UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Metadata Quality Evaluation in Cultural Heritage Domain

Matteo Lorenzini

Advisor:

Dr. Sara Tonelli

Fondazione Bruno Kessler

Co-Advisor:

Prof. Marco Rospocher

Università degli Studi di Verona

February 2022

To my wife Alina

Abstract

Metadata are fundamental for the indexing, browsing, and retrieval of cultural heritage resources in digital repositories. Since the manual control of metadata quality in digital repositories may not be feasible, especially when working with large collections, this Ph.D. thesis focuses specifically on the problem of automatic metadata quality assessment. Taking as the main reference the Metadata Quality Framework developed by Thomas Bruce and Diane Hilmann, we propose to evaluate metadata information according to three aspects. The first is metadata Completeness, approached as a statistical analysis. We compute the ratio of the filled elements with respect to the metadata schema taking into account its structure as well as the specific topic of a collection. The second is metadata Accuracy of the textual description of a given cultural heritage object, approached as a binary classification problem. We determine whether the field contains a high-quality or low-quality description, measured as the compliance of the textual content with the description rules from the guidelines used to implement metadata information. The last aspect concerns metadata Coherence, where we investigate the feasibility to use high-quality metadata at source while implementing metadata information. We assess the metadata Coherence of the subject element recommending the three most likely subjects of the resource analyzing the iconography of the resource. Applying this methodology to the Italian digital library "Cultura Italia", we noticed overall that it is indeed possible to automatically evaluate metadata quality. However, despite the promising results we obtained, to have a more detailed picture about automatic metadata quality evaluation, our methods should be also tested on a wider range of digital repositories.

Keywords Digital Library, Metadata Quality, Machine Learning, Cultural Heritage.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cultural heritage often brings to mind artifacts (paintings, drawings, prints, mosaics, sculptures), historical monuments and buildings, as well as archaeological sites. But the concept of cultural heritage is even wider than that, and has gradually grown to include all evidence of human creativity and expression: photographs, documents, books and manuscripts, instruments, etc. either as individual objects or as collections.

Thanks to the digitization, this huge variety of cultural resources are accessible on a large scale using dedicated infrastructure as a digital library or digital archive meant to aggregate cultural content from different collections and cultural institutions as museums and foundations. The duty of metadata is to describe and categorize the digital cultural resources available throughout a given digital archive [29, 11]. Metadata are defined as "data that provides information about other data". In other words, it is, "data about data"[1]. The main purpose of metadata is to increase the information related to a resource or to index the existing data according to a standard metadata schema such as Dublin Core[2]. Metadata comes in many shapes and flavors, carrying additional information about where a resource was produced, by whom, when was the last time it was accessed,

[1] https://en.wikipedia.org/wiki/Metadata

[2] `https://dublincore.org/`

what it is about, and many more details around it. Tables A.1,A.2 in the Appendix A shows the metadata standard schema used to describe the digital resources in the cultural heritage domain.[3]

Similar to museum cards describing a painting, metadata describes objects and adds more granularity to the way they are represented.

Managing and maintaining correct information in metadata throughout their entire life-cycle plays a fundamental role [56].

The International Federation of Library Association and Institutions[4](IFLA)identified four activities where metadata are involved:

- Find relevant elements, for example search for all the paintings made by Sandro Botticelli;
- Identify an element or discriminate between elements, for example to distinguish between two resources with the same title, e.g. "L'annunciazione" by Leonardo da Vinci and "L'annunciazione" by Beato Angelico;
- Select the most appropriate elements, e.g. to select a version of a book that is available at the library;
- Retrieve or obtain access to the selected element, e.g. to provide the URL of the desired online resource.

Additional to these four activities, modern metadata-based information systems also have new uses for the metadata:

- To cluster elements, for example recommend similar resources based on the subject of the cultural artifact;
- To improve the efficiency of the system, for example detect image duplication through the title.

[3]Source Wikipedia `https://en.wikipedia.org/wiki/Metadata_standard#Available_metadata_standards`

[4]`https://www.ifla.org/`

Metadata quality is directly related to how well metadata facilitates the six activities previously mentioned. Understanding how different metadata characteristics affect how to use, find, identify, select and retrieve the described elements and how they help or hinder the provision of extended or improved functionality is the main goal of research about metadata quality.

Resource	Metadata
	• Title: La Primavera • Author: Sandro Botticelli • Creation Date: 1482 • Subject: Allegory of Spring • Description: Large panel painting in tempera paint by the Italian Renaissance painter Sandro Botticelli ...

Table 1.1: Example of descriptive metadata of the painting "La Primavera" by Sandro Botticelli

Taking as an example the painting "La Primavera" by Sandro Botticelli in Table 1.1, if we have high-quality metadata, users are able to retrieve all the information about "La Primavera", or perform a specific search for Subject, e.g. retrieving all the resources referring to an Allegory. On the contrary, by having low-quality metadata, the information will be partial or, in the worst case, not correct; e.g. having Caravaggio instead of Sandro Botticelli as the author. In this case, the information provided to the end-users will be not only wrong, but will also represent a limit to a more specific faceted search for "Author", since the resource related to the Author Botticelli will be incomplete.

As an additional problem, the process of quality control still lacks of

a clear definition and workflow. This has several implications, including the impossibility of introducing systematic approaches to its automatic measurement and enhancement [23]. Day [16] assess metadata quality in e-print archives according to the functional requirements defined at two separate levels: compliance with the specifications of metadata schema used to describe the digital objects and compliance with the needs of end-users. At the first level, an object must be described strictly following the rules and guidelines of a metadata schema (or application profile) in order to be considered correct. The second, higher level of correctness, requires the rightness of the values of metadata fields.

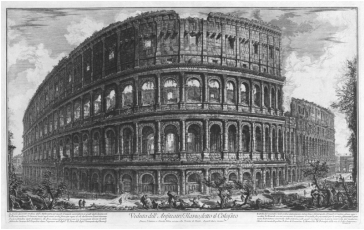
Resource	Metadata
	• Title: Veduta dell'Anfiteatro Flavio detto il Colosseo • Author: Giovan Battista Piranesi • Creation Date: 1757 • Subject: Roman Building • Description: The Colosseum is the most famous remnant of ancient Rome. Also called the Flavian Amphitheater ...

Table 1.2: Example of descriptive metadata of the drawing "Veduta dell'Anfiteatro Flavio detto il Colosseo" by Giovan Battista Piranesi

Consider for example the drawing "Veduta dell'Anfiteatro Flavio detto il Colosseo" by Giovan Battista Piranesi (Table 1.2). According to the guidelines of the Dublin Core Metadata Schema,[5] the title [6] should refer to a name given to the resource, in this case, "Veduta dell'Anfiteatro Flavio

[5] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#title`
[6] dc:title

detto il Colosseo", which is the original title of the drawing. However, the "Anfiteatro Flavio" is also known as "Il Colosseo". This could contribute to the creation of ambiguous metadata as in the case of the "Veduta dell'Anfiteatro Flavio detto il Colosseo" from the Minneapolis Institute of Art[7] in Figure 1.1, where the title of the resource is "The Colosseum".

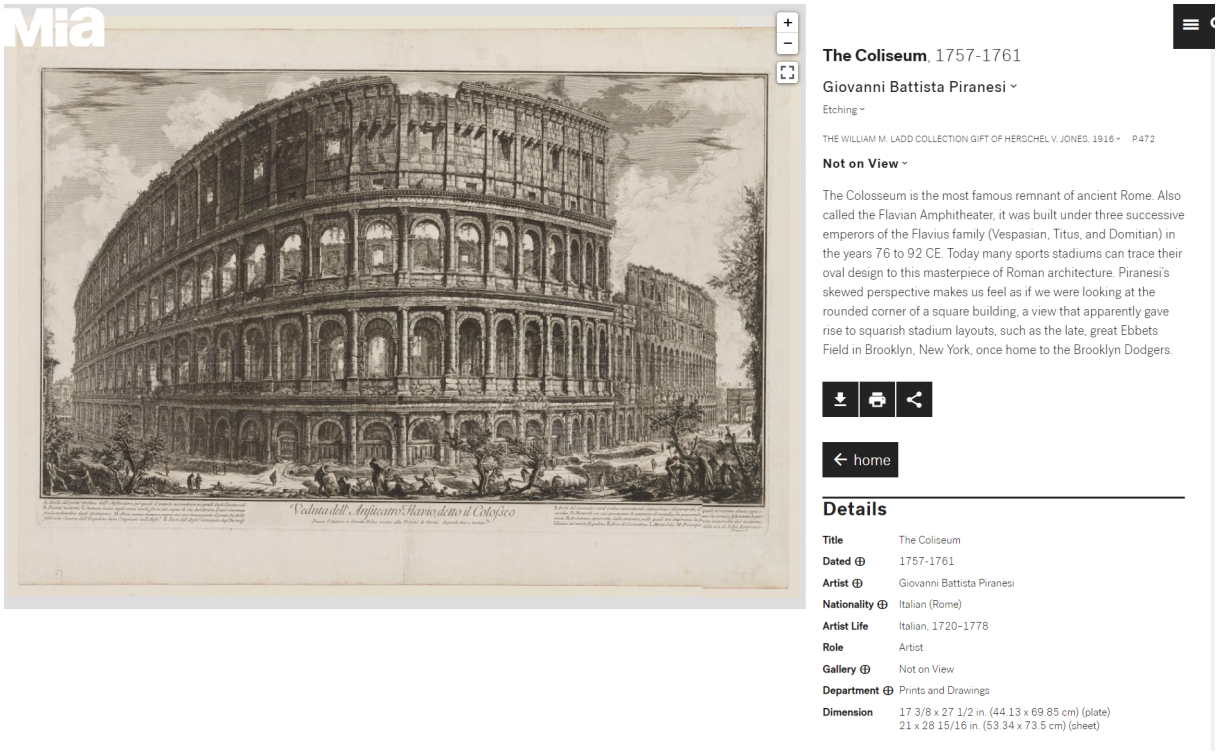


Figure 1.1: Screenshot of the record id.55193 from the Minneapolis Institue of Art

Evaluating the quality of the title on the basis of the compliance with metadata specifications, only "Veduta dell'Anfiteatro Flavio detto il Colosseo" should be considered as high-quality, since it is the original title used by the author to identify the drawing; instead, if we evaluate the quality of the title based on the compliance with users' needs, also "Il Colosseo/The Colosseum" should be considered as high-quality, since both appellations are valid for the identification of the resource. Hence, according to this second level of interpretation, quality and correctness are about fitness for

[7] https://collections.artsmia.org/art/55193/the-coliseum-giovanni-battista-piranesi

purpose.

The NISO Foundation[8] defines another approach to assess metadata quality and addresses the problem in the context of metadata creation by machines and by professionals who are not familiar with cataloging, indexing or vocabulary control [56]. The NISO Framework of Guidance for Building Good Digital Collections presents six principles of what is considered "good" metadata [52]. However, these criteria and principles do not provide a clear number of well-defined quality dimensions, so that metadata curators and end-users are not supported in addressing these issues.

The Metadata Quality Framework by Bruce and Hillmann [7] represents the first attempt to operationally define what the evaluation of metadata quality is, where seven dimensions and related characteristics are introduced and described as follows:

- Completeness: Completeness is a commonly referred quality dimension. It measures if the metadata elements are sufficient for a comprehensive and complete representation of the described resource. In this sense, the Completeness of metadata description is conditioned by characteristics of the resource type within a given domain and specifically by local metadata guidelines and best practices;

- Accuracy: It measures the correctness and precision of the provided information about a resource: how well metadata elements describe the object;

- Conformance to Expectations: The degree to which metadata fulfills the requirements of a given community of users for a given task could be considered as a major dimension of the quality of a metadata instance;

[8] https://www.niso.org/standards/

- Logical Consistency and Coherence: Degree to which the resource description matches with metadata standard schema and definition. For example, the resources should be described by the metadata elements from the standard metadata schema, categorical fields should contain only values from a taxonomy or thesaurus, the combination of categorical values and non-fixed values is not recommended;

- Accessibility: This refers to the degree of the logical accessibility or "findability" of a metadata instance. During the resource discovery process, an instance should be reached independently of the method used for its retrieval e.g. using filters, facets, or keywords;

- Timeliness: Metadata should change whenever the described object changes (currency);

- Provenance: The source of metadata can be another factor to determine its quality. Knowledge about who created the instance, the level of expertise of the indexer, what methodologies were followed at indexing time, and what transformations metadata has passed through, could provide in-sight into the quality of the instance.

However, there is no formal definition of the quality aspects that should be measured by each dimension. The authors note that it is not possible to state which of the seven dimensions they describe is most important for a given application, since the importance of each quality criterion is strictly influenced by the nature of the resource to be described, as well as by the environment in which metadata is to be constructed or derived. Thus, great emphasis is put on the fact that perception of quality strictly depends on context.

Besides the framework by Bruce and Hillmann, few other approaches have been proposed to automatically compute quality dimensions. The

ones that are more related to this work are the Framework for Information Quality Assessment by Stvilia [61], metadata Quality Framework by Ochoa and Duvall [53], the fine-grained dimension system for the Completeness of metadata by Margaritopoulos [46], and the Metdatata Quality Assurance Framework by Péter Király [32, 35].

Stvilia et al. [61] propose a framework that overlaps with the Metadata Quality Framework by Bruce and Hillmann. The author identifies four major sources of information quality problems: mapping, changes to the information entity, changes to the underlying entity or condition, and context changes. To address mapping, Stvilia adopts the definition from Wand [65] according to which this issue arises when there is incomplete or ambiguous alignment between the information source and the information entity from the metadata schema. Changes, instead, may occur in the information entity itself or in the real-world entity that it represents. Based on that, the authors develop a taxonomy of 22 dimensions, systematically organized into three categories: Intrinsic, e.g., dimensions that can be assessed by measuring information aspects in relation to reference standards (e.g., spelling mistakes); Relational, e.g. dimensions that measure relationships between the information and some aspects of its usage (e.g., Accuracy); and Reputational, e.g. dimensions that measure the position of an information entity in a given structure (e.g., authority). However, there is no implementation of these dimensions as the algorithms that can be operationally applied to different cases.

Ochoa and Duvall's framework [53] is inspired by the parameters introduced by Bruce and Hillmann and Stvilia. However, it is more detailed and specific, since it presents several automatic calculable dimensions of quality associated with the seven parameters in Bruce and Hillmann's framework. The authors point out that the proposed dimensions are not intended to be a comprehensive or definite set, but should be considered as a first step

towards the automatic evaluation of metadata quality.

The framework proposed by Margaritopoulos et al. [46] focuses on the evaluation of metadata Completeness following two dimensions of analysis both at the field level. The first dimension classifies a field as single or multi-value (e.g. *dc:language*). A multi-value field is considered complete if all the values indicated by metadata profile specification are filled. The second dimension of analysis goes deeper into the hierarchical structure of metadata schema, taking into consideration also the sub-elements of a given root element (e.g., the "file" section from METS metadata schema[9], which is composed of eight additional attributes). In this second dimension, a field is considered complete if all the sub-elements are filled. In both cases, Completeness is computed as the weighted average of the filled elements with respect to the metadata schema.

The last metadata quality framework we consider in this Ph.D. thesis is developed in collaboration with the Data Quality Committee (DQC) from the European Digital Library, "Europeana"[10] by Péter Király [35]. Metadata Quality Assurance Framework is an ongoing project tailored to measure metadata quality of the Europeana digital library and is based on the Europeana Data Model (EDM) metadata profile. The framework consists of four different dimensions, namely Completeness, Multilinguality, Uniqueness, (e.g. frequency of the duplicated values) and Record Patterns (e.g. density distribution of filled fields among all Europeana content providers).

The work presented in this Ph.D. thesis adopts an operational definition of metadata quality, considering it as a way to measure how much of the information describing a cultural heritage object supports a given purpose [57] focusing on Completeness, Accuracy, and Coherence dimensions.

[9] `https://www.loc.gov/standards/mets/`

[10] `https://www.europeana.eu/portal/en`

The contents of this thesis are based on the concepts and ideas presented in the following publications:

- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Automatically evaluating the quality of textual descriptions in cultural heritage records. *International Journal on Digital Libraries*, 22(2):217–231, 2021

- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Proposta per una valutazione automatica della completeness dei metadati nel contesto delle biblioteche digitali. *DigItalia*, 2:159–167, 2020

- Matteo Lorenzini. Automatic metadata curation of the cultural heritage resources. *Proceedings of AIxIA Doctoral Consortium*, (2249):33–37, 2018

- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Computer Assisted Curation of Digital Cultural Heritage Repositories. In *Proceedings of DH2019*. DataverseNL, 2019

- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. On assessing metadata completeness in digital cultural heritage repositories. *Digital Scholarship in the Humanities*, 36:182–188, 11 2021

1.1 The Problem

Despite the key role played by metadata in cultural heritage collections, how to systematically identify the data features that need to be improved and fixed is still a debated argument. There are two main reasons for this. The first is the context-specific notion of metadata quality that changes depending on the domain of the digital archive [63, 28]. The second is

the lack of a standard definition of the quality dimensions. For example, the Accuracy dimension has a different interpretation among various frameworks:

- Metadata quality frameworks from Bruce and Hillmann [7]: *The information provided about the resource in metadata instance should be as correct as possible [...] Typographical errors, as well as factual errors, affect this quality dimension*;
- Ochoa and Duvall [53]: *The Accuracy corresponds to the degree to which metadata values are "correct", e.g. how well do they describe the object*;
- Stvilia et al. [61]: *The Accuracy represents a dimension that measures the relationship between the information and some aspects of its usage.*

With different interpretations comes a different approach to evaluate the quality of metadata. As a consequence, metadata curators are required to follow a generic and "fitness for use" workflow [5] based on personal interpretation and manual intervention: they should check the content of each record and, depending on the metadata type and the correction required, report errors to metadata creators or fix metadata themselves, relying for instance on a controlled vocabulary. Given the growing amount of digital cultural heritage records available, this is a very time-consuming process, which cannot be adopted at scale.

Although the definition of metadata quality is still unclear, it is possible to identify the most frequent issues related to metadata quality evaluation:

- Issues related to the human factor: Metadata are often implemented by hand by the human operators following the guidelines provided by the digital archive and the standard definition of metadata schema. This condition, even if it can be considered a standard procedure,

could contribute to the production of low-quality metadata in two different aspects:

  - Each metadata element from the used metadata schema could be interpreted in a different way depending on the operator's point of view;
  - Depending on the domain, the operator could provide more detailed information for some metadata elements. Consider for example the textual description of a Greek epigraph: for an archaeologist, it represents an archaeological items related to a specific event; for a philologist, it represents a document. So, describing the epigraph the archaeologist will emphasize the physical aspects of the item, such as the shape, the material, and the conservation status. The philologist will emphasize the linguistic aspects of the item, such as the subject of the epigraph, the narrative style, and the syntax.

- Issues related to the context: often low metadata quality depends on the fact that metadata curators and creators are not able to retrieve the information about a specific resource or, in a specific context, such that metadata elements are simply not useful to cover a domain of interest and hence, they are not used;

- Issues related to the aggregation process: The aggregation process occurs between the content provider, e.g. a local digital archive as the Uffizi Museum, and the aggregator, e.g. a centralized digital archive as the Italian digital library Cultura Italia[11]. This practice could lead to the following issues:

  - Due to the dissimilar evaluation methodology, high-quality re-

[11] http://www.culturaitalia.it/

sources from the content provider can be wrongly categorized once aggregated. For example, during the assessment of the metadata Completeness, some resources could be evaluated as low-quality because of the different definition of the compulsory element by the aggregator;

- Multilingualism: it applies when the same resources are aggregated from different countries. An example could be represented by a dataset of postcards from Venice: the place Venice could be defined as Venezia, Venice, or Venedig according to the provenance of the data provider;
- Duplication of the resources: an example is represented by the same cultural object harvested from two different providers as for the painting "Sacra Famiglia (Tondo Doni)" by Michelangelo Buonarroti in Cultura Italia which has been harvested from the provider "Scala Group" (id.0093475), and from the provider "Galleria degli Uffizi" (id.work_63917).

Considering the limitations of the actual frameworks used to assess metadata quality and issues that have an impact on the processes of metadata quality check, the goal of this work is to define a hybrid methodology capable of evaluating metadata quality considering the structure, the formal aspect, and the content of the provided information.

1.2 Research Objectives

As described in the Section 1.1, issues such as the human factor, the context and the aggregation processes have an impact on metadata quality evaluation criteria.

The goal of this PhD work is therefore to define a hybrid methodology

that aims at automatically checking the metadata quality of a repository addressing the following main **Research Question:**

- Can computational methods support the evaluation and improvement of metadata quality in cultural heritage archives?

With the Completeness dimension this thesis aims to analyze the structure of metadata information assessing the issues related to the context of the resources addressing the following **Research Question 1 (RQ1):**

- Can we define a way to compute Completeness that takes into account the context and relevance of metadata elements for a metadata collection?

With the Accuracy dimension this thesis aims to analyze the formal aspect of the provided information assessing the issues related to the human factor while describing the cultural object. The problem is investigated specifically through the metadata element *dc:description* addressing the following **Research Question 2 (RQ2):**

- Can we effectively assess metadata Accuracy of textual descriptions using computational methods based on machine learning?

With the Coherence dimension the goal of this thesis is to analyze the content of the provided information examining the feasibility to prevent the creation of low-quality metadata. Taking as reference the textual description of the cultural resource we aim to recommend the correct information to be used to fill the metadata element *dc:subject*. This task is investigated by the following **Research Question 3 (RQ3):**

- Is it possible to predict the most coherent subject of a given cultural heritage resource from its description?

1.3 Methodology

This section provides a discussion of the chosen research design in order to answer the research questions outlined in the Section 1.2.

The main goal of this work is to propose and verify the feasibility of defining a hybrid methodology to automatically evaluate metadata quality of digital cultural heritage.

We start by identifying the main aspects of a metadata to be analyzed with respect to the structure and type of information provided. Afterwards, we evaluate to which dimensions from past frameworks these aspects could map for the computational evaluation. We remark that our goal is to create a methodology to evaluate metadata quality and not to develop another metadata quality framework.

