeana Collections (2020b). https://www.europeana.eu/pt. Accessed 16 Oct 2020

Europeana Foundation: Europeana Data Exchange Agreement (2020). https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Data%20Exchange%20Agreement.pdf. Accessed 27 Sept 2020

Europeana Foundation: the more you give, the more you get. Europeana Publishing Framework 2.0 (2019). https://pro.europeana.eu/files/Europeana_Professional/Publications/Publishing_Framework/Europeana%20Publishing%20Framework%20V2.0%20English.pdf. Accessed 27 Sept 2020

Europeana Pro: available rights statements (2015). https://pro.europeana.eu/page/available-rights-statements. Accessed 27 Sept 2020

Europeana Pro: Europeana Aggregators (2020a). https://pro.europeana.eu/page/aggregators. Accessed 16 Oct 2020

Europeana Pro: Europeana Foundation, About us (2020b). https://pro.europeana.eu/about-us/foundation#europeana-foundation. Accessed 28 Sept 2020

Europeana Pro: Share your data (2020c). https://pro.europeana.eu/share-your-data/process. Accessed 27 Sept 2020

Europeana Think Culture: Europeana Licensing Framework. Institute for Information Law, the Bibliothèque nationale de Luxembourg and Kennisland in cooperation with Europeana (2011). https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Licensing%20Framework.pdf. Accessed 23 Sept 2020

GITHUB: Europeana - Metis Framework (2020). https://github.com/europeana/metis-framework. Accessed 12 Oct 2020

Gilliland, A.J.: Setting the stage. In: Baca, M. (ed.) Introduction to Metadata, 3 edn. Getty Research Institute, Los Angeles (2016)

GOV.BR: Projeto Tainacan. Ministério do Turismo - Instituto Brasileiro de Museus (Ibram) (2020). https://www.museus.gov.br/acoes-e-programas/projeto-tainacan/. Accessed 12 Oct 2020

Gueguen, G., et al.: Metadata Application Profile, version 5.0. DPLA - Digital Public Library of America (2017). https://drive.google.com/file/d/1fJEWhnYy5Ch7_ef_-V48-FAViA72OieG/view. Accessed 23 Sept 2020

Kirsten, T., Kiel, A., Rühle, M., Wagner, J.: Metadata Management for Data Integration in Medical Sciences. Datenbanksysteme für Business, Technologie und Web (2017). https://dl.gi.de/handle/20.500.12116/627. Accessed 05 Oct 2020

Mannocci, A., Casarosa, V., Manghi, P., Zoppi, F.: The Europeana network of ancient Greek and Latin epigraphy data infrastructure. In: Closs, S., Studer, R., Garoufallou, E., Sicilia, M.-A. (eds.) MTSR 2014. CCIS, vol. 478, pp. 286–300. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13674-5_27

Medeiros e Sá, A., Ibañez Vila, A.B., Rodriguez Echavarria, K., Marroquim, R., Luiz Fonseca, V.: Accessible digitisation and visualisation of open cultural heritage assets. In: Selma, R., Karina, R.E. (eds.) Eurographics Workshop on Graphics and Cultural Heritage. The Eurographics Association (2019). https://doi.org/10.2312/gch.20191349

Miles, A., Zhao, J., Klyne, G., White-Cooper, H., Shotton, D.: OpenFlyData: an exemplar data web integrating gene expression data on the fruit fly Drosophila melanogaster. J. Biomed. Inform. **43**(5), 752–761 (2010). https://doi.org/10.1016/j.jbi.2010.04.004

Potenziani, M., Callieri, M., Dellepiane, M., Scopigno, R.: Publishing and Consuming 3D Content on the Web: A Survey, vol. 10, no. 4. Foundations, Trends R in Computer Graphics, and Vision (2018). https://doi.org/10.1016/10.1561/0600000083
Santarem Segundo, J.E.: Web Semântica, Dados Ligados e Dados Abertos: Uma Visão dos Desafios do Brasil Frente às Iniciativas Internacionais. XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB) (2015). http://www.brapci.inf.br/index.php/article/download/43838. Accessed 17 Oct 2020
Scholz, H.: Europeana Publishing Guide v1.8. Europeana Foundation (2019). https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Publishing%20Guide%20v1.8.pdf. Accessed 23 Oct 2020
Sczholz, H., et al.: Europeana Content Strategy. Europeana Foundation (2017). https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Content%20Strategy.pdf. Accessed 23 Sept 2020
Scopigno, R., Callieri, M., Dellepiane, M., Ponchio, F., Potenziani, M.: Delivering and using 3D models on the web: are we ready? Virtual Archaeol. Rev. **8**(17), 1–9 (2017). https://doi.org/10.4995/var.2017.6405
Siqueira, J., & Martins, D. L.: Workflow models for aggregating cultural heritage data on the web: A systematic literature review. J. A. Inform. Sci. Tech. 1–21 (2021). https://doi.org/10.1002/asi.24498
Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3** (2016). Article number: 160018. https://doi.org/10.1038/sdata.2016.18
Zeng, M., Qin, J.: Metadata, 2nd edn. ALA Neal-Schuman, Atlanta (2016)
Ziébelin, D., Hobus, K., Genoud, P., Bouveret, S.: Heterogeneous data integration using web of data technologies. In: Brosset, D., Claramunt, C., Li, X., Wang, T. (eds.) W2GIS 2017. LNCS, vol. 10181, pp. 35–47. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55998-8_3