For the purpose to have a real-world usable case against which to test and validate the results, the resources from the digital repository Cultura Italia[12] are used. CulturaItalia is the Portal of Italian Culture, managed by the Central Institute for the Union Catalogue of Italian Libraries (ICCU) and the Italian Ministry of Cultural Heritage (MiC); it involves cultural institutions from all sectors and levels (national, regional, and local), and gives access to a metadata repository, which gathers and organizes the information harvested from Cultura Italia's providers and consists of around 4,500,000 records including images, audio visual content and textual resources. All the resources are accessible via the OAI-PMH handler or via the SPARQL endpoint[19]. Metadata are ingested into Cultura Italia using the PICO metadata schema [9, 8], a qualified Dublin Core (DC) which consists of 94 elements. Given the nature of the records aggregated in Cultura Italia, most of the terms used are related to the Visual Art, Archaeology and Architecture, Sound, and Video domains. CulturaItalia, as

[12] `http://www.culturaitalia.it/`

a national aggregator, plays an important role in the development of Europeana, making available cooperative networks and agreements and coordinating technical activities leading to the establishment of Europeana [18].

The choice of this specific use case is due to two reasons: the first one concerns the quantity and the variety of metadata managed by the repository, which allows testing the proposed methodology and a wider range of domain and context. The second reason is related to the fact that in the past, I personally worked at Cultura Italia as a Data Ingestion Manager, so I am able to benefit from a direct contact with the technical unit of the portal. This condition is fundamental to assess the specific tasks explained in the Chapters 3, 4, 5 and to validate the results obtained by the proposed methodology.

1.3.1 Definition of Qualitative Desiderata

Aiming to define a hybrid and scalable evaluation methodology, the first phase of the project focuses on identifying the general aspects according to which a metadata should be evaluated.

While analyzing the metadata curation process usually done by curators from Cultura Italia, we identify three different perspectives that should be always considered into the evaluation process:

- Evaluation of metadata structure: to check if metadata structure used to describe the resources complies with the standard guidelines provided by metadata schema and with the guidelines provided by the digital archive on how to implement metadata schema in order to successfully harvest metadata into the repository;
- Evaluation of formal aspects: to check if the information provided by metadata complies with the guidelines on how to fill metadata elements;

- Evaluation of the provided information: to check if the information provided by metadata is correct.

After the analysis of previous approaches described in Chapter 2 to evaluate metadata quality and the real impact that specific dimensions have in metadata evaluation, we use (i) the Completeness dimension, to evaluate the metadata structure used to describe the resources, (ii) the Accuracy dimension to evaluate the formal aspect of metadata information and finally (iii) the Coherence dimension to evaluate and enhance the content of metadata information. Moreover, dimensions like Conformance to the Expectations or Provenance are too abstract for a quantitative definition. For example, in the case of the Provenance, it is not possible to assume that high-quality metadata can only be produced by renowned Institutes. The resource id. 1890:5312 from Cultura Italia[13], referring to the Italian masterpiece "Bacco" made by Caravaggio is an example. As is possible to see in Figure 1.2 even though the digital resource belongs to the "Polo Museale Fiorentino", the provided information is for sure not sufficient to categorize this resource as high-quality: the image preview, the subject and the description of the painting are missing and the rest of the available information are few and incomplete.

1.3.2 Quantitative Definition of the Quality Dimension

Completeness Dimension

The evaluation of the Completeness dimension often refers to a specific metadata profile [33] or either focuses on the analysis of a specific metadata element [46]. These approaches, even if they allow metadata curators to check the quality status of a single record or, more generally, of a dataset, do not try to embed in the computation elements assessing whether low

[13] `https://bit.ly/3669SRk`

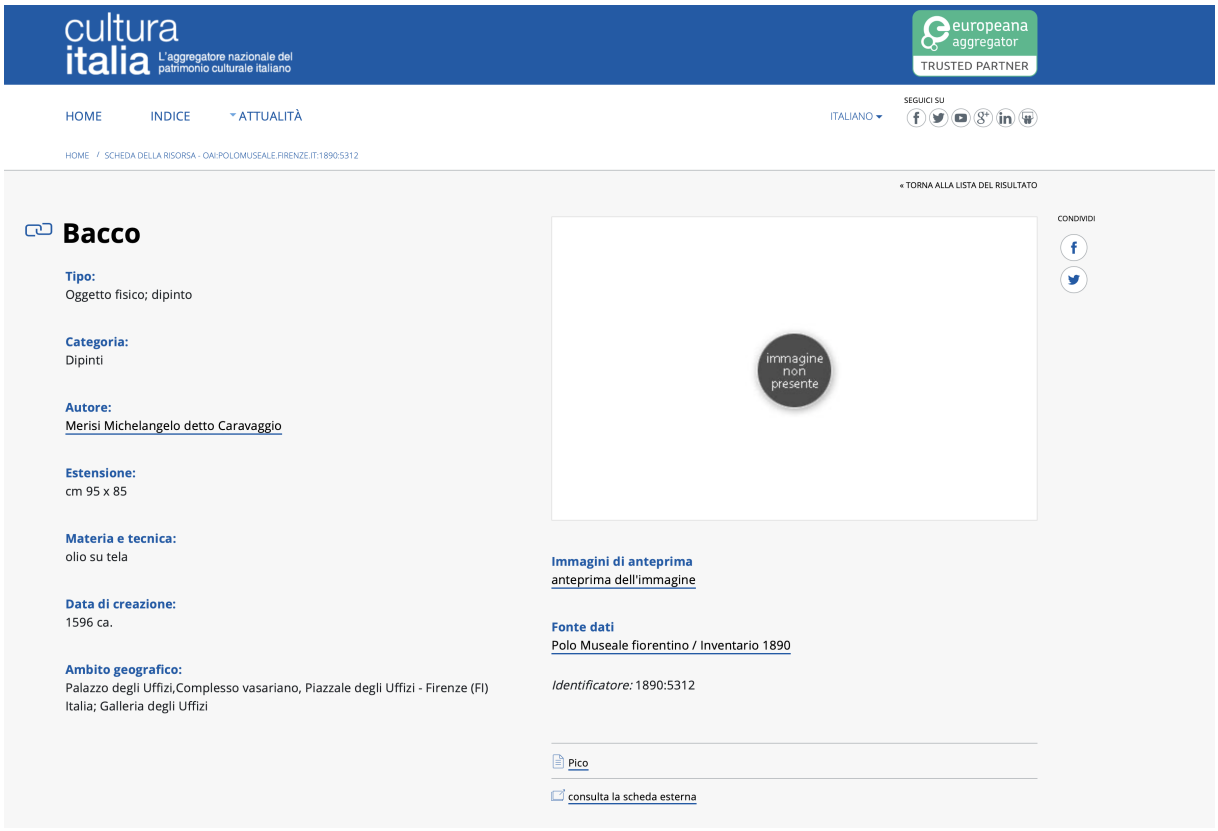


Figure 1.2: Screenshot of the resource id. 1890:5312 from Cultura Italia "Bacco" by Caravaggio

metadata quality is due to the lack of few metadata with high relevance or to the lack of many metadata elements with low relevance. However, the Completeness dimension should enable users and curators to define in a flexible way what metadata they deem more relevant in the overall evaluation of Completeness, to make this value comparable across different repositories, and to allow for a fine-grained analysis of metadata elements. In this work, we compute the Completeness dimension as the ratio of filled elements with respect to a metadata profile. In this computation, several variables are taken into account: for instance, metadata elements that are mandatory and those that are optional, the context and the domain of a collection, as well as the preferences of curators when evaluating Completeness.

The approach to evaluate metadata Completeness consists of the following key steps:

- Given a repository to be evaluated, we divide metadata elements into groups, representing their importance (e.g., compulsory/recommended/optional metadata);
- For each object o in the repository, we compute a separate Completeness score $C_g(o)$ for each metadata group G as follows: the number of filled G metadata for that object is divided over the total number of G metadata elements;
- These Completeness scores are computed (one for each metadata group) for each element of the dataset.

Accuracy Dimension

We measure the metadata Accuracy as the degree to which the data values in the metadata record match with the characteristics of the described object [61]. The focus of the proposed approach is to determine the Accuracy of the textual description (typically encoded using the *dc:description* element from the Dublin Core[14] metadata schema) of a given cultural heritage object. More specifically, we propose to assess the Accuracy of such description metadata by determining whether the field contains a high-quality or low-quality description of the considered object, measured as the compliance of the textual content with the description rules from Istituto Centrale per il Catalogo e la Documentazione[15] (ICCD), adopted as cataloging standard on a national level in cultural heritage domain. As a first step, by taking the resources from Cultura Italia, we develop a dataset for training and testing machine learning approaches: the dataset consists

[14] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[15] http://www.iccd.beniculturali.it/

of object descriptions manually labeled as high/low quality according to the adherence to the cataloguing guidelines of the digital repository indexing the objects. The resulting annotated dataset is publicly available [40] under the terms of the Creative Commons Attribution-ShareAlike 4.0 Generic (CC BY-SA 4.0) license.

Secondly, to address RQ2 we execute several experiments, comparing two different classification algorithms: Support Vector Machine (SVM) [14] and the FastText logistic regression classifier [30]. We perform the comparison on the three cultural heritage domains, "Visual Artworks", "Archaeology" and "Architecture", assessing system performances using well-known metrics (e.g., precision, recall, F1-measure) and adopting evaluation protocols aiming to reduce possible biases (e.g., cross-validation setting and the removal of duplicates).

As the third step, we analyze the learning curve of the best classification model, by incrementally adding new instances to the training data to check how many annotated resources are needed to create a good quality dataset to assess description quality.

All the code used for running the classifiers and preprocessing the dataset is available on the GitHub code repository of the Accuracy dimension.[16]

Coherence Dimension

Unlike the standard usage of the Coherence dimension, which points to the evaluation of the degree to which the fields describe the same object similarly [7], in this work, Coherence prevent the creation of erroneous information during the metadata creation process, exploiting the possibility to automatically create high-quality metadata at the source.

More precisely, we focus on enhancing the quality of the information related to the subject (typically encoded using the *dc:subject* element from

[16] `https://github.com/matteoLorenzini/description_quality`

the Dublin Core[17] metadata schema) of a given cultural heritage resource. Referring to the definition of the Coherence dimension provided by Bruce and Hillmann[18] and the guidelines[19] from Istituto Centrale per il Catalogo e la Documentazione (ICCD), we investigate the feasibility to predict the correct subject of a given digital resource from the iconography illustrated in the textual description[20].

As the first step in this direction, we train a specific model based on the Iconclass[21] vocabulary. Then we integrate the trained model into a multi-label classification system able to suggest the three most likely subjects of a certain resource using the categories of the Iconclass vocabulary.

All the code implemented is available on the GitHub code repository of the Coherence dimension.[22]

1.3.3 Results Evaluation

The last aspect concerns the evaluation and validation of the results we obtain from the methodology we propose. We validate our methodology by taking as a reference the resources from the Cultura Italia portal. We select two of the most representative datasets in terms of number of resources: "Regione Marche" and "MuseID-Italia", in total 149,768 resources. We apply on this dataset the algorithm we implemented for the computation of the Metadata Completeness, Accuracy, and Coherence. The results are validated by metadata curators from Cultura Italia.

[17] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`

[18] 'Metadata should be consistent with standard definitions and concepts used in the domain. The information contained in metadata should also have internal coherence, which means that all the fields describe the same resource'

[19] The subject must refers to the scene or the subject represented on the cultural resource

[20] dc:description

[21] `https://rkd.nl/nl/collecties/services-tools/iconclass`

[22] `https://github.com/matteoLorenzini/coherence_quality`

1.4 Structure of the Thesis

This thesis is structured as follows. In Chapter 2, we provide a detailed overview of the problems related to metadata quality evaluation and of the past attempts used to evaluate metadata analyzing metadata quality framework defined by Moen [49] in Subsection 2.2.1, Bruce and Hillmann [7] in Subsection 2.2.2, Margaritopoulos et al. [45] in Subsection 2.2.3, Ochoa and Duvall [53] in Subsection 2.2.4, Stvilia et al. [61] in Subsection 2.2.5 and Király [32] in Subsection 2.2.6.

In Chapter 3, we assess the problem of the evaluation of the Completeness dimension. After discussing how Completeness is computed in the other frameworks, we present in Section 3.3 our methodology to compute the Completeness dimension. In Section 3.4 we then present the results of the computation of Completeness in Cultura Italia.

In Chapter 4, we present the computation of metadata Accuracy. After introducing the state of the art in evaluating the Accuracy dimension, we present our solution based on a machine learning approach describing in details how we framed the problem in Section 4.3, the construction of the dataset in Subsection 4.3.1 and, in Section 4.5 the evaluation of the obtained results by comparing two different algorithms 4.3.1.

In Chapter 5, we present our solution to suggest information coherent with the cultural resources. After critically analyzing the solution adopted so far we describe the methodology used to implement our multi-label classification system capable of predicting the correct metadata information that should be used by the metadata creator during the creation of digital resources.

Finally, in Chapter 6, we present our conclusions and discuss the limitations of the proposed methodology.

Chapter 2

Metadata Quality

As already underlined in Section 1 of this work, descriptive metadata represents the backbone through which users can navigate information and improve their knowledge of specific topics, also reusing data coming from external sources [29, 11]. For this reason, managing and maintaining correct information in metadata throughout their entire lifecycle plays a fundamental role [56]. The necessity of a lifecycle approach, to ensure the continuity of digital material, is discussed by Pennock [8]. A lifecycle approach ensures that all the required stages are identified and planned, and necessary actions implemented, in the correct sequence. This can ensure the maintenance of authenticity, reliability, integrity and usability of digital material, which in turn ensures maximization of the investment in their creation.

The DCC Curation Lifecycle Model [27] represented in Figure 2.1 offers a graphical high-level overview of the lifecycle stages required for successful curation. Generic in nature, the model is indicative rather than exhaustive. It can be used to define roles and responsibilities, build frameworks of standards and technologies and ensure that the processes and policies are adequately documented. The model identifies *i)* curation actions that are applicable across the whole digital lifecycle *ii)* those which need to be

undertaken sequentially if curation is to be successful, and *iii)* those which are undertaken occasionally, as circumstances dictate.

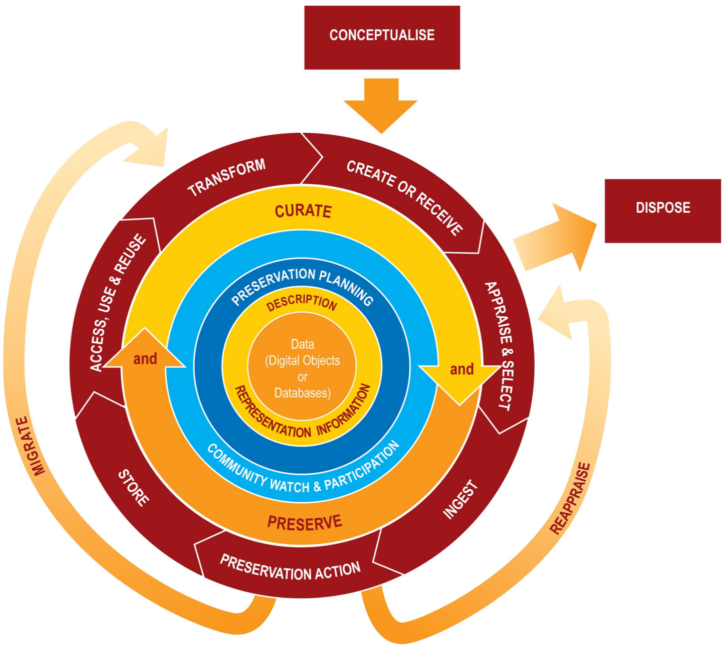


Figure 2.1: DCC Curation Lifecycle Model

Consider for example the resource id:1890:8346, "Presentazione di Gesú al Tempio" from Cultura Italia by Ambrogio Lorenzetti in Figure 2.2, which is one of the Italian visual artwork masterpieces. Despite metadata quality control made before the ingestion by metadata curators, this record still lacks in description, the preview image is missing and the creation date is wrong, since the painting was created in 1342.

This means that the result presented to end-users will be only partially correct. However, despite the wide agreement on the need to assure high-quality metadata, there is less consensus on what high-quality means and how metadata quality should be measured. Traditionally, quality has been defined as the degree of excellence of an object [24]. However, metadata quality does not only depend on some objective characteristics of the resource, but also on the needs and uses of a given domain and community

Figure 2.2: Example of low-quality metadata record from Cultura Italia: Presentazione di Gesù al Tempio

of practice. What in some domains is considered as high metadata quality, in others could be considered as low quality. An example is represented by the description of the archaeological object id. 0600044578 from Cultura Italia repository: *Plate fragment in Terra Sigillata Chiara A, shape Hayes 31*. For an archaeologist it represents a high-quality description, since the type of the archaeological item (Plate fragment), the typological classification to which it belongs (Terra Sigillata Chiara A) and finally, the shape (Hayes 31) have been specified. However, even if the description provided is fully compliant with the guidelines[1] by the Istituto Centrale per il Catal-

[1]RA card *OGT elements `http://bit.ly/ICCD_RA_card`

ogo e la Documentazione (ICCD), for those who are not archaeologists, the same description could be interpreted as low-quality. As a matter of fact the plate is described by the identifier from the collection made by John W. Hayes (Hayes 31) and not by a textual description. This could lead to a wrong interpretation of metadata quality. Due to this dependence on the context and domain, nowadays it is a common practice to define the evaluation of metadata quality as "Fitness for purpose" [5].

In this chapter we present the past approaches and methodologies used to assess metadata quality in the digital library domain. They can be grouped as follows:

- Guidelines: The usage of guidelines aims to prevent the creation of low-quality metadata by cataloguers. Guidelines have been defined with the purpose to create a set of best practices on how to implement metadata e.g. FAIR principles [2] and NISO[3] or, on a more granular level, on how to implement metadata according to a specific metadata schema or application profile, as for the technical guidelines provided by each digital archive to the content provider;

- Implementation of metadata quality frameworks: The goal of metadata quality frameworks is to define a set of qualitative metrics that are able to evaluate the quality of metadata once they have been ingested into the digital archive. The metadata Quality Framework from Bruce and Hillmann [7] is an example. However, few attempts have been proposed to automatically compute quality metrics [53, 32, 35, 23, 46] providing a quantitative definition of the same metrics;

- Usage of interoperable semantic models: this approach is used in semantic archives, that is to say when the information is structured

[2] https://www.go-fair.org/2017/12/11/metrics-evaluation-fairness/
[3] http://www.niso.org/publications/understanding-metadata-2017

according to an ontology or a conceptual reference model. Here, the goal is to guarantee the structural metadata quality using an interoperable semantic model and to standardize the information using controlled vocabularies to fill specific entities as places or actors [10].

2.1 Guidelines

Guidelines provide best practices useful for metadata creators and archivists on how to implement and maintain a metadata repository or dataset.

The FAIR principles represent one of the last formulation. They provide guidelines for the publication of digital resources such as datasets, code, workflows, and research objects, in a manner that makes them Findable, Accessible, Interoperable, and Reusable [66]. The complete definition of the FAIR principles is summarized in Table 2.1

The NISO[4] Foundation[5] defines another approach which focuses on the context of metadata creation by machines and by professionals who are not familiar with cataloging, indexing, or vocabulary control [56]. The NISO Framework of Guidance for Building Good Digital Collections presents nine principles of what is considered "good" metadata [52]. However, these criteria and principles do not provide a clear number of well-defined quality dimensions, so that metadata curators and end users are not supported in addressing these issues. The complete definition of the NISO principles is summarized in Table 2.2

[4] http://www.niso.org/publications/understanding-metadata-2017
[5] https://www.niso.org/standards/

Principles		
Findable	F1	(meta)data are assigned a globally unique and persistent identifier
	F2	data are described with rich metadata (defined by R1below)
	F3	metadata clearly and explicitly include the identifier of the data it describes
	F4	(meta)data are registered or indexed in a searchable resource
Accessible	A1	(meta)data are retrievable by their identifier using a standardized communications protocol
	A1.1	the protocol is open, free, and universally implementable
	A1.2	the protocol allows for an authentication and authorization procedure,where necessary
	A2	metadata are accessible, even when the data are no longer available
Interoperable	I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
	I2	(meta)data use vocabularies that follow FAIR principles
	I3	(meta)data include qualified references to other (meta)data
Reusable	R1	meta(data) are richly described with a plurality of accurate and relevant attributes
	R1.1	(meta)data are released with a clear and accessible data usage license
	R1.2	(meta)data are associated with detailed provenance
	R1.3	(meta)data meet domain-relevant community standards

Table 2.1: FAIR Principles

The purpose of these principles is to cover the widest variety of desiderata while defining a metadata set. On the other hand, the technical guidelines provided by the single digital archive are delineated with the purpose to facilitate the implementation of a metadata set according to a specific metadata schema or application profile. An example is represented by the Italian digital library Cultura Italia, which adopts two metadata schema to describe the digital resources: the full Dublin Core, which is a standard metadata schema, and the PICO[6] application profile, a qualified Dublin

[6] `http://purl.org/pico/1.1/picotype.xsd`

Principles	
Collections Principle 1	A good digital collection is created according to an explicit collection development policy.
Collections Principle 2	Collections should be described so that a user can discover characteristics of the collection, including scope, format, restrictions on access, ownership, and any information significant for determining the authenticity of the collection, the integrity, and the interpretation.
Collections Principle 3	A good collection is curated, which is to say, its resources are actively managed during their entire lifecycle.
Collections Principle 4	A good collection is broadly available and avoids unnecessary impediments to use. Collections should be accessible to persons with disabilities, and usable effectively in conjunction with adaptive technologies.
Collections Principle 5	A good collection respects intellectual property rights.
Collections Principle 6	A good collection has mechanisms to supply usage data and other data that allows standardized measures of usefulness to be recorded.
Collections Principle 7	A good collection is interoperable.
Collections Principle 8	A good collection integrates into the users own workflow.
Collections Principle 9	A good collection is sustainable over time.

Table 2.2: NISO Principles

Core specification. In this case, especially regarding PICO elements, the role of the guidelines is crucial to produce high-quality metadata.

2.2 Metadata Quality Frameworks

The goal of metadata quality frameworks is to provide metrics that are able to measure the quality of information at the collection and record level. This type of measurement presents three main issues. First, metadata quality is inherently multi-dimensional. Indeed, there are several independent dimensions of metadata records that affect quality. For example, dimensions such as metadata Completeness or Accuracy. Second,

metadata quality is user and task-dependent. The assessment or measurement made for one community of practice and system is maybe not valid for another community or even the same community using a different system. Finally, quality is not static. The aging of the record, the addition of new metadata records in the collection or the change in the usage patterns could affect how well the records enable the different functions of the systems. To deal with multidimensionality and to reduce subjectivity in the assessment of information quality, several quality evaluation frameworks have been developed. These frameworks define parameters that indicate whether information should be considered as high or low-quality.

In the following sections we describe metadata quality frameworks worthy of attention. These frameworks will be used as a comparison for the work we present in this thesis except the Virtual Language Observatory Curation Module (VLO) [54] described in the Subsection 2.2.7. Despite the fact that the methodology and the approach used by the VLO framework have returned excellent results [62], it only focuses on the analysis of the linguistic content, while the goal of the work presented in this Ph.D. thesis is to evaluate metadata information with respect to a digital resource.

2.2.1 Moen et al.'s Metadata Quality Framework

Moen, Stewart and McClure [49] framework' s represents the first attempt to assess metadata quality. They proposed 23 assessment criteria: Access, Accuracy, Availability, Compactness, Compatibility,Comprehensiveness, Content, Consistency, Cost, Data Structure, Ease of Creation, Ease of Use, Economy, Flexibility, Fitness For Use, Informativeness, Protocols, Quantity, Reliability, Standard, Timeliness, Transfer and Usability. While the purpose was not to describe in detail each one of these dimensions, this list represents the first approach to define the different dimensions of metadata quality.

2.2.2 Bruce and Hilmann's Metadata Quality Framework

We consider the curation framework by Bruce and Hillmann [7] as a benchmark in the pursuit of quality assessment[7]. This framework defines seven parameters to measure the quality of metadata:

- Completeness: Metadata should be complete in two senses. First, the element set used should describe the target objects as completely as economically feasible. It is almost always possible to imagine describing things in more detail, but it is not always possible to afford the preparation and maintenance of more detailed information. Second, the element set should be applied to the target object population as completely as possible; it does little good to prescribe a particular element set if most of the elements are never used, or if their use cannot be relied upon across the entire collection.

- Accuracy: Metadata should be accurate in the way it describes objects – a uniquely non-controversial statement that houses platoons of devils. Minimally, the information provided in the values needs to be correct and factual. At the next level, Accuracy is simply high-quality editing, since it should lead to the elimination of typographical errors, conforming to the expression of personal names and place names, use of standard abbreviations, and so on, in large or heterogeneous collections, Accuracy may not be directly verifiable; sampling techniques, statistical profiles, or other alternatives to laborious inspection may be needed.

- Logical Consistency and Coherence: Consistency and coherence are usually seen as problems only for heterogeneous, federated collections,

[7]The domain independence that the framework offered delegated the responsibility for a contextual wrap to posterior attempts.

or perhaps for single collections that are presented in successive "releases" over time. But in fact, very few collections exist in isolation, even at their inception. There is almost always a need to ensure that elements are conceived in a way that is consistent with standard definitions and concepts used in the subject or related domains and presented to the user in consistent ways

- Accessibility: Metadata that cannot be read or understood by users has no value. The obstacles may be physical or they may be intellectual. Barriers to physical access come in several forms. Metadata may not be readily associated with the target objects, perhaps because it is physically separated, comes from a different source, or is not properly keyed or linked to the object being described. Or it may be unreadable for a wide variety of technical reasons, including the use of obsolete, unusual, or proprietary file formats that can only be read with special equipment or software.

- Timeliness: We use two different terms to refer to two different aspects of metadata timeliness: "currency", and "lag". "Currency" problems occur when the target object changes but metadata does not. "Lag" problems occur when the target object is disseminated before some or all metadata are knowable or available.

- Provenance: The provenance of metadata often provides a useful basis for quality judgments. Sometimes this is a matter of knowing who prepared metadata, how experienced he or she might be, how good his or her judgment is, or of having some sense of their expertise in the relevant domain and with metadata standards generally. We may also rely on well-understood or certified methodologies as proxies that ensure reliability and quality. Scientists and statisticians are quite at home making judgments about the quality of data based

on the methods used to create and handle it. This is particularly true in situations where individual items cannot be directly verified. However, the use of creation and handling methodology as guarantor of quality is not limited to the sciences; all sorts of content standards and best-practices guides exist, the Anglo-American Cataloging Rules (AACR2) not least among them.

- Conformance to the expectations: Standard metadata element sets and application profiles that use them are promises from the metadata provider to the user. Moreover, they are promises surrounded by the expectations of a particular community about what such promises mean, how realistic they are, and how they are to be carried out.

2.2.3 Margaritopoulos et al.'s Metadata Quality Framework

Margaritopoulos et al. [46] assess the problem of metadata quality according to two levels. The first level concerns the correctness of the information considering the requirements that the values of metadata fields must obey the grammatical and syntactical rules of the language and metadata standard or the application profile used. Missing letters, misspelled words, inconsistent formatting or representation of the same fields, fields containing inappropriate values with respect to metadata standard schema, are among the problems of this level. According to Margaritopoulos at al., the first level of correctness concerns objective information and can be resolved, for example, by using any relevant validator parser. The second level concerns the relevance of the information that should be appropriate and complete. This condition is measured by Completeness, which is evaluated at the field level, following two dimensions of analysis. The first dimension classifies a field as single or multi-value (e.g., *dc:language*). A multi-valued field is considered complete if all the values indicated by

metadata profile specification are filled. The second dimension of analysis goes deeper into the hierarchical structure of metadata schema, taking into consideration also the sub-elements of a given root element (e.g., the "file" section from METS metadata schema, which is composed of 8 additional attributes). In this second dimension, a field is considered complete if all the sub-elements are filled. In both cases, Completeness is computed as the weighted average of the filled elements with respect to metadata schema [46, 45].

2.2.4 Ochoa and Duvall's Metadata Quality Framework

Ochoa and Duvall [53] framework' s reflects metadata Quality Frameworks from Bruce and Hillmann already described in Section 2.2.2. While Bruce and Hillmann proposed their framework to guide metadata curators in evaluating the quality of metadata elements, Ochoa and Duval's work aims at proposing a framework that comprises meaningful quality parameters, e.g., quality parameters that might be used by human reviewers, associated with automatic calculable measures of quality. In particular, they complement Bruce and Hillmann's framework by proposing automatic measurement methods for each of the seven parameters of such a framework. Ochoa and Duvall tested the framework on the ARIADNE repository. They perform experiments to evaluate their metrics and to find that they correlate well with human evaluation. After testing their metrics, Ochoa and Duval found several of them to correlate well with human evaluation. However, the authors concluded that their proposed metrics were not a solution but a baseline for further metrics of automatic evaluation

2.2.5 Stvilia at al.'s Metadata Quality Framework

In contrast to context-specific quality assessment models depending on variables determined by local needs, the evaluation framework by Stvilia et al. [61] focuses on studying the causes of quality change, and presents a framework consisting of typologies of Information Quality (IQ) problems, related activities, and a systematically organized taxonomy of information Quality dimensions. In this framework, an IQ problem occurs when the quality of the information provided by metadata does not meet the requirements from one or more IQ quality dimensions. Four major sources of IQ problems are identified: mapping, changes to the information entity, changes to the underlying entity or condition, context changes. From the analysis of these sources, the authors develop a taxonomy of 22 IQ dimensions, systematically organized into three categories:

- Intrinsic: Some dimensions of information quality can be assessed by measuring attributes of information items themselves, in relation to a reference standard. Examples include spelling mistakes (dictionary), conformance to formatting or representation standards (HTML validation), and information currency (age with respect to the standard index date, e.g., 'today"). In general, Intrinsic IQ attributes persist and depend little on context, hence can be measured more or less objectively.

- Relational/Contextual IQ: This category of IQ dimensions measures relationships between information and some aspects of its usage context-dependent and their attributes may change in time and space as the metadata provenance.

- Reputational IQ: This category of IQ dimensions measures the position of an information artifact in cultural or activity structure, often determined by its origin and its record of mediation.

2.2.6 Király's Metadata Quality Framework

The Metadata Quality Assurance Framework by Péter Király in 2015 [33] is developed in collaboration with the European Digital Library Europeana. It represents the last implementation in the panorama of metadata quality framework based on the metrics proposed by Bruce and Hillmann. Since the mission of Europeana is to aggregate metadata from Institutional, National and Regional archives across Europe, the goal of metadata Quality Assurance Framework is to check the quality of metadata on an higher level covering the most relevant issues derived from the aggregation process as the duplication of the resources, multilingualism and the usage of multiple metadata standards to describe the resources.

Later in 2019, Király analysed the problem on a deeper level in his Ph.D. thesis focusing metadata quality evaluation on:

- Completeness: the ratio of the filled fields with respect to metadata schema;
- Multilinguality: the average number of languages per property for which there is at least one language tagged e.g. "Olpe Chigi" @it;
- Uniqueness: frequency of the duplicated values, e.g. the painting of "La Gioconda" harvested from two different content providers;
- Record patterns: e.g. density distribution of filled fields among all Europeana content providers.

2.2.7 VLO Curation Module

The Virtual Language Observatory (VLO) Curation Module [54] aims to check the discoverability and accessibility of valuable linguistic contents fa-

cilitating metadata ingestion and curation process [63][8]: it provides a systematic method to measure metadata quality and a user-friendly interface (Figure 2.3) to inspect profiles, records, and collections of the Component MetaData Infrastructure (CMDI)[9] used for the VLO.

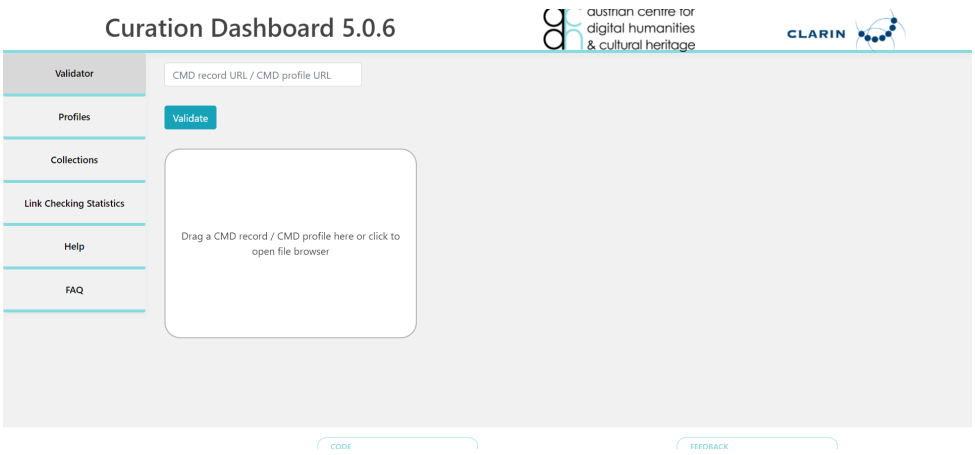


Figure 2.3: VLO Curation Module dashboard

The design of the module is guided by the following four use cases:

- Metadata editor checks (on the fly) the quality and validity of a newly created record;
- Metadata modeller evaluates the quality of profiles (especially facet coverage), when selecting an existing profile or creating a new profile for new resources;
- The data provider, repository administrator, or collection manager checks the overall quality of metadata in his/her repository, including the facet coverage;

[8]The Curation Module heavily depends on other CLARIN infrastructure services such as the Component Registry from where it fetches the XSD schema files of the CMD profiles and the Concept Registry from where it retrieves information about concepts.

[9]https://www.clarin.eu/content/component-metadata

- All records ingested into the VLO undergo a systematic process of curation, validation, normalization and quality assessment (benchmarking).

Then the improvement of the resources is demanded to metadata curation task force (MCTF).

2.3 Interoperable Semantic Models

The usage of interoperable semantic models in metadata curation and quality control aims to improve metadata quality at the source. The usual pipeline is to create semantic data from multiple structured data sources such as SQL databases and provide to the end-users the semantic data via specific semantic infrastructures and tools such as the Research Space[10] platform developed by The British Museum. Specifically, this process involves the following steps:

- Data Mapping: The goal of the mapping process is to interpret each object type from the source repository as a semantic model through the alignment of the object elements into the respective entity from the ontology used to represent the information. To create a standard and interoperable mapping schema, it is a common practice to use a specific mapping template. The Swiss Art research Infrastructure[11] (SARI) as described in the Section 2.3.1 is an example;
- Data Transformation: Data are transformed into RDF graphs by a specific mapping tool. When possible, RDF data are also reconciled to the terminology from a fixed-URI controlled vocabulary. This is to avoid duplication of the resources and to make the information standard and interoperable;

[10] `https://researchspace.org`
[11] `https://swissartresearch.net`

- Data Ingestion: The final step is the data ingestion of the resulting RDF graphs into the platform.

The usage of mapping templates and standard ontologies ensures that metadata are both interoperable and structurally homogeneous. However, apart from the process of reconciliation, which allows a semantic enrichment of the resource and the use of a vocabulary with standard terms, making reliable information possible, this methodology does not foresee metadata quality control. This process assumes that all information are validated and checked at the source.

In this respect, we mention two projects, the Swiss Art Research Infrastructure (SARI) and the Mapping Manuscript Migrations project [12], below are the details.

2.3.1 Swiss Art Research Infrastructure

The Swiss Art Research Infrastructure (SARI) provides access to domain-specific research data, collection data, digitized visual resources, and related reference data in the field of art history, design history, history of photography, film studies, architecture and urban planning, archaeology, history studies, religious studies, and other disciplines related to the visual studies. The core of the project is the implementation of the Semantic Reference Data Model[13] (SRDM) which allows mapping the domain specific data into 8 main entities: Persons, Artworks, Group, Built Work, Place, Digital Document, Events and Bibliographic Entity. The entities are semantically defined according to 8 different mapping templates that are implemented using the standard ontology CIDOC-CRM[14][20]. Each

[12] https://mappingmanuscriptmigrations.org/en/

[13] https://airtable.com/apposIclAqCiaDQK5/tbluwq931mwZr0nvp/viwOm3KN7oDfzObey?blocks=hide

[14] http://www.cidoc-crm.org

entity is defined by a fixed URI. The aim is manifold: to provide reference implementations to be used by institutions and projects not familiar with CIDOC-CRM and, in general, with semantic data models, to create a usable schema to generate input semantic data source and to guide mapping processes from extant sources into the CRM-conformant reference model. For example, Figure 2.4 represents the mapping schema of the entity ArtWorks defined using the CIDOC-CRM ontology.

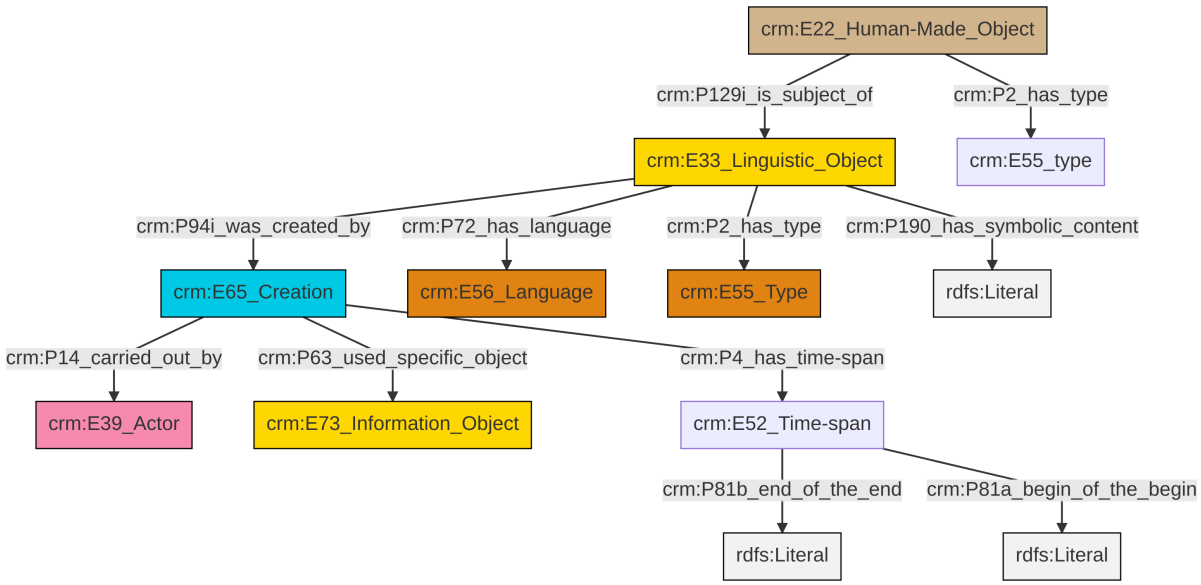


Figure 2.4: Artworks mapping schema in CIDOC-CRM

Once the source data are materialized into RDF graphs, the Place and Actor entity are reconciled respectively using the Getty Thesaurus of Geographic Names (TGN)[15], which provides a good basis for hierarchical reasoning when exploring the data geographically, and the Virtual International Authority File (VIAF)[16] wherever possible.

In SARI the metadata quality evaluation and curation are manually done by metadata curators at two different stages of the data ingestion process. The first is before transforming data into semantic models: the

[15] https://www.getty.edu/research/tools/vocabularies/tgn/
[16] https://viaf.org

control process focuses on the quality check of the source information before the data mapping. The second is before the final ingestion to check the quality of the reconciled data. Thanks to the usage of mapping templates and the generation of fixed URIs, the information provided by SARI infrastructure can be shared and reused across different domains.

2.3.2 Mapping Manuscripts Migration

The Mapping Manuscript Migrations (MMM) project [10], implements a Linked Open Data (LOD) framework to aggregate three datasets related to the history and provenance of medieval and renaissance manuscripts:

- Schoenberg Database of Manuscripts [17]: A relational database containing more than 240,000 records for manuscript observations;
- Bibale [18]: A relational database containing nearly 13,000 manuscript records;
- Medieval Manuscripts in Oxford Libraries [19]: A collection of more than 10,000 XML documents.

Data are aggregated using a set of mapping templates developed in CIDOC-CRM and FRBROO[20] ontologies and consisting of five main entities: Manuscripts, Agents, Works, Places, and Events. In addition, Agents and Places are automatically reconciled through the use of the Thesaurus of Geographic Names from Getty for the Places and the Virtual International Authority File for Agents. Therefore, we cannot define this workflow as a real pipeline to control the quality of metadata but rather as a solution to make available metadata according to a certain data structure in a specific domain.

[17] https://sdbm.library.upenn.edu/
[18] http://bibale.irht.cnrs.fr/
[19] https://medieval.bodleian.ox.ac.uk/
[20] http://www.cidoc-crm.org/frbroo/home-0

2.4 Chapter Summary

This Chapter focuses on the problem of metadata quality providing a detailed description of the approaches and methodologies used to deal with this issue.

Two main aspects are highlighted:

- It is not possible to consider or treat metadata quality control as a generic process that can be materialized in a hybrid framework or tool. It must be contextualized based on the domain and how the digital resources are catalogued;

- The methodologies implemented so far are not standardized and interoperable with each other. They cannot give comparable results to establish which framework is the most effective. This is mainly because of the different interpretation that has been given to the dimensions for metadata quality control. Figure 2.5 summarizes this concept. For example, Completeness is used in all analyzed frameworks while Accuracy is used by Moen et al., Bruce and Hillmann, Stvilia et al. and Ochoa, and Duvall. Lastly, some frameworks that have specific dimensions as Moen et al. and Stvilia et al.

In Section 2.1 we present how the existing guidelines are used by metadata creators and archivists to implement and maintain metadata repositories or dataset. In Section 2.2 we analyze the six most used frameworks used to evaluate the quality of metadata information, underlining how the seven quality dimensions proposed by Bruce and Hillmann in metadata Quality Framework are differently used and interpreted in each framework.

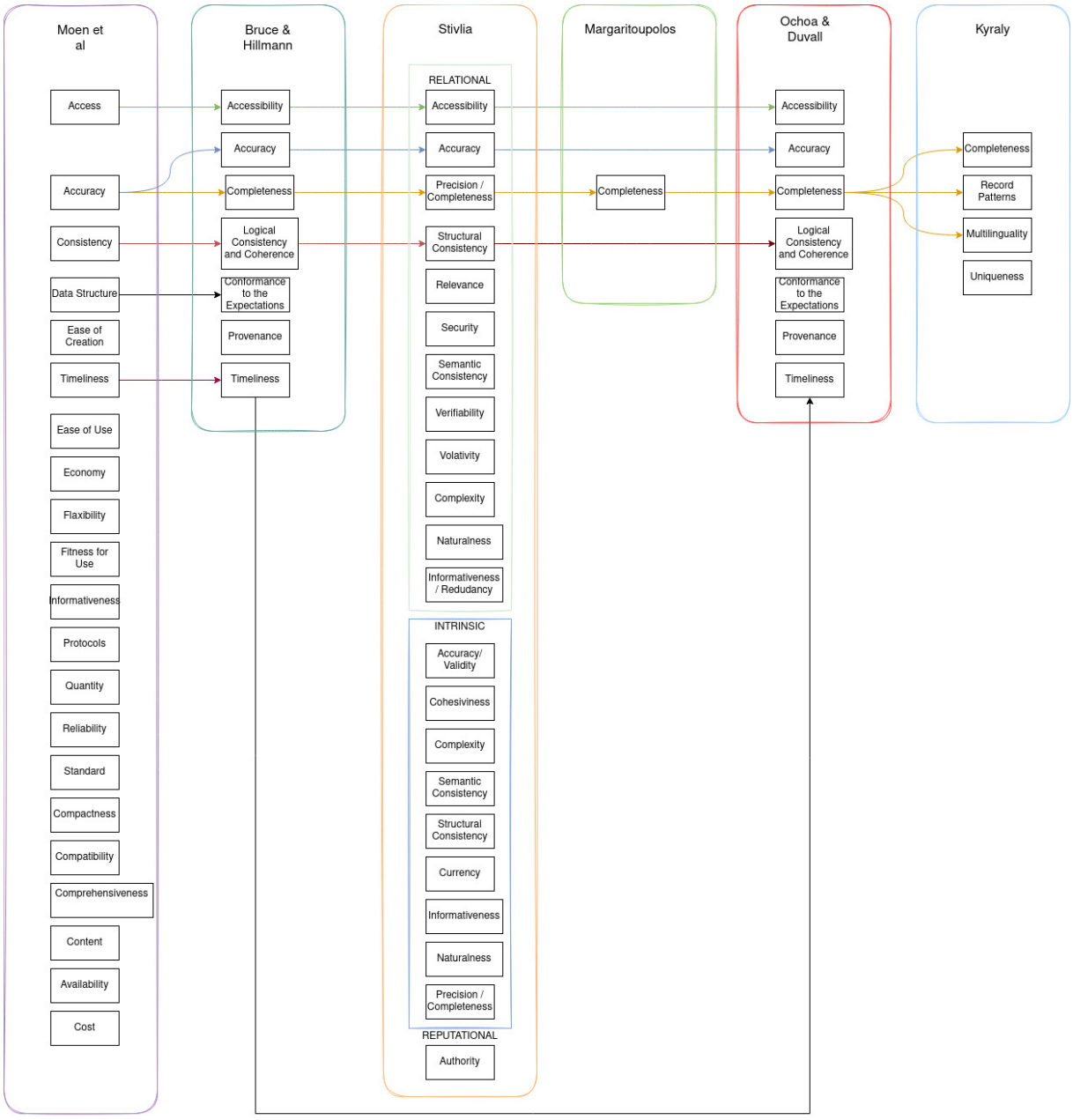


Figure 2.5: Metadata frameworks graph with alignment between quality dimensions

Chapter 3

The Completeness dimension

This Chapter is based on our work presented in:

- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Proposta per una valutazione automatica della completeness dei metadati nel contesto delle biblioteche digitali. *DigItalia*, 2:159–167, 2020
- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. On assessing metadata completeness in digital cultural heritage repositories. *Digital Scholarship in the Humanities*, 36:182–188, 11 2021

We investigate the role of context and the relevance of metadata elements into the evaluation of the Completeness dimension.

3.1 Introduction

Completeness is the *condicio sine qua non* to have meaningful object descriptions and it can be considered as a prerequisite to assess quality: incomplete records are by default low-quality resources due to lack of essential information. Numerous metadata standards have been established in an attempt to define sufficient descriptions of a resource from different perspectives and satisfy diverse functionalities. Theoretically, a sufficient description exists when all metadata elements of a standard are populated

with values. However, in practice, this is not what happens in the real world. Relevant surveys [26, 50] shows that metadata creators tend to fill out only particular metadata elements depending on the following factors:

- Popularity: The usage of specific guidelines should be considered as a standard procedure in implementing metadata records in a digital archive. However, this practice is not always adopted and metadata creators have the tendency to fill the most valuable elements like the title, the subject, or the description, while they ignore other less popular elements such as the provenance or the material and technique;

- Domain-Specific: In cultural heritage, the majority of digital archives are multi-domain and the adopted metadata schema is used to manage different domains, as for for example visual artworks, archaeology and architecture. However, beside the compulsory elements which are cross-domain, some of the recommended or optional elements might be applied, for example, to the description of visual artworks but not to the description of archaeology resources. Such metadata elements are simply not necessarily useful to cover a domain of interest. The element *dc:author* from Dublin Core is an example. Normally the field is used to provide the name of the author of a certain painting, drawing etc. However, in archaeology, the attribution of an artifact, aside from the production of some specific ceramics where the name of the potter is known as for *Maiolica di Montelupo*, it is almost impossible to identify the author of the artifact. Therefore the element is not filled or used;

- Metadata schema or Application Profile: Depending on metadata schema, some elements are considered as compulsory for the description of the resources, while others are considered as recommended or optional. For example, the 94 elements of the PICO application pro-

file[1] used to describe the resources in Cultura Italia are divided into compulsory (8 elements), recommended (10 elements) and optional (76 elements). In such cases, the trend is to fill mostly the compulsory elements.

Moreover, the creation of metadata is a task requiring major labor and financial cost and, most importantly, the involvement of knowledgeable and experienced people [38, 3]. Since all these requirements are generally difficult to be fully met, it is rather common, in the majority of digital repositories, to have incomplete metadata.

3.2 State of the Art

Within the framework discussed in the Section 2.2, the Completeness dimension is commonly evaluated as the *ratio* of the filled elements with respect to the metadata profile used for the description of the resources. The more elements are used, the higher is the quality of the information.

To operationally compute the Completeness dimension, two main approaches have been proposed [45, 34, 53]:

Binary Assessment

The presence or absence of metadata elements is computed with a binary assessment, assigning either 0 or 1 depending on the presence of specific metadata [53]. This methodology follows two dimensions of analysis. The first dimension focuses on evaluating if the metadata element is filled (value 1) or not (value 0)[53], while the second dimension of analysis goes deeper into the hierarchical structure of metadata schema, taking into consideration also the multi-valued elements with cardinality greater than 1 as the

[1] `http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/picoap/picoap1.0.xml`

language. In this second dimension, a field is considered complete if all the sub-elements are filled [45].

Consider for example the following *dc:type* element:

```
<dc:type xml:lang=‘‘it”>dipinto</dc:type>
<dc:type xml:lang=‘‘en”></dc:type>
```

According to the first dimension, it could be considered as complete since the *dc:type* element is filled. On the contrary, according to the second dimension it is considered non-filled since the attribute “en” has no value. In both cases, Completeness is computed as the weighted average of the filled elements with respect to a metadata schema.

Customized Score

For each metadata element a custom score is defined according to its importance with respect to the metadata profile. An example of this approach is represented by the computation of the Completeness dimension of the aggregated resources in the Europeana collection [33].

On the first level, all metadata records are analyzed. The score value assigned to each field is 1 except for the *rdf:about* attribute (that identifies the individual entities) to which a value score 10 is assigned. To obtain the overall weight, the sum of the obtained score is divided then by the number of entities from the metadata profile.

This kind of approaches, even if they allow metadata curators to check the quality status of a single record or, more generally, of a dataset, do not try to represent whether low metadata quality is related to the lack of few metadata with with high relevance or to the lack of many metadata elements with low relevance. However, such metrics should enable users and curators to define in a flexible way what metadata they deem more relevant in the overall evaluation of Completeness, to make this value com-

parable across different repositories, and to allow for a fine-grained analysis of metadata elements.

3.3 Completeness Assessment

As pointed out in Sections 1, 2, and 3, to have a clear and exhaustive picture of the overall Completeness quality, a simple check on the presence or absence of metadata elements is not enough. In the evaluation, the workflow should be considered, as well as other aspects such as the the resource domain and the context in which resources are created. The goal of the proposed methodology is to define a flexible way to compute Completeness that takes into account the structure of the metadata scheme as well as the specific topic of a collection. Specifically, we address the following research question **RQ1**:

- Can we define a way to compute Completeness that takes into account the context and relevance of metadata elements for a metadata collection?

The methodology we propose to evaluate metadata Completeness consists of the following key steps: *i)* The examination of metadata schema, *ii)* the score definition and computation, *iii)* the graphical representation of the obtained results.

Examination of metadata Schema

Depending on the adopted metadata standard, each metadata schema or application profile has its own specifications and characteristics. Therefore, in this first phase, the usage of metadata standard guidelines and the technical documentation on how to implement metadata schema requested for the metadata harvesting process is essential.

Given a repository to be evaluated, metadata elements are divided into groups, representing their importance:

- Compulsory: The minimal set of entities that must be implemented to guarantee metadata harvesting into a digital repository. For example *dc:title* or *dc:type*;
- Recommended: The elements which are not compulsory for metadata harvesting but are strongly recommended to improve the description of the resources. For example *dc:description, dc:subject* and *dc:coverage*;
- Optional: The elements which are not meant to provide descriptive information about the resource and its content but only technical information; for example, about the license or the date of the last modification;
- Domain-Specific: The peculiar metadata elements which are used to provide specific domain information as material and technique.

Score Definition and Computation

The presence or absence of metadata elements is computed with a binary assessment, assigning either 0 or 1 depending on the presence of the metadata element.

For each object o in the repository, a separate Completeness score $C_g(o)$ is computed for each defined metadata group G as follows: the number of filled G metadata for that object is divided over the total number of G metadata elements. For instance, if an object o has 3 out of 10 of the compulsory metadata filled, $C_compulsory(o) = 0.3$.

The resulting value is a real number between 0 and 1: the closer this value is to 1, the more complete the description of the object for that metadata group.

Checking metadata quality according to Completeness scores for the various metadata groups gives to metadata curators the possibility to have a complete view of the overall status of metadata quality. Curators can subsequently fix the objects with a low score, evaluating the different problems which contribute to the quality of the dataset.

Completeness Representation

To have also a graphical overview of the Completeness of a dataset, for each metadata group a separate barplot is drawn, having on the x-axis 10 intervals representing Completeness score ranges (e.g., 0-0.1; 0.1-0.2; ..., 0.9-1.0) and on the y-axis the percentage of the objects in the whole dataset having that Completeness score.

3.4 Metadata Completeness in Cultura Italia

To put in practice the proposed metadata Completeness methodology and prove its applicability and effectiveness, we apply the metrics on two specific datasets from Cultura Italia: MuseID-Italia (containing 76,828 records) and Regione Marche datasets (containing 90,602 records). Both datasets mainly deal with the visual artwork domain.

The datasets used to test this methodology consist only of non-aggregated and single fields. Like all the other datasets in Cultura Italia's repository, the metadata resources in MuseID-Italia and Regione Marche are directly implemented as PICO resources and ingested in Cultura Italia without intermediate aggregation steps as specified by the metadata aggregation policy adopted in Cultura Italia[2]. Moreover, to avoid redundancy of metadata elements, multiple values should be grouped into one metadata element

[2] http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/progetto/sintesi_progettotecnicoscientifico.pdf

[47]. Taking as example the metadata element "*pico:materialAndTechnique*" and the values "argilla, argilla depurata, tornio", the information should be represented using one element for the three values:

```
<pico:materialAndTechnique xml:lang=''it">
argilla , argilla depurata , tornio
</pico:materialAndTechnique>
```

and not three separate definitions as

```
<pico:materialAndTechnique xml:lang=''it">
argilla
</pico:materialAndTechnique>
<pico:materialAndTechnique xml:lang=''it">
argilla depurata
</pico:materialAndTechnique>
<pico:materialAndTechnique xml:lang=''it">
tornio
</pico:materialAndTechnique>
```

Without this rule, in the case of multi-valued fields, it should be considered whether the field is complete when only one instance is filled or not as mentioned by Margaritopoulos et al. in [46]. However, depending on the granularity of the quality check that curators want to apply, Completeness can be computed with this approach both at the level of the root element and at the level of the aggregated elements (e.g., defining mandatory only some sub-elements, and optional the others), as well as for multi-value fields (e.g., considering a multi-value field complete if it consists of at least a value, or an expected number of values).

Examination of metadata Schema

Relying on the guidelines[3] on how to implement metadata schema in Cultura Italia, the 94 PICO elements are divided into compulsory (8 elements),

[3] http://www.culturaitalia.it/opencms/documentazione_tecnica_it.jsp?language=it&tematica=static

recommended (10 elements) and optional (76 elements). The domain-specific elements are chosen instead among the PICO optional elements that are relevant for a specific domain, and therefore should be preferably filled for objects of datasets in that domain. In this case visual artwork (11 elements).

While the distinction between compulsory, recommended and optional elements is defined into metadata harvesting guidelines adopted in Cultura Italia, the 11 domain-specific elements are validated together with metadata curators from Cultura Italia considering also the standard guidelines[4] provided by the ICCD about the implementation of the Art Object card (OA). The table 3.1 shows in detail the metadata group.

Completeness Computation and Results

In this section we present the results obtained by applying the methodology described in the Section 3.3. Before analyzing in detail the Completeness of the considered datasets according to the proposed metadata group, the methodology investigates the frequency of usage of PICO metadata elements in the records of the collections.

[4] http://www.iccd.beniculturali.it/it/ricercanormative/29/oa-opere-oggetti-d-arte-3_00

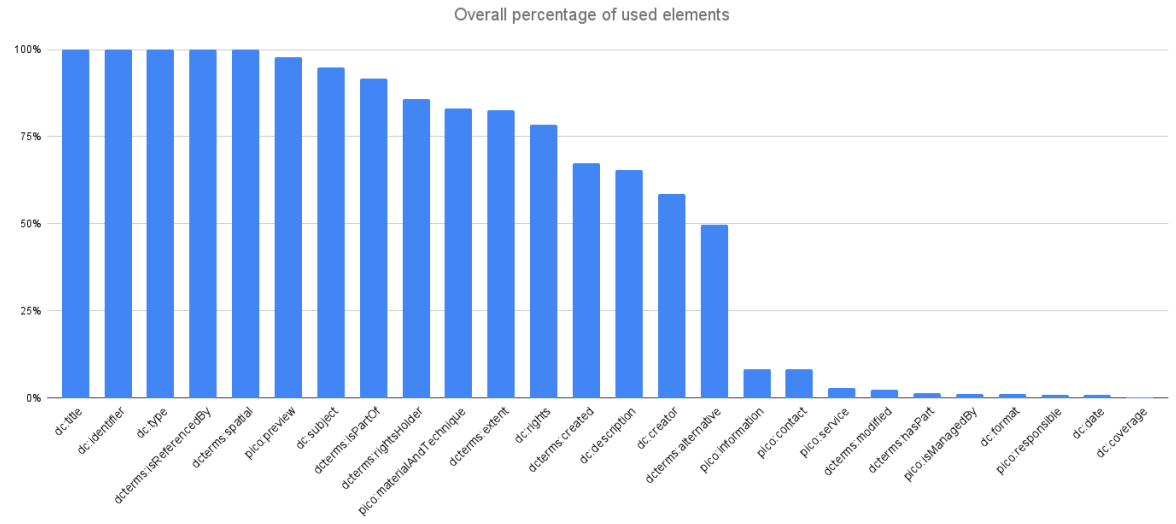


Figure 3.1: Percentage of records in the MuseID-Italia dataset having a given metadata element

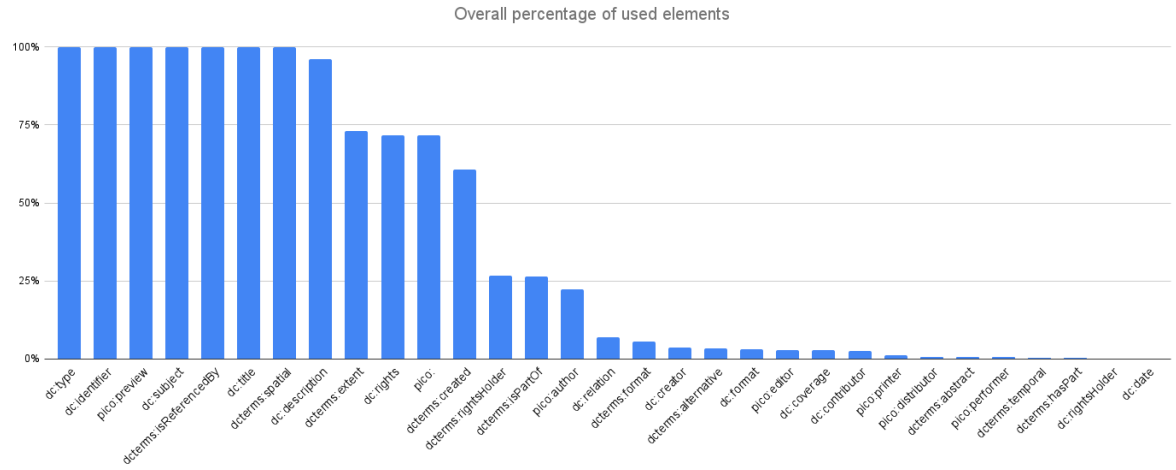
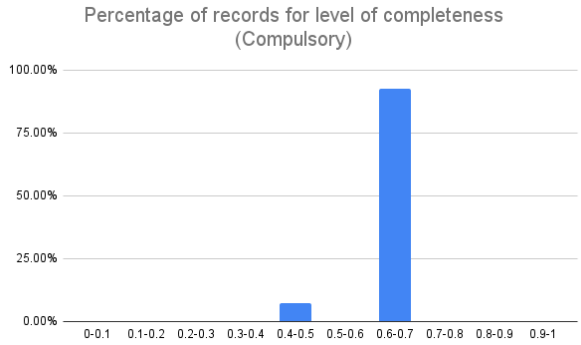


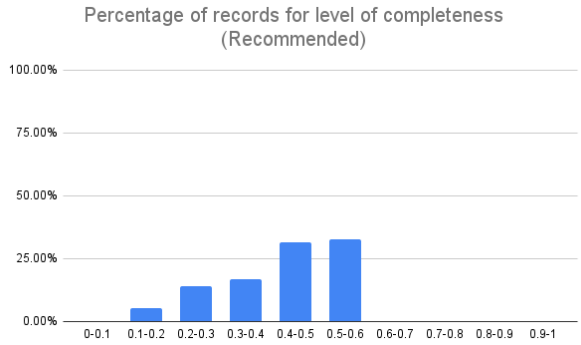
Figure 3.2: Percentage of records in the Regione Marche dataset having a given metadata element

Figures 3.1 and 3.2 graphically represent the percentage of records in the two datasets (MuseID-Italia e Regione Marche) having the given metadata elements. It is possible to notice that many (but not all) of the compulsory metadata are filled for all records in the datasets.

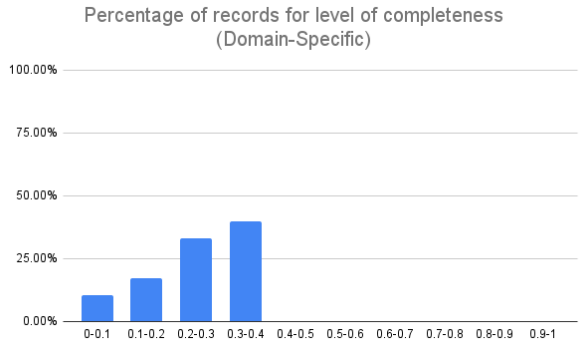
For metadata elements in the other groups, the percentage of records having those elements filled is substantially lower.



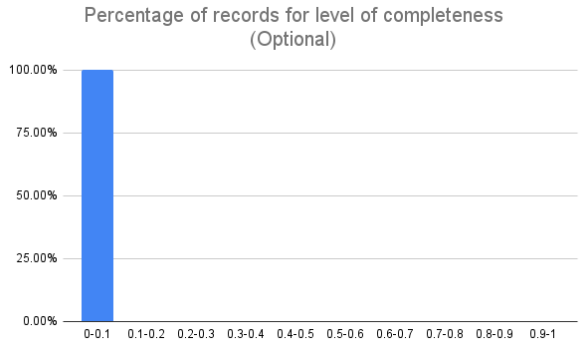
(a) Compulsory elements



(b) Recommended elements



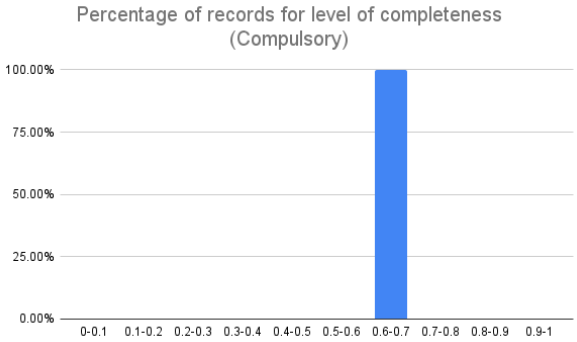
(c) Domain Specific elements



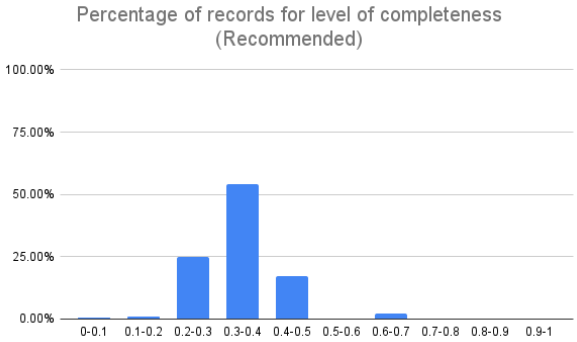
(d) Optional elements

Figure 3.3: Completeness plots for MuseID-Italia dataset

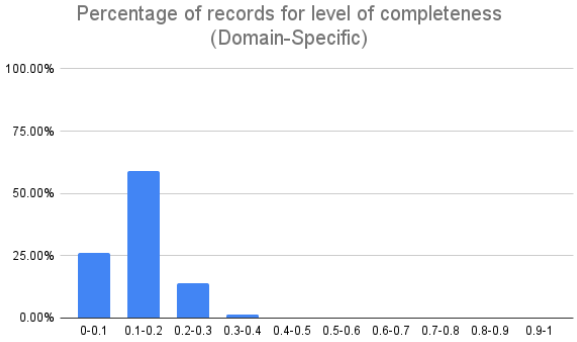
Then, given the four metadata groups proposed in Section 3.3, we compute for each record in the datasets four completeness scores (one for each metadata group), and separately for each metadata group, we analyze the distribution of the resulting completeness scores over the datasets, by plotting the aggregated bar plots as described in Section 3.3. These plots are reported in Figures 3.3 and 3.4. The plots show that the results obtained for Completeness on the two datasets are generally low. For example, in the MuseID-Italia dataset in Figure 3.3a, most of the records (92%) obtain a Completeness score for the compulsory group in the range between 0.6



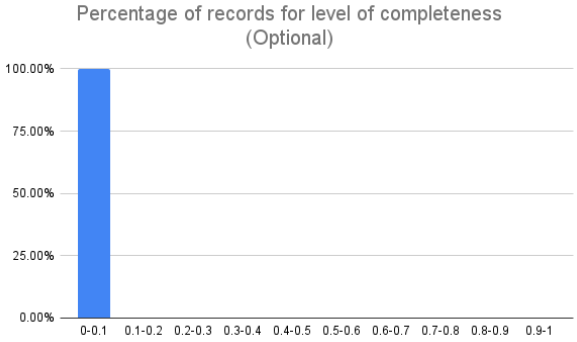
(a) Compulsory elements



(b) Recommended elements



(c) Domain Specific elements



(d) Optional elements

Figure 3.4: Completeness plots for Regione Marche dataset

and 0.7, while for Regione Marche in Figure 3.4a all records achieve for the same metadata group a score between 0.5 and 0.6. The same can be observed for the domain-specific schema (between 0.1 and 0.4). All records in the datasets rarely use elements of the optional metadata group, while the usage of recommended and domain-specific metadata elements varies.

These results also underline that the metadata of the two datasets are partially FAIR. As reported in Section 2.1, to have meaningful and useful data, metadata should guarantee Accessibility, Findability, Interoperability and Usability of the information provided by the digital object. By having a low Completeness only the Accessibility principle is covered. In detail:

- Findability: Data should be easy to find and discoverable with meta-

data. To have this condition, a complete data structure is a necessary condition. For example, in the Regione Marche dataset, only 21% of the resources are filled by using the *pico:author* element. This means that, in most cases (79%), the end-users will not be able to filter the resources by "Author" or perform a full-text search by typing the name of the artist.

- Interoperability: Allows data exchange and reuse between researchers, institutions, organisations, or countries. Often the concept of interoperability is limited to the usage of the same standard metadata schema to describe a certain type of resource; the Dublin Core scheme for example is one of them. However, especially in the digital archive domain, to guarantee an interoperable data structure [12], it is essential to harvest and merge metadata between repositories [36], as for the case of Cultura Italia and Europeana. Having low Completeness does not facilitate the metadata harvesting process. For example the *dc:language* and *dc:source* elements are not represented in MuseID-Italia and Regione Marche dataset since they are both considered as Optional fields in Cultura Italia. However, the *dc:language* and *dc:source* elements are treated as compulsory by Europeana to allow proper metadata harvesting. In this case the metadata curator from Cultura Italia must manually intervene to fix the missing value.

- Usability: Resources should be sufficiently described and shared with the least restrictive licenses, allowing the widest reuse possible and the least cumbersome integration with other data sources. Uncomplete resources are per se unusable.

A line of action to solve the problem of low Completeness could be the adoption of metadata templates as in the Swiss Art Research Infrastructure and Mapping Manuscripts Migrations project by the metadata

creator to facilitate the creative process and guarantee, at least, the full interoperability with the others digital infrastructure.

3.5 Chapter Summary

This chapter focuses on the analysis of metadata Completeness for the description of digital resources. In Sections 3.1 and 4.2, we provided an overview of the main approaches used to address the Completeness evaluation with respect to metadata quality framework taken as a reference in this thesis and the causes that contribute to describing objects using a few metadata elements such as the popularity of the single metadata elements, the domain, and metadata schema.

Considering the current limitations of the approaches followed so far in Section 3.3, we described the methodology defined in this work to evaluate the Completeness dimension. The proposed methodology uses the context and the domain of metadata as the main focus of the analysis, highlighting how, using four different levels of evaluation, this analysis can be more precise and tailored for the need of metadata curator. Finally, in Section 3.4, we present the results of the metadata Completeness evaluation performed in the Regione Marche and MuseID-Italia datasets from the Cultura Italia repository. In the next chapter, we will focus our analysis on the second dimension: Accuracy.

Compulsory	Recommended	Optional	Domain Specific
dc:identifier	dc:creator	dc:contributor	dc:contributor
dc:isReferencedBy	dc:creator	dc:coverage	dc:coverage
dc:subject	dc:date	dc:format	dcterms:alternative
dc:title	dc:date	dc:language	dcterms:bibliographicCitation
dc:type	dc:description	dc:publisher	dcterms:created
dcterms:license	dc:rights	dc:relation	dcterms:isPartOf
pico:licenseMetadata	dc:rights	dc:source	dcterms:modified
pico:preview	dcterms:extent	dcterms:abstract	dcterms:replaces
dc:identifier	dcterms:rightsHolder	dcterms:accessRights	pico:commissioner
dc:isReferencedBy	dcterms:spatial	dcterms:accrualMethod	pico:materialAndTechnique
dc:type	dcterms:spatial	dcterms:accrualPeriodicity	pico:printer
pico:licenseMetadata	dcterms:temporal	dcterms:accrualPolicy	
	pico:author	dcterms:alternative	
	pico:object	dcterms:audience	
	pico:object	dcterms:available	
		dcterms:bibliographicCitation	
		dcterms:conformsTo	
		dcterms:created	
		dcterms:dateAccepted	
		dcterms:dateCopyrighted	
		dcterms:dateSubmitted	
		dcterms:educationLevel	
		dcterms:hasFormat	
		dcterms:hasPart	
		dcterms:hasVersion	
		dcterms:instructionalMethod	
		dcterms:isFormatOf	
		dcterms:isPartOf	
		dcterms:isReferencedBy	
		dcterms:isReplacedBy	
		dcterms:isRequiredBy	
		dcterms:issued	
		dcterms:isVersionOf	
		dcterms:mediator	
		dcterms:medium	
		dcterms:modified	
		dcterms:provenance	
		dcterms:references	
		dcterms:replaces	
		dcterms:requires	
		dcterms:tableOfContents	
		dcterms:valid	
		pico:anchor	
		pico:commissioner	
		pico:contact	
		pico:contributesTo	
		pico:dateOfBirth	
		pico:dateOfDeath	
		pico:digitises	
		pico:distributor	
		pico:editor	
		pico:hasAsContributor	
		pico:hasAsResponsible	
		pico:information	
		pico:isDigitisedBy	
		pico:isManagedBy	
		pico:isOwnedBy	
		pico:isOwnerOf	
		pico:isPerformedBy	
		pico:isProducedBy	
		pico:isPromotedBy	
		pico:isResponsibleFor	
		pico:manages	
		pico:materialAndTechnique	
		pico:performer	
		pico:performs	
		pico:placeOfBirth	
		pico:placeOfDeath	
		pico:printer	
		pico:producer	
		pico:produces	
		pico:promotes	
		pico:responsible	
		pico:service	
		pico:translator	

Table 3.1: Detailed PICO metadata group

Chapter 4

The Accuracy Dimension

This chapter is based on our work presented in:

- Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Automatically evaluating the quality of textual descriptions in cultural heritage records. *International Journal on Digital Libraries*, 22(2):217–231, 2021

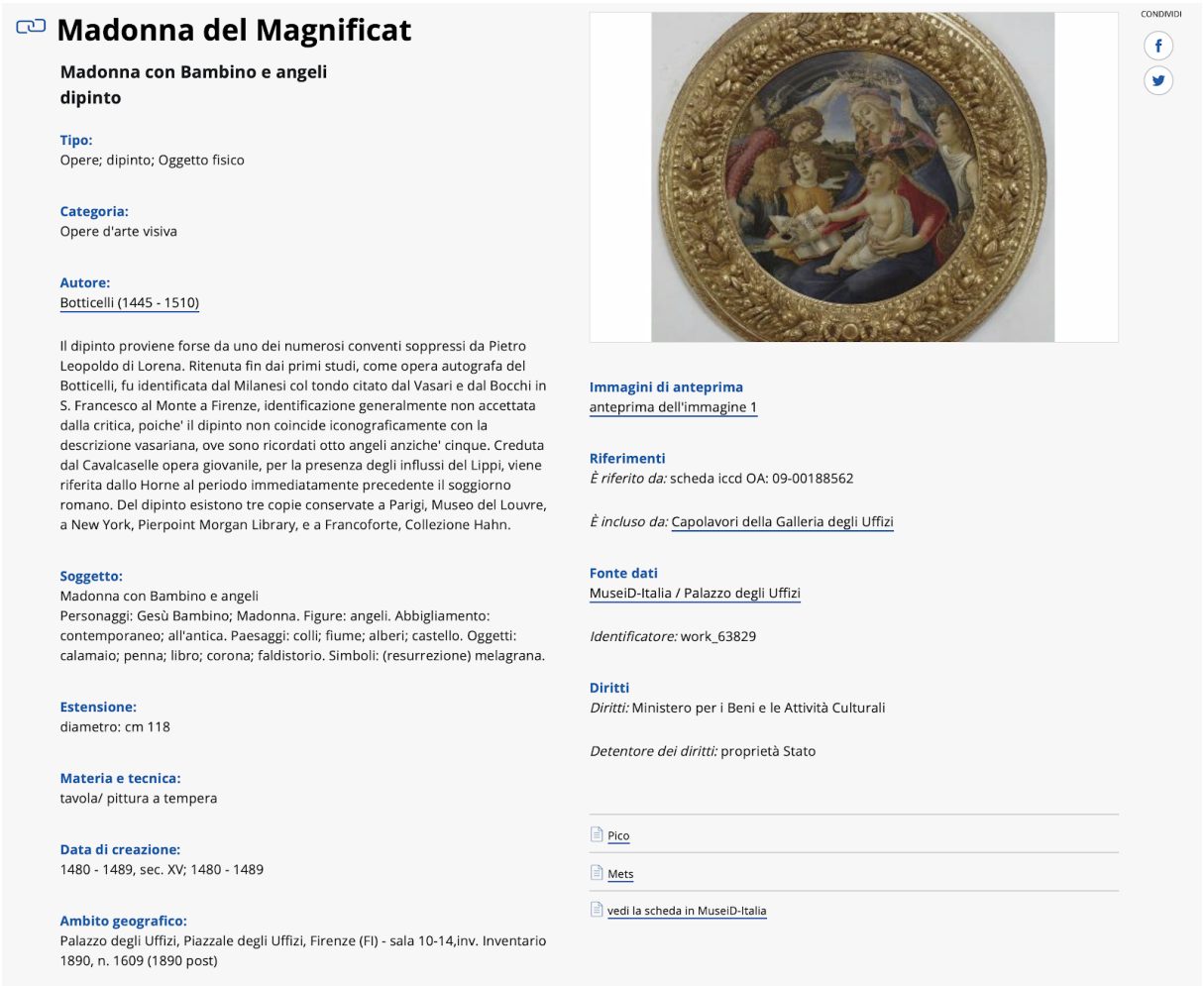
This article investigates the assessment of metadata Accuracy using computational methods based on machine learning.

4.1 Introduction

As already anticipated in Section 1, in evaluating the Accuracy, the context and the domain of a resource play a key role. This is from both the creator's and the curator's point of view. Depending on the domain, the metadata creator could indeed provide more detailed information for some metadata elements.

Let us consider as an example the resource id work_63829[1] from Cultura Italia. The record refers to the Italian masterpiece, "Madonna del Magnificat" from Sandro Botticelli represented in Figure 4.2.

[1] `https://bit.ly/3cKUcXu`



Madonna del Magnificat

Madonna con Bambino e angeli
dipinto

Tipo:
Opere; dipinto; Oggetto fisico

Categoria:
Opere d'arte visiva

Autore:
Botticelli (1445 - 1510)

Il dipinto proviene forse da uno dei numerosi conventi soppressi da Pietro Leopoldo di Lorena. Ritenuta fin dai primi studi, come opera autografa del Botticelli, fu identificata dal Milanesi col tondo citato dal Vasari e dal Bocchi in S. Francesco al Monte a Firenze, identificazione generalmente non accettata dalla critica, poiche' il dipinto non coincide iconograficamente con la descrizione vasariana, ove sono ricordati otto angeli anziche' cinque. Creduta dal Cavalcaselle opera giovanile, per la presenza degli influssi del Lippi, viene riferita dallo Horne al periodo immediatamente precedente il soggiorno romano. Del dipinto esistono tre copie conservate a Parigi, Museo del Louvre, a New York, Pierpoint Morgan Library, e a Francoforte, Collezione Hahn.

Soggetto:
Madonna con Bambino e angeli
Personaggi: Gesù Bambino; Madonna. Figure: angeli. Abbigliamento: contemporaneo; all'antica. Paesaggi: colli; fiume; alberi; castello. Oggetti: calamaio; penna; libro; corona; faldistorio. Simboli: (resurrezione) melagrana.

Estensione:
diametro: cm 118

Materia e tecnica:
tavola/ pittura a tempera

Data di creazione:
1480 - 1489, sec. XV; 1480 - 1489

Ambito geografico:
Palazzo degli Uffizi, Piazzale degli Uffizi, Firenze (FI) - sala 10-14,inv. Inventario 1890, n. 1609 (1890 post)

CONDIVIDI

Immagini di anteprima
anteprima dell'immagine 1

Riferimenti
È riferito da: scheda iccd OA: 09-00188562

È incluso da: Capolavori della Galleria degli Uffizi

Fonte dati
MuseiD-Italia / Palazzo degli Uffizi

Identificatore: work_63829

Diritti
Diritti: Ministero per i Beni e le Attività Culturali

Detentore dei diritti: proprietà Stato

Pico

Mets

vedi la scheda in MuseiD-Italia

Figure 4.1: Madonna del Magnificat from Cultura Italia

As can be seen, the information provided by the resource is complete and the quality of the information is accurate: all the compulsory, optional, and recommended elements are filled and the provided information is correct. In this case, metadata quality can be considered as high-quality. However, even though all the information is correct, the textual description of the painting focuses more on the historiographic aspect of the painting by telling some aspects of the history behind the artifact, for example, how it was attributed to Botticelli. Per se it is certainly an interesting information but also incomplete, since the real description of the main scene

depicted on the painting is not described. This issue affects mainly the textual field, since the description has no particular restrictions, while in the fields where the information should be preferably be managed via a taxonomy or a controlled vocabulary as the subject this problem is obviously less evident. Using the same methodology also in the case of descriptions would lead to an underlying problem: the descriptions contained in the vocabulary should be also be compliant with the guidelines provided by the archive for the metadata ingestion. For example in Cultura Italia the description should be provided according to the standard guidelines from the Istituto Centrale per il Catalogo e la Documentazione (ICCD) (See section 4.3.1). Therefore the terms of the vocabulary should also follow the same description structure. Otherwise the description will be, in any case, considered as low-quality.

4.2 State of the Art

With respect to the metadata quality frameworks analyzed in this thesis and described in Chapter 2, the Accuracy dimension is adopted, with three different definitions, namely by Bruce and Hillmann, Ochoa and Duvall and Stvilia et al. while in the framework presented by Király Accuracy is not introduced as a separate metric, but only mentioned as a dimension that can be inferred from the other dimensions.

Bruce and Hillmann's framework points to the fact that "*The information provided about the resource in metadata instance should be as correct as possible [...] Typographical errors, as well as factual errors, affect this quality dimension.*" This is however a very narrow definition of Accuracy, which only takes into account some surface features of a description (e.g., presence of mistakes), without considering that a description can be formally perfect without containing useful information, therefore being of

low-quality.

Ochoa and Duvall define it as *"the degree to which metadata values are "correct", e.g. how well they describe the object."* [53]. Similar to the approach we present in this thesis, they make use of text processing techniques and apply them to textual fields of metadata. However, they propose a general unsupervised method based on Vector Space Model (VSM), aimed at finding the semantic distance between two resources according to the keywords stored in a vocabulary. Our approach, instead, is supervised and does not rely on external resources, because this information is already inferred by the trained classification model. Furthermore, Ochoa and Duvall's proposal to assess metadata Accuracy may be affected by issues related to the length of the descriptions. Longer texts tend to contain more words than shorter ones, and this has an impact on the computation of the semantic distance with the keywords stored in the external vocabulary: the longer the text, the higher the chances that it contains some of the keywords in the vocabulary, and thus the higher the Accuracy score (due to the way the VSM works). This means that lengthy (but not necessarily accurate) descriptions containing many keywords may score higher Accuracy than shorter (but accurate) descriptions. Moreover, Ochoa and Duval present also three validation studies to evaluate the proposed metrics with respect to human-made quality assessment. In general, the quality metrics do not seem to correlate with human ratings.

For Stvilia et al., Accuracy represents a dimension that measures relationships between the information and some aspects of its usage. However, there is no implementation of the Accuracy dimensions as an algorithm that can be operationally applied to different cases. In general terms, metadata Accuracy is measured as the extent to which the data values in the metadata record match with the characteristics of the described object [61].

4.3 Accuracy Assessment

The methodology we present in this thesis aims to automatically assess the Accuracy of the textual description element (typically encoded using the *dc:description* from the Dublin Core[2] metadata schema) of a given cultural heritage object. More specifically, we assess the Accuracy evaluation of such description metadata by determining whether the field contains a high-quality or low-quality description of the considered object, measured as the compliance of the textual content with the description rules from Istituto Centrale per il Catalogo e la Documentazione (ICCD), adopted in the Cultura Italia portal.[3]

As the first step in this direction, we create a large dataset of object descriptions, (semi-)automatically labelled as being of high-quality or not. An example of high-quality and of low-quality descriptions are reported in Table 4.1. In the first, all and only the necessary information related to the object (e.g., the frame) and the subject (the person portrayed in the painting) are reported. The second description, instead, is a lengthy text that focuses first on the painter and only towards the end mentions the subject of the painting. More details on the methodology and guidelines we followed for judging the quality of a description are discussed in Section 4.3.1.

As a second contribution, we exploit natural language processing techniques and machine learning to create a binary (high-quality vs. low-quality) classification model that can assess the quality of unseen descriptions by predicting the class they should belong to. To this purpose, we compare two different classification algorithms — Support Vector Machine (SVM) [14] and the FastText logistic regression classifier [30] — leveraging the representation of descriptions as word embeddings, e.g., as real-valued

[2]https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[3]`http://www.culturaitalia.it`

Quality	Record ID	Original Italian Description	English Translation
High	iccd2225343	Dipinto entro cornice lignea verniciata ocra con bordo interno dorato. Amedeo III è raffigurato di profilo in armatura scura con ceselli in oro, mascheroni dorati sulle spalle e sull'elmo, cimiero con piume rosse e bianche. Nella parte inferiore del dipinto fascia con iscrizione a caratteri stampatello. Personaggi: Amedeo III di Savoia	Painting within an ocher painted wooden frame with a inner golden border. Amedeo III is depicted in profile with a dark armor chiseled in gold, golden figurehead on the shoulders and on the helmet. Crest with white and red plumage. On the lower part of the painting inscription with block letters. Characters: Amedeo the 3rd of Savoy.
Low	work82865	Congdon si è raramente dedicato al disegno come forma espressiva autonoma, cosi la mole di disegni raccolti sui taccuini non sono altro che appunti visivi presi durante numerosi viaggi. In questo senso non è possibile, se non raramente, assegnare al singolo disegno un'opera finita direttamente corrispondente, cosi questi disegni non vengono nemmeno ad essere schizzi preparatori. La sommatoria di tutti i disegni relativi a un luogo danno origine a una serie di dipinti che non hanno un corrispettivo oggettivo nei disegni stessi. Tutto questo giustifica la presenza degli appunti all'interno delle immagini (colori, sfumature e spiegazioni di vario genere). Nel caso probabile veduta di Napoli eseguita durante un viaggio del 1951.	Congdon has rarely devoted himself to drawing as an autonomous expressive form, so the drawings in his notebooks are nothing more than visual sketch taken during his numerous trips. Rarely it is possible to assign to the single drawing the corresponding attributes as finished art work since they represents the base idea for others drawings or paintings. The collection of all the drawings related to a place give rise to a series of paintings that do not have a direct mapping to the drawings themselves. All this justifies the presence of notes inside the images (colors, shades and explanations of various kinds). In this case, probably, a view of Naples from 1951.

Table 4.1: Example of high-quality and low-quality descriptions from the dataset we built starting from Cultura Italia portal.

vectors in a predefined vector space that compactly captures meaning similarity. We perform the comparison on three different cultural heritage domains: visual artworks, archaeology and architecture. While text analysis and machine learning have already been applied to metadata quality assessment [51], recent advances in language modelling, in particular the use of word embeddings [48], have not been explored for the task. This novel way to capture the semantic content of descriptions, together with supervised machine learning, is exploited in this work addressing the main research question (RQ2):

- Can we effectively assess metadata Accuracy of textual descriptions using computational methods based on machine learning?

The goal is to provide some insights into which techniques and algorithms can be effectively used to support curators in the manual quality control of cultural heritage descriptions.

Moreover the methodology presented here aims also to provide guidance in the creation of datasets for performing this task in a supervised setting, taking into account also the characteristics of different domains. We investigate these peculiar aspects addressing the following specific research questions:

- Research Question 2.1 (RQ2.1): Which machine learning algorithm should be used to assess the quality of cultural heritage descriptions approximating as much as possible human judgement?
- Research Question 2.2 (RQ2.2): Can a classification model trained with descriptions in a given cultural heritage domain be effectively applied to automatically assess description quality in other domains?
- Research Question 2.3 (RQ2.3): How many annotated resources are needed to create enough training data to automatically assess the quality of descriptions?

With RQ2.1 we compare different classification algorithms and natural language processing techniques. With RQ2.2 we investigate how classification performance changes when using data from different domains, even in a combined way. Finally, with RQ3 we aim to provide guidance in applying supervised techniques to novel datasets, by assessing how the dimension of the training data affects classification quality, and therefore suggesting how many instances should be manually annotated.

First, given the lack of an adequate resource, we develop a dataset for training and testing machine learning approaches: the dataset consists of object descriptions manually labeled by an expert annotator as high/low-quality according to the adherence to the cataloguing guidelines of the digital repository indexing the objects. Secondly, we run several experiments to address the aforementioned research questions, assessing system performances using well-known metrics (e.g., precision, recall, F1-measure) and adopting evaluation protocols aiming to reduce possible biases (e.g., cross-validation setting, removal of duplicates). Finally, we analyze the learning curve of the best classification model, by incrementally adding new instances to the training data.

4.3.1 Dataset Description

The usage of machine learning techniques in the cultural heritage domain is still limited, especially in the evaluation of metadata quality. The lack of adequate datasets is one of the main reasons for this. A recent survey [22] has shown that most of the datasets are not publicly available and are focused on images rather than on textual elements. For this reason, we develop a new dataset tailored to evaluate the Accuracy of metadata quality.

Resource Collection

To create the dataset, we use the textual descriptions[4] from the Cultura Italia repository. These records include mainly data from "Musei d'Italia" and "Regione Marche" datasets, which have been chosen because they contain a high number of non-empty *dc:description* elements [5].

Dataset Structure

The dataset is composed of 100,821 descriptions after duplicates removal. Duplicates are removed for two reasons: this reduces annotation effort in the subsequent manual annotation, and avoides that the same examples appear both in the training and in the test set, a situation that could make classification biased and lead to an inaccurate evaluation in supervised settings.[6] Duplicated descriptions were mainly short and of low-quality, reporting few generic words to describe an item (e.g., "Mensola.", "Dipinto.").

All these descriptions are about objects of different typologies and from different domains, and thus, are further organized into three specific domains:

- Visual Art works (VAW) (59,991 descriptions).
- Archaeology (Ar) (29,878 descriptions).
- Architecture (A) (10,952 descriptions).

[4]Encoded by the dc:description element from the Dublin Core metadata schema

[5]Only 47.8% of the resources of Cultura Italia have a filled dc:description element.

[6]This is a technical aspect to address in order to properly assess the classification performance, and does not hinder the application of the approach for assessing the quality of descriptions in collections where multiple items share the same textual content.

Dataset Annotation

To determine the quality of the collected descriptions, we rely on the standard cataloguing guidelines provided by the Istituto Centrale per il Catalogo e la Documentazione (ICCD), e.g. the same guidelines that should be followed by the data providers of Cultura Italia portal. More precisely, a specific section of the guidelines[7] addresses how to describe any cultural item, clarifying that both the object and the subject of the item must be presented in the description as follows:

Object : the object typology and shape must be described. To this purpose, the cataloguer must refer to the vocabularies provided by ICCD, using specific terminology (e.g., the technique used for paintings and drawings, or the material for the archaeological items);

Subject : the cataloguer must report the iconographic and decorative settings of the item, such as the characters of the depicted scene in a painting and their attribution. Other aspects (e.g., the history behind the painting or the painter) should not be included.

Following the above cataloguing guidelines, each textual description in our dataset is (semi-)automatically annotated as "high-quality" if object and subject of the item are both described according to the ICCD guidelines, and as "low-quality" in all other cases. More specifically, we annotate each harvested description according to the following rules:

- If the length of the description is less than 3 words, it is labeled as "low-quality" (e.g., 'Painting', 'Rectangular table', 'View of harbour'). This is done automatically based on the assumption that in few tokens it is not possible to describe both the object and the subject of a record. This concerns 5,349 descriptions, automatically labeled as "low-quality";

[7]OA card, DESO and DESS element: `http://bit.ly/ICCD_OA_card`

Dataset	High-Quality	Low-Quality (manual)	Low-Quality (auto)	Total
Visual Art Work	30,383	19,824	9,784	59,991
Archaeology	19,280	6,334	4,264	29,878
Architecture	6,908	1,842	2,202	10,952
Overall dataset	56,571	28,000	16,250	100,821

Table 4.2: Number of descriptions per domain labelled as High-Quality or Low-Quality. Low-quality descriptions have been identified both manually and following an automatic selection.

- If descriptions are coming from a collection not updated after 2012, they are very likely to be "low-quality". This assumption is based on my domain knowledge, being aware of the history of Cultura Italia collections and therefore being able to identify less curated batches of records. This assumption is practically confirmed randomly sampling 500 records from such collections, and manually checking each of them, confirming that none of the samples can be classified as "high-quality". This way 10,901 descriptions are automatically labeled as "low-quality";

- The remaining descriptions are then manually annotated one by one and labeled as "high-quality" or "low-quality".

Table 4.2 summarizes statistics of the annotated dataset and the size of the three domains. We show in a separate column ('Low-Quality (auto)') the number of descriptions with poor quality automatically identified based on their length or the year of the last update, as described above. Although low-quality descriptions are less represented than high-quality ones, there are enough examples in both classes to train a supervised system. Regarding human effort, the manual labelling task spanned around two years (partial time), at a pace of approximately 150 annotations per hour.

Validation of the Annotated Resources

As analyzed in the introduction to Chapter 4.1, the human factor is an aspect which contributes to different evaluations of metadata Accuracy. Although the annotation are made following the ICCD rules, following best practices in linguistic annotation and dataset creation [58], we compute inter-annotator agreement, in order to assess whether the task is sound or the concept of low and high-quality metadata is too subjective.

From the annotated dataset we select a balanced sample of 1,500 descriptions to be manually annotated also by one of the metadata curators of Cultura Italia. We then compare our annotation with the one from Cultura Italia. The inter-annotator agreement, computed according to Cohen's kappa [37], shows a very high level of agreement (16 diverging annotations over 1,500 description, $\kappa = 0.979$) between the two annotators. This confirms that the task can be confidently carried out by domain experts and that the quality of the resulting annotations is accurate. The resulting annotated dataset is publicly available[8] [40] under the terms of the Creative Commons Attribution-ShareAlike 4.0 Generic (CC BY-SA 4.0) license.

4.3.2 Classification Framework

Text classification is one of the fundamental tasks in machine learning and natural language processing with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. Consider for example this description of a Roman amphora: *A tall cylindrical amphora with angular shoulders, long straight handles and a collar rim. This form is the most important Italian wine amphora of the late Republican period, with a wide distribution around the Mediterranean (with many examples*

[8] https://figshare.com/articles/dataset/Annotated_dataset_to_assess_the_accuracy_of_the_textual_description_of_cultural_heritage_records/13359104

from shipwrecks) and across the north-west provinces. A text classifier can take this phrase as an input, analyze its content, and then automatically classify the text according to a specific label, such as "Archaeology" or "Roman Pottery". With respect to metadata Accuracy evaluation this approach is formulated as a binary classification task, where the two labels are "low-quality" vs "high-quality". How to benefit from this technology in metadata curation and evaluation processes have not yet been systematically analyzed or applied to a real world use case. The RQ2.1 aims to fill this gap analyzing the feasibility to use a classification framework to automatize Accuracy evaluation.

We experiment and compare two algorithms: Support Vector Machines (SVM) [14] and the FastText multinomial logistic regression classifier [30] (hereafter, MLR_{ft}). Both approaches are fed with the FastText embeddings [6] as input features. This means that no manually-engineered features have been used, but only those represented through word embeddings. We remark that in the FastText word embeddings, each word is represented as a bag of character n-grams in addition to the word itself, so that also out-of-vocabulary words (e.g., words never seen during the training of the model) are included in the representation, and information on suffixes and prefixes is captured.

4.3.3 SVM and MLR_{ft} Comparision Scenario

The goal of the classification framework discussed in this thesis is to automatically identify high-quality and low-quality descriptions in cultural heritage records. Methodologically the problem is addressed as a binary classification task adopting the annotated data presented in the Section 4.3.1 to train a supervised system able to assign an unseen description to one of the two classes, low-quality or high-quality.

Therefore, since our input data are natural language descriptions, we

first convert them into numerical vectors using the FastText word embeddings [6]: each word is assigned to a real-valued vector representation for a predefined fixed sized vocabulary, capturing the fact that words that have similar meaning have a similar vector representation, and the vector representation for each description (e.g., a collection of words) is obtained by averaging the vector representations of its words. The vector representation of each description can then be directly fed to machine learning classification algorithms.

Before sending the descriptions to the classifiers, a pre-processing step is performed, following best practices in text classification:

- Stopword removal: Stopwords include all terms that do not convey a semantic meaning such as articles, prepositions, auxiliaries, etc. These are removed from each description by comparing each token against a pre-defined list of Italian words imported from the NLTK Python library.[9]

- Punctuation removal: Following the same principle of stopword removal, each punctuation is removed from the descriptions.

Support Vector Machine (SVM)

Considering a binary classification problem, SVM learns to separate an n-dimensional space with a hyperplane into two regions, each of which corresponds to a class. The idea behind SVM is to select the hyperplane that provides the best generalization capacity: the SVM algorithm first attempts to find the maximum margin between the two data categories and then determines the hyperplane that is in the middle of the maximum margin. Thus, the points nearest the decision boundary are located at the same distance from the optimal hyperplane [1, 60, 59]. Different kernels

[9] `https://www.nltk.org/`

(e.g., learning strategies) can be used in a SVM, such as radial basis function (RBF) or linear: for this classification task, the best kernel is selected via grid search in the classifier optimization phase using the implementation available in the scikit-learn library [55]. The advantage of using RBF kernel for the training processes is that it restricts training data to lie in specified boundaries mostly used to solve problems related to a binary classification as in this case[10].

Since the classifier takes a feature vector in input, we convert each record description into a FastText embedding. The embedding of each description is built by averaging the FastText word embeddings of the single words in the description. For this step, we rely on pre-trained continuous word representations, which provide distributional information about words and have been shown to improve the generalization of models learned on limited amount of data [13]. This information is typically derived from statistics gathered from a large unlabeled corpus of textual data like Wikipedia or the GigaWord corpus. In our case we compare two different models, a *domain-specific* and a *general-purpose* one. The first is obtained by creating FastText embeddings from the corpus obtained by merging all textual descriptions used in our experiments, while the second is the Italian pre-trained model of FastText embeddings,[11] created from Wikipedia. Both models are trained in the same way, using continuous bag-of-word with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives. We also experiment with two different vector dimensions: 300, e.g. the default FastText number of dimensions, and 50, which we obtain by applying principal component analysis (PCA)

[10]The RBF kernel non linearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The RBF kernel has less numerical difficulties than polynomial kernel. Polynomial kernels are less widely used than the RBF kernel. This might be because under similar training and testing cost, a polynomial kernel may not give higher Accuracy.

[11]`https://fasttext.cc/docs/en/crawl-vectors.html`

[64] to the 300-dimensional embeddings.

FastText Implementation of the Multinomial Logistic Regression ($\mathbf{MLR_{ft}}$)

A second classification algorithm we consider is the implementation of multinomial logistic regression included in the FastText library[12] [30]. This is a linear classifier, developed by the Facebook Research Team, that was evaluated on various classification tasks (e.g., sentiment analysis, tag prediction) achieving performance score comparable to advanced deep learning models in terms of Accuracy, but orders of magnitude faster for training and evaluation.

Like in the SVM scenario, we compare two variants of MLR_{ft}: one fed with the FastText embeddings obtained by merging all textual descriptions of our corpus, and one fed with the Italian pre-trained FastText embeddings created from Wikipedia. Also in this case, embeddings of different dimensions, e.g. 300 and 50, are created and compared.

All the code used for running the classifiers and preprocessing the dataset is available on the GitHub code repository of the Accuracy evaluation dimension[13].

4.3.4 Baseline

According to the guidance provided by the ICCD, in object descriptions both the characteristics of the object and of the subject should be specified. This particular structure should also have an impact on the length of the resource descriptions. The barplot in figure provides an overview of the description length in the annotated dataset: on the x-axis the different length bins are reported, while on the y-axis the number of objects in the annotated dataset having the corresponding length range are shown. So

[12] `https://fasttext.cc/`

[13] `https://github.com/matteoLorenzini/description_quality`

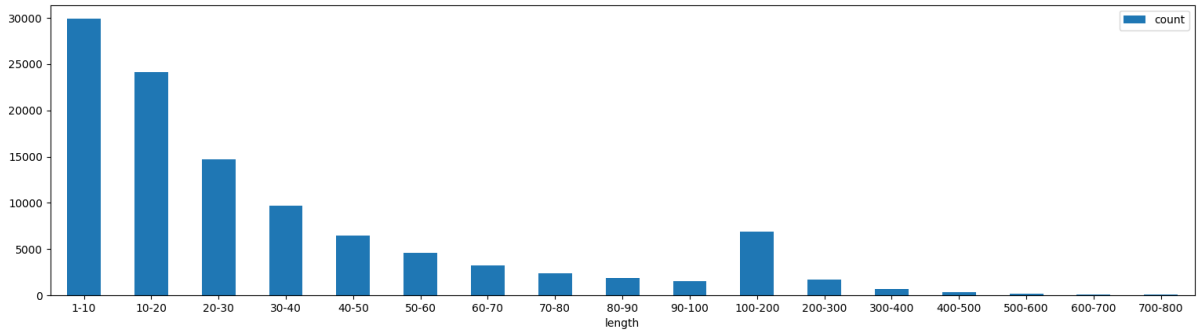


Figure 4.2: Number of records in the annotated dataset (y-axis) per description length bin (x-axis) measured in tokens. Note that a bin size of 10 is used up to length 100, while a size of 100 is used for the remaining bins.

the length is considered as a reasonable baseline to compare with other classifiers as, intuitively, low-quality descriptions tend to be shorter than accurate ones, assessing whether this feature alone could be a good indicator of the description quality.

Therefore, as a baseline, the SVM classifier is trained using as single feature the length of the description in tokens, computed using the TINT tool [2].

4.4 Experimental Setup

4.4.1 Parameter Setting

We run our classification experiments on the three domains in isolation (Visual Art Works, Archaeology and Architecture) and then on the whole dataset. We compare SVM and MLR_{ft}, considering word embeddings of 50 and 300 dimensions in two variants: domain-specific, and general-purpose.

All experiments are run using ten-fold cross-validation. This means that the dataset is first randomly shuffled and then split (preserving the same high-quality / low-quality proportion of the whole dataset) into 10 groups. Each group is used once as test set, while the remaining ones are merged

Dataset	C	G	Kernel
Visual Art Work	3	3	RBF
Archaeology	3	3	RBF
Architecture	32	8	RBF
Overall dataset	1	3	RBF

Table 4.3: SVM C, G and Kernel parameter settings used on each dataset, as result of grid search optimization

Dataset	Learning Rate	Epoch	n-grams	Bucket
Visual Art Work	1.0	100	2	20,000
Archaeology	1.0	100	2	20,000
Architecture	1.0	100	2	20,000
Overall dataset	1.0	100	2	20,000

Table 4.4: MLR_{ft} default parameter settings

into a training set. The evaluation scores obtained on each test set are then averaged to obtain a final, single performance evaluation.

For the SVM, three parameters need to be set, e.g. cost (C), gamma (G) and the Kernel to use. We compute them for each in-domain training set by using the grid search function in scikit-learn. The best parameter combination, which we then adopt in our experiments, is reported in Table 4.3.

With MLR_{ft}, instead, we use the predefined hyper-parameter setup concerning *learning rate*, *epoch*, *n-grams* and *bucket*. The best parameter combination is reported in Table 4.4.

4.4.2 Evaluation Measures

We evaluate the performance of the classifiers using a standard approach for binary tasks: we first compute Precision, Recall and F1 on each of the two classes separately (e.g. high-quality and low-quality) and then average

them. In a 10-fold cross-validation setting, the above evaluation metrics are computed on each fold, and then averaged. More specifically, for each class we count: true positives (TP) – correctly recognized class examples; true negatives (TN) – correctly recognized examples that do not belong to the class; false positives (FP) – examples that are incorrectly assigned to the class; and false negatives (FN) – examples of the class that are not recognized. Then, *Recall*, *Precision* and *F1* are computed as follows:

- Recall $(R) = \frac{TP}{TP+FN}$. It measures how extensively a certain class is covered by the classifier;
- Precision $(P) = \frac{TP}{TP+FP}$. It measures how precise a classifier is, independently from its coverage;
- $F1 = 2 \times \frac{P \times R}{P+R}$.

Overall measures are then obtained by (macro) averaging the scores of both classes. All the metrics are computed using the Python scikit-learn "classification_report" method.[14]

4.5 Evaluation Results

In our evaluation, we address the three research questions introduced in Section 4.

We report in Table 4.5 the classification results obtained with the different algorithms and configurations presented in the previous sections. We include both the within-domain setting, e.g. training and test belong to the same domain (Visual Art Works, Archeology or Architecture), and the global one, considering the three datasets altogether.

[14] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Dataset	system	embeddings	Dim.	Low-quality			High-quality			Overall		
				P	R	F1	P	R	F1	P	R	F1
	baseline			.505	.446	.474	.515	.574	.543	.510	.510	.508
	SVM	Wikipedia	50	.809	.762	.785	.781	.824	.802	.795	.793	.793
	SVM	Wikipedia	300	.850	.826	.838	.835	.858	.846	.843	.842	.842
	SVM	in-domain	50	.809	.762	.785	.780	.824	.802	.794	.793	.793
VAW	SVM	in-domain	300	.850	.826	.838	.835	.858	.846	.843	.842	.842
	MLR_{ft}	Wikipedia	50	.834	.876	.854	.873	.830	.851	.853	.853	**.853**
	MLR_{ft}	Wikipedia	300	.832	.875	.853	.872	.828	.849	.852	.852	.851
	MLR_{ft}	in-domain	50	.834	.860	.847	.859	.834	.846	.847	.847	.847
	MLR_{ft}	in-domain	300	.838	.848	.843	.850	.840	.845	.844	.844	.844
	baseline			.547	.194	.286	.673	.912	.774	.610	.553	.530
	SVM	Wikipedia	50	.814	.659	.728	.830	.918	.872	.822	.788	.800
	SVM	Wikipedia	300	.850	.752	.798	.872	.927	.899	.861	.839	.848
	SVM	in-domain	50	.815	.656	.727	.829	.918	.871	.822	.787	.799
Ar	SVM	in-domain	300	.850	.752	.798	.872	.927	.899	.861	.839	.848
	MLR_{ft}	Wikipedia	50	.861	.848	.854	.917	.925	.921	.889	.886	**.888**
	MLR_{ft}	Wikipedia	300	.862	.843	.852	.915	.926	.920	.888	.884	.886
	MLR_{ft}	in-domain	50	.860	.844	.852	.915	.925	.920	.888	.884	.886
	MLR_{ft}	in-domain	300	.861	.845	.853	.916	.925	.920	.888	.885	.886
	baseline			.530	.288	.373	.671	.850	.750	.600	.569	.562
	SVM	Wikipedia	50	.796	.786	.791	.875	.882	.879	.836	.834	.835
	SVM	Wikipedia	300	.816	.799	.807	.883	.895	.889	.850	.847	.848
	SVM	in-domain	50	.799	.791	.795	.878	.883	.880	.838	.837	.838
A	SVM	in-domain	300	.816	.799	.807	.883	.895	.889	.850	.847	.848
	MLR_{ft}	Wikipedia	50	.845	.822	.833	.890	.905	.897	.868	.864	.865
	MLR_{ft}	Wikipedia	300	.843	.821	.831	.889	.903	.896	.866	.862	.864
	MLR_{ft}	in-domain	50	.843	.812	.828	.884	.905	.895	.864	.859	.861
	MLR_{ft}	in-domain	300	.844	.825	.834	.891	.904	.897	.868	.864	**.866**
	baseline			.493	.255	.336	.577	.795	.669	.535	.525	.502
	SVM	Wikipedia	50	.755	.609	.674	.734	.845	.786	.744	.727	.730
	SVM	Wikipedia	300	.794	.693	.740	.782	.860	.819	.788	.776	.780
	SVM	in-domain	50	.757	.609	.675	.735	.847	.787	.746	.728	.731
All	SVM	in-domain	300	.794	.693	.740	.782	.860	.819	.788	.776	.780
	MLR_{ft}	Wikipedia	50	.769	.738	.753	.801	.826	.813	.785	.782	**.783**
	MLR_{ft}	Wikipedia	300	.767	.740	.753	.801	.824	.812	.784	.782	**.783**
	MLR_{ft}	in-domain	50	.769	.734	.751	.798	.827	.812	.784	.781	.782
	MLR_{ft}	in-domain	300	.771	.732	.751	.798	.829	.813	.784	.781	.782

Table 4.5: Classification results on Visual Art Works (VAW), Archaeology (Ar) and Architecture (A) records, and on the whole dataset. Results are reported as Precision (P), Recall (R) and F1

Overall, MLR_{ft} substantially outperforms SVM in every within-domain setting and configuration, with the former always achieving better F1 score over the latter (with improvements from 0.002 to 0.088 on the overall F1 score). Its performance is consistent for all single domains (best F1 scores ranging from .853 to .888), showing that it is robust despite the different topics mentioned in the descriptions. Also with SVM we observe a compa-

rable performance in the three domains. While for SVM, however, feature vectors with 300 dimensions yield substantially better results, different embedding sizes do not affect much MLR_{ft} output. This means that, even limiting the computation to 50 features dimensions, and hence reducing training time, it is possible to reach good classification performances. The choice of different pre-trained embeddings does not seem to affect much the classification performance, with F1 scores that are substantially similar (with minor, negligible differences) when using in-domain or Wikipedia word embeddings.

When training and testing are performed on the whole dataset, combining descriptions from different domains, the overall scores are lower than on the single domains, suggesting that description quality is something inherent to the different cultural heritage domains, an aspect we investigate more in details with RQ2 in Section 4.5.1.

The baseline results, e.g. a classifier taking into account only description length, are different in the three domains (from .508 to .562 of F1 score). For Architecture it achieves .562 F1, meaning that in most cases longer descriptions tend to correspond to high-quality ones. This is not the case for the Visual Art Work domain, instead, where description length does not correlate with high or low-quality. A possible explanation for this different behaviour may be the fact that in the domain of Architecture, or even Archaeology, descriptions of the cultural artifacts tend to be more standardised, with the same kind of structure and information, therefore description length can be a good indicator of quality. This could explain also why classification performance on the Architecture and the Archaeology datasets is higher than on the Visual Art Work data, even if the latter contains more training instances. We also observe that for the Visual Art Work domain low-quality and high-quality instances can be classified with substantially equal performance, while for the other domains high-quality

descriptions are recognized more accurately. This difference has two possible explanations: first, the two classes are more balanced in the VAW dataset, with roughly the same amount of instances per class. Secondly, classification is equally challenging on the two classes because descriptions are less standardized than in the Ar and A domains.

4.5.1 Specific Domain Analysis

In Table 4.6 we report a second evaluation aimed at assessing what is the impact of the different domains on classification performance. Indeed, for the first set of experiments only descriptions from the same domain were used for training and testing (with the exception of the "All" configuration of Table 4.5). In this second set of experiments, we aim at assessing to what extent quality can be associated with specific domains, and what performance can be achieved by training and testing using data from different domains. In particular, we evaluate the performance of one of the best scoring classifiers of Table 4.5 (namely, MLR_{ft} with Wikipedia embeddings of 50 dimensions) using training data from one or more domains, and testing on one or more (possibly) different domains (e.g., not among the ones used for training). The detail of the various considered combinations is reported in Table 4.6. All experiments are conducted preventing data overlap between train and test datasets.

The results, which should be interpreted according to the dimensions of the domain-specific datasets considered, show that using out-of-domain data greatly affects classification performance. The F1 scores are in general substantially lower than the values reported in Table 4.5, ranging from .371 to .831. The highest value is achieved training on VAW and testing on data from all the domains, an outcome partly justified by the substantially larger size of the VAW dataset with respect to the others. The worst classification performance is achieved using data from the Architecture dataset (A) for

training, both when used in isolation and when added to data from other domains: when training on Ar+A and VAW+A, the scores are lower than when training on Ar and VAW alone, respectively.

Overall, the results show that description quality is something inherent to the different cultural heritage domains, and does not hold in general, because it must be contextualized according to each domain specification. This, as already pointed out in the Chapter 1 and in Section 4.1, is one of the aspects not covered by the automatic evaluation approaches previously proposed in the literature. In general, it is still possible to achieve reasonably good results when a good amount of test data comes from the same domain used for training, as shown by the last two rows of Table 4.6.

Test	Train	P	R	F1
VAW	Ar	.653	.645	.640
VAW	A	.488	.498	.371
Ar	VAW	.644	.654	.617
Ar	A	.447	.488	.414
A	VAW	.551	.552	.550
A	Ar	.560	.562	.556
VAW	Ar+A	.610	.609	.609
Ar	VAW+A	.624	.635	.613
A	VAW+Ar	.573	.576	.572
VAW+Ar	A	.464	.494	.383
VAW+A	Ar	.637	.633	.627
A+Ar	VAW	.610	.617	.596
VAW+Ar+A	A	.661	.556	.495
VAW+Ar+A	Ar	.738	.741	.735
VAW+Ar+A	VAW	.833	.838	.831

Table 4.6: Cross-domain evaluation: Classification results obtained using training data from one or more domains, and testing on one or more (possibly) different domains (e.g., not among the ones used for training).

4.5.2 Comparison of Different Sizes of Training Data

Since manual annotation is, in most cases, a time-consuming task (see Section 4.3.1), the goal of RQ3 is to check how many annotated resources are needed to create a good quality dataset to assess description quality. We address this question by analysing the learning curve of MLR_{ft}, that shows how much the performance improves as the number of training samples increases (from 0.5% to 100%), and therefore estimates when the model has learned as much as it can about the data.

To run this experiment, we proceed as follows. In order to be able to compare the different sizes of training data on the same test set, we manually split the whole dataset according to the classical 80-20 Pareto principle, keeping 20% of the whole dataset (roughly 20K samples out of 100K) for testing.[15] Data are split by preserving their balance both in terms of high/low-quality descriptions as well as source domain. We then train the MLR_{ft} classifier (Wikipedia, 50 dimensions) with increasing sizes of training instances, from 0.5% (~400 descriptions) to 100% (~80K descriptions), and compute the evaluation scores. Figure 4.3 plots the F1 scores obtained (y-axis) by varying the proportion of training data used (x-axis).

The F1 score consistently grows while adding more data to the training set. The higher score is obtained using all the available training material (F1=.845). The curve substantially flattens out at about 35% of the training material (~28K descriptions), and the F1 score is ~.800 already with 10% of the training material (~8K description). This means that, even if the full training set is ten times larger, the classifier does not improve with

[15]Note that this makes the results for this experiment not directly comparable with the values reported in Table 4.5, which instead are obtained following the cross-validation evaluation protocol. Indeed, the results plotted in Figure 4.3 can be considered as a single split (but 80-20 instead of 90-10) of the 10 ones averaged in Table 4.5, and based on the actual split the score obtained may be higher or lower than the ones reported in Table 4.5.

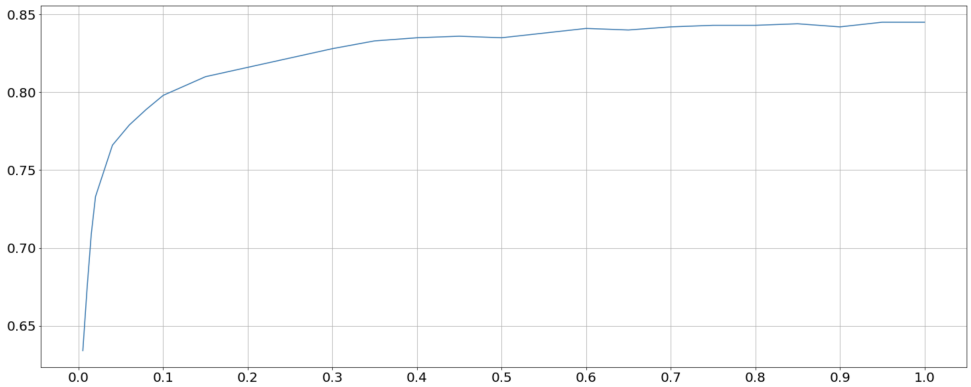


Figure 4.3: Learning curve with F1 on the y-axis, obtained by progressively increasing the number of training instances (x-axis).

the same proportion (less than 5%). Therefore, in a scenario in which no training data are available, we would suggest a domain expert to manually annotate around 8-10,000 in-domain descriptions to still yield good classification results. At the annotation rate described in Section 4.3.1, developing a manually validated dataset of this size would required approximately 53-67 hours of human effort.

4.5.3 Discussion

Even if the classifier presented in this work may still be improved, the obtained results are very promising, suggesting that an automated analysis of description quality is feasible and it would be possible to provide a first check of the descriptions in cultural heritage records before expert validation. The obtained results show also that more training data are not necessarily the best solution, especially if they are not from the same domain. On the contrary, around 8-10,000 annotated instances, possibly from the same domain of interest, are enough to achieve reasonably good classification performances. Another insight from the experiments is that FastText multinomial logistic regression classifier (MLR_{ft}) outperforms SVM for this task. Moreover, the domain of the pre-trained embeddings used for build-

ing the numerical vectors of the descriptions fed to the classifiers seems to have little impact on the performances, as both general-domain embeddings (trained on Wikipedia) and in-domain ones achieve comparable scores.

In general, the advantage of this approach is that no feature engineering and no language-specific processing of the descriptions are needed, apart from stopword and punctuation removal. This means that this approach is easily applicable to descriptions in any language, provided that training data are manually annotated by a domain expert.

Record ID	Description	Gold	Predicted	Error
work_48470	Oinochoe *a corpo baccellato*. Applique with female protome matrix at the handle attachment.	HQ	LQ	A
124472	Black-figure painted attican Kylix , Siana type.	HQ	LQ	A
10530	Corintian Amphoriskos with zoomorphic decoration.	HQ	LQ	A
iccd3415758	The Saint, kneeled down looks up. on the bottom, to the left, there is a winged putto.	LQ	HQ	B
iccd3145858	the base lies on a parallelepiped-shaped base; [...] high volute handle.	LQ	HQ	B
iccd3165805	Brocade satin; checkered pattern. The compositional unit derives by [...] with flowers and leafs.	LQ	HQ	B
iccd3908065	Rich Oriental with mustache and half-closed mouth, head slightly oriented to [...] Figure: man	LQ	HQ	B
iccd4413810	The cycle includes three illustrated tondos, [...] .	LQ	HQ	C
iccd3913506	Wooden little angels sitting on a cloud, wrapped in a blue mantle, with wings [...]	HQ	LQ	C

Table 4.7: Sample of high-quality (HQ) and low-quality (LQ) annotated records wrongly classified in our classification experiments.

As regards the mistakes done by the classifiers, the wrongly classified instances produced by one of them (MLR_{ft}, Wikipedia, 50 dimension) have been manually inspected and they almost exclusively (95% of them) fall in one of the following three categories:

- Error type A: *Descriptions containing Latin and/or Greek terms:* mis-

Record ID	Description	Gold	Predicted
work_15736	The big polyptych commissioned by the Guidalotti family for their chapel [..]	LQ	LQ
work_63812	Thanks to Shearman it was verified that the painting was located in the building in via Larga where it remained [...]	LQ	LQ
iccd3906852	Crib statuette depicting an angel in a flying posture, dressed [...]	HQ	HQ
iccd2307693	[...] The man depicted has a mustache and beard and wears a wide-brimmed hat [...]	HQ	HQ

Table 4.8: Sample of high-quality (HQ) and low-quality (LQ) annotated records correctly classified by the approach.

classifications in these cases (e.g., work_48470 and work_48471 in Table 4.7) may be due to the fact that these words are not frequent and therefore are not represented in a meaningful way in the embedding space;

- Error type B: *Descriptions only partially compliant with the cataloguing guidelines provided by the ICCD:* these descriptions are typically annotated as low-quality in our gold standard, even if the description does not contain factual errors per se on the item. In our experiments they tend to be automatically annotated as being of high-quality (see for example the record iccd3908065 in Table 4.7);[16]

- Error type C: *Descriptions where the subject is implicit:* in these cases the classifier is not able to properly identify the domain of the item, as there may be no reference about the typology of the cultural object (see record iccd3913506 in Table 4.7).

Additional examples of incorrect classifications are reported in Table 4.7. As regards correctly classified instances, the Table 4.8 shows some example.

[16]The iccd3908065 description is of low-quality in the gold standard according to the ICCD guidelines as it does not provide a description of the object: there is no mention in the text that the item refers to a statue, nor to its material characteristics

Among them, the description of the Italian masterpiece "La Primavera" by Sandro Botticelli (record work_63812 in Table) consists of an articulated explanation on how the painting joined the Uffizi Gallery's collection rather than describing the painting itself, hence it has been correctly classified as having low-quality by the system.

4.6 Chapter Summary

This chapter analyzed the problem of metadata Accuracy evaluation of digital resources, assessed as a binary classification of textual descriptions. In Section 4.1 we presented the problem of Accuracy evaluation describing why it is a key factor to have high-quality metadata. In Section 4.2 we provided a critical analysis of the different approaches used to evaluate the Accuracy dimension, introducing our point of view on why a machine learning approach could be better than what has been developed so far. Section 4.3 is devoted to the assessment of the problem presenting three specific sub-research questions. Moreover we described in detail how we developed the dataset used to train and test the machine learning model and how we compared the SVM and MLR_{ft} algorithms. This comparison, given the novelty of the proposed approach, can be considered as the starting point for future improvements. In Section 4.5 we presented the results of the comparison providing a detailed comment about the different performances of the SVM and MLR_{ft} algorithms. In the final section of this Chapter, we then presented the most common errors obtained by using the machine learning model.

Chapter 5

The Coherence Dimension

5.1 Introduction

The dimension related to Logical Consistency and Coherence relies on two types of evaluation. At the dataset level, Logical Consistency checks if the resources are described using the same set of the metadata element. At the level of the single record, the Coherence, evaluates if a resource is described according to the guidelines provided by the chosen metadata standard schema. Consider for example the Stamnos in Figure 5.1, record id. 5588 from the dataset Regione Umbria in Cultura Italia.

The Logical Consistency and Coherence dimension evaluates if the values provided by the metadata elements used to describe the resource are logically connected and mutually consistent with the goal to improve the information retrieval process. This allows end-users to retrieve the record id. 5588 for example by using the typology, the subject as well as the material and technique element as terms for a faceted or a full text research as for "Archaeological items with black paint figure".

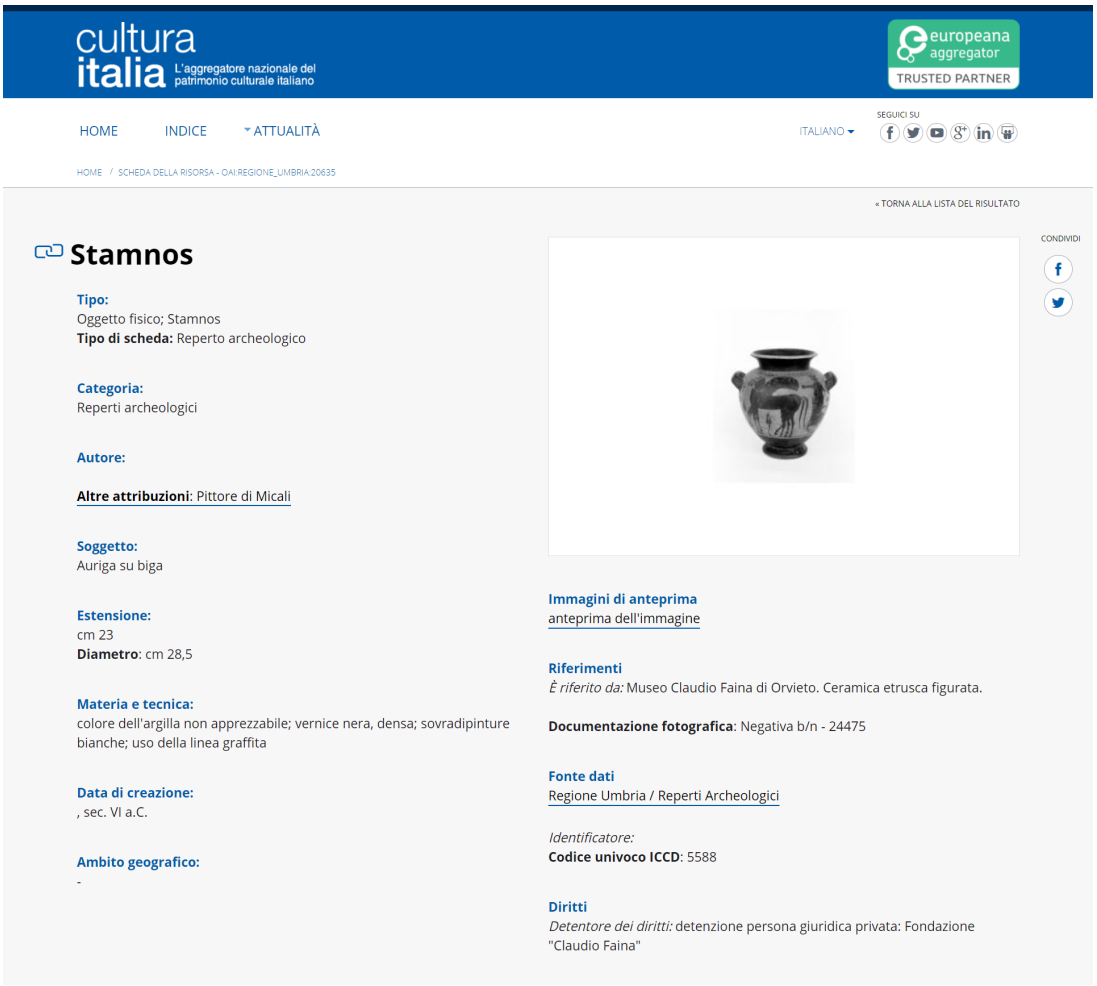


Figure 5.1: Stamnos, source record id 5588 from Regione Umbria dataset.

5.2 State of the Art

Within the seven frameworks described in the section 2.2 this dimension is adopted in the metadata quality evaluation process by Bruce and Hillmann [7], in the metadata quality framework developed by Stvilia et al. [61] and in metadata quality framework by Ochoa and Duvall [53].

For Bruce and Hillmann (See Subsection 2.2.2), Logical Consistency and Coherence dimension is defined as the degree to which the resource description matches with metadata standard schema and definition. In Stvilia et al. (See Subsection 2.2.5), Logical Consistency and Coherence is named

Semantic Consistency dimension and provides two types of metadata quality check: about the correspondence with metadata schema and about the coherence of the information provided by all the elements used to describe the resource. The proposed methodology is a theoretical assessment and no implementation of this dimension is tested on a real repository. In the metadata quality framework by Ochoa and Duvall (See Subsection 2.2.4), Logical Consistency and Coherence is computed as two separate dimensions: the Logical Consistency aims to evaluate the degree to which the resource matches with metadata standard definition while the Coherence aims to evaluate the degree to which all metadata fields describe the same object similarly.

Ochoa and Duvall define three rules to identify low Consistency in metadata records:

- Instances include fields not defined in the standard or do not include fields that the community sets as mandatory;
- Categorical fields, that should only contain values from a fixed list, are filled with a non sanctioned value;
- The combination of values in categorical fields is not recommended by the standard definition.

They propose the following Consistency dimension to capture these type of rules: 0 if the instance complies with the rule (all possible errors are made) otherwise 1 (there are no consistency problems). The Consistency dimension will be equal to 1 minus the average of fraction of problems found, for each type of problem.

The Coherence is measured as the degree to which all the fields describe the same object in a similar way. It is computed evaluating the semantic distance between textual elements according to the words they have in common.

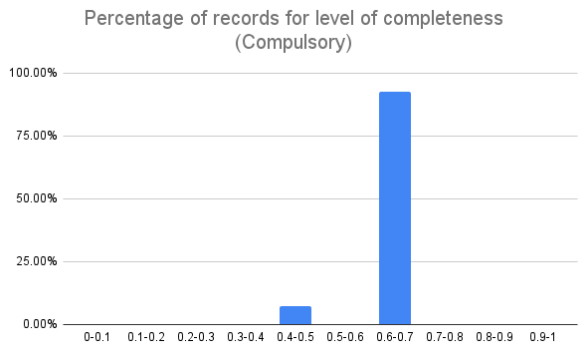
In our view, taking the definition from Bruce and Hillmann as the main reference "*The logical consistency and coherence dimension is defined as the degree to which the resource description matches with metadata standard schema and definition...*", This dimension must be addressed separating Logical Consistency from Coherence during the evaluation, as Ochoa and Duvall did, thus introducing two sub-dimensions: the Consistency to evaluate the structure of metadata information, and the Coherence to evaluate the formal aspect of metadata information. Contextually to the methodology presented in this Ph.D. thesis we argue that Consistency can be partially evaluated while computing metadata Completeness presented in the Chapter 3. As previously mentioned, Ochoa and Duvall identify three rules that identify elements with low Logical Consistency. One of them explicitly refers to the compulsory metadata elements used to describe the resource as follows:

- Instances include fields not defined in the standard or do not include fields that the community sets as mandatory.

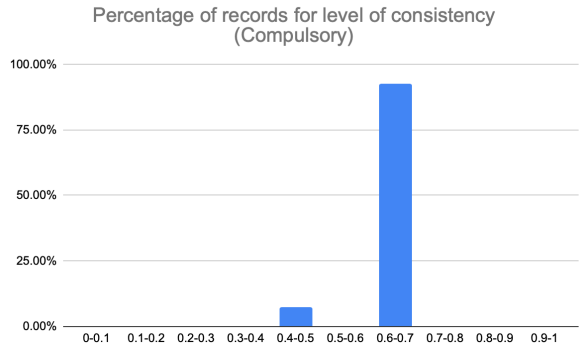
Therefore the Completeness score for the compulsory group could be also interpreted as the Consistency score with respect to the metadata profile. We consider as an example Figure 5.2

The plot in Figure 5.2a shows the results for the Completeness on the MuseID-Italia datasets. Most of the records (92%) obtain a Completeness score for the compulsory group in the range between 0.6 and 0.7. This also means that the (92%) of the records are also consistent with respect to the 8 compulsory metadata elements of the PICO application profile (5.2b).

While investigating the frequency of usage of the PICO metadata elements in the records of the collections, the metadata curator should be able also to check if the metadata resources are described using metadata elements not defined in the standard schema.



(a) Completeness score for the compulsory elements



(b) Consistency score for the compulsory elements

Figure 5.2: Completeness and Consistency plots for MuseID-Ialia dataset

The plot in Figure 5.3 represents the overall percentage of used metadata elements per dataset. This representation could be also helpful to metadata curators for this specific metadata quality check.

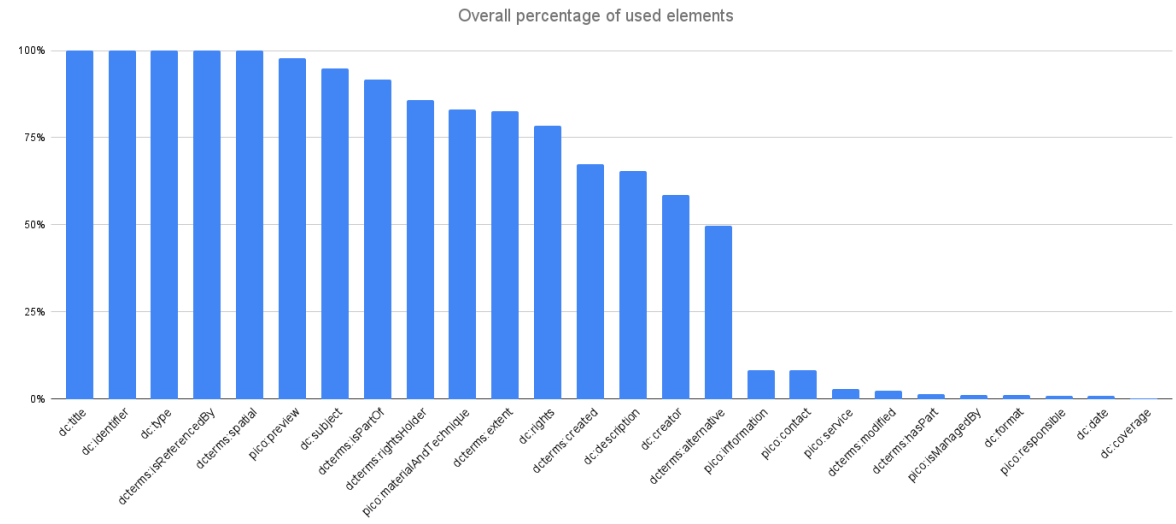


Figure 5.3: Percentage of records in the MuseID-Italia dataset having a given metadata element

The quality check of metadata Coherence requires a more complex analysis. In the definition from Bruce and Hillmann there is no refer-

ence whether Coherence should be evaluated against the metadata schema specifications or comparing the other elements used to describe the digital object. In Ochoa, for example, the Coherence is evaluated among metadata information. In this work we measure the metadata Coherence with respect to the topic of a cultural object.

5.3 Coherence Assessment

The subject is one of the most important metadata elements to describe a cultural resource. Thanks to this information users are able to have useful information about the main topic of a painting as for example "Deposition of Christ" and to obtain a specific resource retrieval filtering the results for a given subject or topic. Moreover the *dc:subject* metadata element is considered as a compulsory field from the majority of digital archives. So, it is crucial to check the quality of this information.

The methodology presented here, instead of evaluating the quality of metadata when they are already ingested into the digital archive as for metadata Completeness and Accuracy evaluation, points to define a methodology that prevents the creation of low-quality metadata information. More specifically, we assess metadata Coherence of the subject element predicting the 3 most likely subjects of the resource analyzing the iconography illustrated in the textual description. The main contribution of this Chapter is to investigate the feasibility to use natural language processing techniques and machine learning solutions to create a multi-label classification model that is able to suggest the top-K appropriate subjects in the Cultural Heritage domain.

According to the ICCD guidelines, the subject element should be defined by using the terminology from a controlled vocabulary[1] that can be

[1]SGT element.`http://www.iccd.beniculturali.it/getFile.php?id=7508`

integrated[2] with the Iconclass iconographic classification dataset[3] So, as the first step in this direction, we use the definition from the Iconclass iconographic classification dataset to train a classification model [15]. In this dataset the descriptions are pre-labelled according to the 10 codes (notions) that refer to a specific iconographical group in Iconclass:

- (0) Abstract, Non-representational Art;
- (1) Religion and Magic;
- (2) Nature;
- (3) Human Being, Man in General;
- (4) Society, Civilization, Culture;
- (5) Abstract Ideas and Concepts;
- (6) History;
- (7) Bible;
- (8) Literature;
- (9) Classical Mythology and Ancient History.

An example of annotated resources from Iconclass is reported in Table 5.1.

Iconclass dataset	
Label	Description
Religion and Magic	Father and Holy Ghost holding the crucified Christ (Trinity)
Bible	The battle between David and Absalom in the forest of Ephraim.

Table 5.1: Example of annotated resources using Iconclass definition

[2] http://www.iccd.beniculturali.it/getFile.php?id=186

[3] http://www.iconclass.org/help/outline

The Iconclass dataset is described in detail in the section 5.3.1. As a second step, we select 501 high-quality resources from the gold standard dataset used to for the metadata Accuracy assessment to use as a test dataset. This methodology to check metadata coherence is exploited in this work addressing the following Research Question 3 **(RQ3)**:

- Is it possible to predict the correct subject of a certain cultural resource?

5.3.1 Dataset Description

Iconclass [15] is a well-known iconographic classification system which is used to describe and retrieve content in artworks. Iconclass offers a hierarchy of unique codes, associated with keywords and definitions, to encode the presence of objects, people, events and ideas depicted in visual artworks, such as paintings or drawings.

Resource Collection

To work with a representative collection of subject descriptions in the domain of cultural heritage, we used the full Italian Iconclass dataset available on GitHub[4]. The entire dataset is composed of 39,772 subjects organized according to the ten Iconclass categories.

Dataset Structure

Iconclass includes 28,000 hierarchically ordered definitions (notations) and 14,000 keywords in multiple languages [31]. As matching targets, in this thesis we adopt the Italian definitions of the Iconclass codes. Iconclass is

[4] https://github.com/iconclass/data/tree/main/txt/it

divided into ten main categories that are represented with a digit from 0 to 9 in Table 5.2

Iconclass dataset		
Label(EN)	Label(IT)	Topics
(0) Abstract, Non-representational Art	(0) Arte Astratta	Abstract Art
(1) Religion and Magic	(1) Religione e Magia	General Topics
(2) Nature	(2) Natura	General Topics
(3) Human Being, Man in General.	(3) Essere Umano, Uomo in Generale	General Topics
(4) Society, Civilization, Culture.	(4) Societá, Civilizzazione e Cultura	General Topics
(5) Abstract Ideas and Concepts.	(5) Idee e Concetti Astratti	General Topics
(6) History	(6) Storia	History
(7) Bible	(7) Bibbia, Storie dal Vecchio e Nuovo Testamento	Bible
(8) Literature	(8) Letteratura	Literature
(9) Classical Mythology and Ancient History	(9) Mitologia Classica e Storia Antica	Classical Mythology and Ancient History

Table 5.2: Main Iconclass categories with relative code

Each main category can be divided into 9 other subdivisions, thus becoming more specific. In Iconclass, this is done by simply adding a second digit to the notation. For example, (1) Religion and Magic category is divided into:

- 10. (symbolic) representations creation, cosmos, cosmogony, universe, and life (in the broadest sense)
- 11. Christian religion

- 12. Non-Christian religions (including institutions, customs and antiquities)
- 13. Magic, Supernaturalism, Occultism
- 14. Astrology

To specify the subject we then should add a letter to the notation, for example:

- 13A general phenomena magic and supernaturalism (spirits, ghosts etc.)
- 13B witchcraft, sorcery
- etc.

5.3.2 Dataset Annotation

To annotate the Iconclass dataset, we use the classification code of the Iconclass notations and the ten main categories to which they refer [25]. For example the code (1) identifies uniquely the category "Religion and Magic" so we annotate all the subjects starting with the notation (1) as "Religion and Magic". Table 5.3 shows an example of an annotated subject definition.

Notion	Subject Description(IT)	Subject Description(EN)	Label (EN)	Label (IT)
11A3	Ira di Dio	God's wrath	Religion and Magic	Religione e Magia

Table 5.3: Example of annotated definition from Iconclass

Table 5.4 summarizes the size of the annotated dataset as well as the size of each category used for multi-label prediction. Among the ten categories that characterize the Iconclass dataset, nine are available also in the Italian

translation since the category nr.0 “Abstract, Non-representational Art” is merged with the category nr.5 “Abstract Ideas and Concepts”.

Dataset Structure		
Label(EN)	Label (IT)	Number of subject per label
Religion and Magic	Religione e Magia	12.246
Classical Mythology and Ancient History	Mitologia Classica e Storia Antica	9.244
Society, Civilization, Culture	Societá, Civilizzazione e Cultura	6.450
Bible	Bibbia, Storie dal Vecchio e dal Nuovo Testamento	5.961
Nature	Natura	2.388
Human Being, Man in General	Essere Umano e Uomo in Generale	2.272
Abstract, Non-representational Art.	Idee e Concetti Astratti	902
Literature	Letteratura	247
History	Storia	62
Overall dataset		39.772

Table 5.4: Structure of the Iconclass dataset used to train the model

5.3.3 Classification Framework

The goal of the classification framework discussed in this Chapter is to automatically identify and suggest the 3 most likely subjects of a cultural heritage resource from its textual description [4].

As a methodological approach, this problem is modeled as a multi-label classification task, employing the annotated dataset described in Section 5.3.1 to train a supervised system that is capable of assigning an unseen object to one of 9 subjects.

We proceed according to the same method used in Section 4.3 for the classification of metadata accuracy: we first remove the stopwords and punctuation from the dataset, then we convert the subject descriptions into numerical vectors using FastText word embeddings.

From the dataset we used for the assessment of metadata accuracy classification, we select 501 high-quality descriptions for testing the model's ability to predict a certain subject. We manually annotate the dataset according to the 9 Iconclass categories, then we validate the annotation again with metadata curators from Cultura Italia. We finally compare the prediction performance of the two algorithms SVM and MLR_{ft}. Table 5.5 summarizes the statistic of the annotated test dataset.

Dataset Structure		
Label(EN)	Label (IT)	Number of subject per label
Society,-Civilization,Culture	Societá, Civilizzazione e Cultura	176
Religion and Magic	Religione e Magia	171
Human Being, Man in General	Essere Umano e Uomo in Generale	83
Classical Mythology and Ancient History	Mitologia Classica e Storia Antica	27
Nature	Natura	15
Abstract, Non-representational Art.	Idee e Concetti Astratti	13
History	Storia	10
Bible	Bibbia, Storie dal Vecchio e dal Nuovo Testamento	6
Overall dataset		501

Table 5.5: Strucutre of the test dataset

5.3.4 Baseline

For the definition of the baseline we proceeded as follows:

- From the description quality dataset, described also in Section 4.3.1 of this thesis, we select 501 high-quality descriptions in the visual artwork domain to use as a test dataset;
- From the Iconclass vocabulary, we select the three categories with the largest number of iconographic descriptions: Religion and Magic (12.246), Classical Mythology and Ancient History (9.244), Society, Civilization, Culture (6.450);
- We annotate each description in the test set according to the three categories, assuming that the baseline predict the correct description in descending order. Table 5.6 shows an example.

Baseline	
High-Quality Description	Labels
Dipinto raffigurante il ritratto di Bartolomeo Castiglioni reso di profilo in veste militare.	Religion and Magic; Classical Mythology and Ancient History; Society, Civilization, Culture.
Dipinto raffigurante Salomé che porge la testa del Battista a Erode.	Religion and Magic; Classical Mythology and Ancient History; Society, Civilization, Culture.

Table 5.6: Example of the baseline annotation

- Then we compute the baseline as follows:
  - Baseline@1(B@1): We assign a value of 1 if the annotated subject matches the first label used in the baseline (in our case Religion and Magic). Otherwise, we assign a value of 0;
  - Baseline@3(B@3): We assign a value of 1 if the annotated subject matches one of the three label used in the baseline (in our case Religion and Magic, Classical Mythology and Ancient History, Society, Civilization, Culture). Otherwise, we assign a value of 0;

– In order to calculate the baseline precision score, we divide the obtained sum by the number of resources annotated.

The baseline results are reported in Table 5.7

Baseline	
B@1	B@3
.341	.746

Table 5.7: Coherence baseline results

5.4 Experimental Setup

5.4.1 Parameter Setting

The process of finding the best parameter setup is done as follows. first we train the model on the Iconclass dataset, then we test it on the 501 annotated high-quality descriptions dataset. We train the model using the best parameters obtained via the grid search function of scikit-learn. The parameter combination, which we then adopt in our experiments, is reported in Table 5.8. With MLR_{ft}, instead, we use the predefined hyper-parameter setup concerning *learning rate*, *epoch*, *n-grams* and *bucket* as reported in Table 5.9.

Dataset	C	G	Kernel
Iconclass	51	1	RBF

Table 5.8: SVM C, G and Kernel parameter settings used on Iconclass, as result of grid search optimization.

Dataset	Learning Rate	Epoch	n-grams	Bucket
Iconclass	1.0	100	2	20,000

Table 5.9: MLR_{ft} default parameter settings used on Iconclass.

5.4.2 Evaluation Results

In our evaluation, we address Research Question 3 **(RQ3)**:

- Is it possible to predict the correct information to be used for a given element?

System	P@1	P@3
MLR_{ft}	.381	.668
MLR_{ft} predicted Subjects	191/501	335/501
SVM	.401	.646
SVM predicted Subjects	201/501	324/501

Table 5.10: Prediction results. Results are reported in percentage on a scale from 0 to 1

From the 501 resources of the test set, we calculate the precision of the predicted label compared with the manual annotation as follows:

- Precision@1(P@1): We assign a value 1 to our annotated subject if the first label predicted by the multi-label classification system matches our annotation (in our case Magic and Religion). Otherwise, we assign a value 0;
- Precision@3(P@3): We assign a value 1 to our annotated subject if one of three labels predicted by the multi-label classification system matches our annotation (in our case Religion and Magic, Classical Mythology and Ancient History, Society, Civilization, Culture). Otherwise, we assign a value 0;
- We divide the obtained sum by the number of the annotated resources to obtain the Precision score.

Table 5.10 summarizes the prediction results we obtain with the different algorithms and configurations presented in Section 5.4.1. Both SVM

and MLR_{ft} algorithms perform similarly. The Precision@1 (P@1) score is higher than the Baseline@1 (B@1) in both SVM and MLR_{ft} while the Precision@3 (P@3) has lower prediction performance with respect to the Baseline@3 (B@3). SVM outperforms MLR_{ft} in the Precision@1 (P@1) with 201 subjects correctly predicted over 501 resources while MLR_{ft} obtain better performances in the Precision@3 (P@3) with 335 subjects correctly predicted over 501 resources. As expected, P@1 has a low precision rate. This is due to the fact that the model is able to recommend the correct concept as the 1st prediction only if the textual resource describes the particular characteristics associated with that concept. Some examples are reported in Table 5.11.

Record ID	Description	Gold	1st prediction	2nd prediction	3rd prediction
Work_752	The painting represents the Dove of the Holy Spirit, inserted in a round stucco frame .Symbols: Dove of the Holy Spirit	Religion and Magic	Religion and Magic	Bible	Nature
Work_1329	The bust represents a Roman emperor, in front, with a very draped tunic over a suit of armor. Portraits. Clothing	Classical Mythology and Ancient History	Classical Mythology and Ancient History	Human Being, Man in General	History

Table 5.11: Sample of correct prediction outcome

In the case of the record id. "Work_752" the painting is explicitly referring to the Dove of the Holy Spirit while the description of the record id. "Work_1329" specifies that the bust represents a Roman Emperor. In the absence of such particular attributes in the textual descriptions, the classifier frequently assigns the correct subject from among the three annotated labels. Other examples are reported in Table 5.12

In this case the description of the cultural objects is generic, without a specific reference to a particular iconography. However, the subject "Human Being, Man in General" is correctly recommended as the 3rd option

Record ID	Description	Gold	1st prediction	2nd prediction	3rd prediction
A548_6	Half-length bust in the round on a rock column pedestal and rectangular base, inserted in an ovoid masonry niche. Represents Pope Pius VII in religious clothing	Human Being, Man in General	Religion and Magic	Society, Civilization and Culture	Human Being, Man in General
Work_120	Portrait of Margherita de 'Medici. Clothing: dress with pearls arranged in the shape of a lily and lace collar strings of pearls jewel-clips	Human Being, Man in General	Society, Civilization and Culture	Human Being, Man in General	History

Table 5.12: Sample of wrong prediction outcome

for the record id A548_6 and as the 2nd option for the record id Work_120, and then the Precision@3 (P@3) score is correctly computed.

A further error of both algorithms that adversely affects the evaluation is that MLR_{ft} and SVM are in some cases unable to distinguish iconography subjects that relate to "Religion and Magic" from those relating to the "Bible" due to their semantic similarity. An example is reported in Table 5.13.

Record ID	Description	Gold	Predicted
Work_2167	Painting depicting the crucifixion of Christ.	Bible	Religion and Magic
49145	Painting depicting Salome offering the head of the Baptist to Herod	Bible	Religion and Magic

Table 5.13: Sample of wrong prediction outcome between the subject "Religion and Magic" and "Bible"

5.4.3 Discussion

The methodology discussed in this Chapter still needs to be improved. As demonstrated by the results presented in the Section 5.4.2, the advantages of using a machine-learning algorithm to predict the correct metadata sub-

ject value are still limited. The baseline (B@3) produces more reliable results than the P@3 (MLR_{ft}.668 vs .746, SVM .646 vs .746), while the performance score of P@1 is better compared to the baseline (B@1). Nevertheless this is not sufficient to guarantee good results by applying our methodology on a large scale. Overall, our experiments shows that suggesting the correct metadata subject(s) can be partly addressed through machine learning.

Two aspects require a deeper analysis and improvement:

- Training dataset: Only 9 main categories were used to annotate the Iconclass dataset. It would be usefull to annotate the dataset with the subcategories described in Section 5.3.1 to provide a more detailed description of iconographic subjects. However, even if we use the subcategories, the total number of annotated resources does not change, leaving a significant imbalance between the annotated terms, as for "Religion and Magic" (12.246) and "Bible" (5.961) or between "Society, Civilization, Culture" (6.450) and " Human Being, Man in General" (2.272). To fill this gap and have a clearer distinction between semantically similar annotations, Iconclass might be integrated with other compatible terms, as suggested by the ICCD regulation[5]. A possible methodology would be to extract all the main entities from the high-quality descriptions of the gold standard dataset and semantically integrate them into the iconclass vocabulary.

- Test dataset: We evaluated the multi-label classification system using 501 high-quality annotated resources.The creation of a larger test dataset could be useful to have a more detailed picture regarding this specific methodology.

We will consider these two aspects for further improvements.

[5] http://www.iccd.beniculturali.it/getFile.php?id=186

5.5 Chapter Summary

This chapter focuses on the analysis of the Logical Consistency and Coherence dimension. We propose a different approach with respect to metadata Completeness and Accuracy evaluation based on the creation of correct information to prevent the creation of low-quality metadata. In Section 5.1 we introduced the problems related to the Logical Consistency and Coherence dimension, describing under which perspectives the evaluation should be assessed. Section 5.2 is devoted to presenting past approaches providing an analysis about the methodology used to evaluate this dimension. In Section 5.3 we presented the methodology we used to assess the problem describing how we defined and annotated the training and test dataset and the baseline we choose. Finally, in Section 5.4, we presented the evaluation of the methodology applied to our use case Cultura Italia.

Chapter 6

Conclusions

The main goal of this work is to develop a comprehensive, generic (e.g., non-domain specific) and machine-processable methodology to evaluate the metadata quality of the cultural resources. To achieve this, the standard guidelines from the Dublin Core Metadata Schema[1] and the standard guidelines from the Istituto Centrale per il Catalogo e la Documentazione (ICCD)[2] have been used as the main reference for the evaluation.

After having introduced the motivation of this work and the main research questions in Chapter 1, in Chapter 2 we have presented the approaches and methodology used in the past. We have critically analyzed two aspects: the first is the usage of standard guidelines to prevent the creation of low-quality metadata by the cataloguers, the second aspect concerns the implementation and usage of metadata quality frameworks to evaluate the quality of metadata information. In Chapter 3 we described our methodology to evaluate the Completeness dimension computed as the percentage of the filled fields with respect to the metadata schema used to describe the resource. The main contribution of our approach is the possibility to have four different levels of analysis with respect to the compulsory, recommended, optional and domain elements allowing metadata

[1] https://www.dublincore.org/specifications/dublin-core/dces/

[2] http://www.iccd.beniculturali.it/it/normative

curator to have a better understanding of the overall metadata Completeness. Moreover we presented also the results obtained by applying the proposed methodology to the Italian digital library Cultura Italia. Chapter 4 is devoted to the description of the methodology used to evaluate metadata Accuracy of the textual description of digital resources. We defined the methodology as a binary classification problem where the textual descriptions are classified as "high-quality" or "low-quality". The methodology we defined for computing metadata Accuracy presents the following main novelties. Firstly, to annotate the dataset we relied on the standard cataloging guidelines by ICCD. Secondly, the annotated dataset was also validated by the technical team of Cultura Italia. Then, we performed three experiments to evaluate the performances of the SVM and MLR_{ft} algorithms on assessing the binary classification problem. Finally, in Chapter 5 we presented the last of the three dimensions we chose for the definition of our metadata quality evaluation methodology: the Coherence. Unlike Completeness and Accuracy, we assessed the validity of our methodology not as a dimension useful to obtain specific information about the quality of the provided information. We concentrated our effort on investigating the feasibility of directly creating high-quality metadata information preventing the definition of erroneous metadata information using a specific multi-label classification system. However, while applying our methodology to a test set of records from Cultura Italia, we noticed that this task is only partially assessable at least with respect to the solution we implemented.

6.1 Answers to the Research Questions

Research Question 1: Can we define a way to compute Completeness that takes into account the context and relevance of metadata elements for a metadata collection?

The evaluation of the metadata Completeness has been performed organizing metadata elements in different groups accounting for their relevance for the considered dataset. Thanks to the computation based on 4 different metadata groups, the system was able to return to the metadata curator as a precise and comprehensive picture of the overall dataset Completeness, and also allowed for comparisons across datasets and different domains.

In respect to RQ1 this approach has demonstrated that the evaluation of the Completeness is a task assessable considering the context and the resource domain. However, despite the promising results we have obtained by applying this methodology to the MuseID-Italia and Regione Marche datasets, the definition of metadata elements for the Domain Specific group, represents an issue that needs a further consideration. The selection of the domain specific element could be compromised by the "Human Factor" as indicated in the Section 1.1 as one of the problems for the creation of low-quality metadata. Depending on the background of the metadata curator the meaning of "Specific Domain" could be interpreted differently. In this case the usage of metadata guidelines and the validation of metadata groups with a domain expert is a fundamental task.

Research Question 2: Can we effectively assess metadata Accuracy using computational methods based on machine learning?

The evaluation of metadata Accuracy of the textual description has been assessed as a binary classification problem aiming to automatically classify textual descriptions in cultural heritage records with the label "high-

quality" or "low-quality". Not only we showed that machine learning approaches yield good results in the task, but we also provided insights into the classifier behaviour when dealing with different domains, as well as into the amount of training data needed for classification, given that manual annotation is a time-consuming activity.

The proposed approach has several advantages: it does not require any in-depth linguistic analysis and feature engineering, since the only features given in input to the classifier are FastText word embeddings. Besides, both SVM and MLR_{ft} are less computationally intensive and energy-consuming than well-known deep learning approaches, and no specific computational infrastructure (e.g. GPU) was needed to launch the experiments. A key finding of this methodology was also the importance of the domain in the classification experiments as well as in the manual creation of training data: without an expert in cultural heritage it would be impossible to create manually annotated data and to judge the performance of the classifiers from a qualitative point of view. Crowd-sourcing approaches to data annotation, which is often adopted to annotate large amounts of linguistic data through platforms such as Amazon Mechanical Turk cannot be used in our scenario, since laypeople would not have the necessary knowledge to judge the compliance of descriptions with the corresponding guidelines. This confirms the importance of multi-disciplinary work in the digital humanities, where technological skills and humanities knowledge are both necessary to achieve the project goals.

In the future, we plan to further extend this work in different research directions. As a short-term goal, we would like to compare the performance of our classifiers with other classification algorithms, including deep-learning ones. Another configuration we would like to evaluate is the use of transformer-based contextual embeddings like BERT [17] instead of word embeddings, since they provide a representation of entire chunks of text and

not just at the word level. This may help in better discriminating different textual contexts, e.g. dealing with different domains. An additional set of experiments could concern extending the evaluation to collections from different countries, therefore tackling descriptions in multiple languages, taking advantage of the fact that our approach does not require language-specific text processing. Moreover, another future research direction we plan to investigate is the benefit of leveraging knowledge beyond the textual content (e.g., knowledge bases, taxonomies, source authorities) to improve the assessment of description quality, especially in combination with the machine learning approaches we considered.

We would also like to address the main limitation of our approach, e.g. the fact that we consider description quality as something that can be observed and measured by only considering the textual component of a cultural heritage record and its compliance with ICCD guidelines. An actual assessment, with broader practical implications, should include also the item image, and check the existing (or missing) correspondences between textual and visual content. This further level of analysis would require multimodal approaches, which we would like to explore as the next step in our investigation, taking advantage of existing infrastructures that support the curation of metadata, record content and images through the same interface [21].

Research Question 3: Is it possible to predict the most coherent subject of a given cultural heritage resource from its description?

In the methodology we proposed, Coherence is considered as the degree to which metadata information is coherent with a digital cultural heritage resource. We defined a methodology which helps metadata creators in generating high-quality values at the source. In doing so we used the Iconclass vocabulary as a dataset to train a machine learning model, with the goal to

predict the top three most likely subjects from the iconography described in the textual description of a digital resource. To evaluate the results of our prediction model, we used a test set composed of 501 high-quality descriptions manually annotated according to the Iconclass categories. We evaluated the performance of the two algorithms SVM and MLR_{ft} by measuring P@1 (i.e. if the manually annotated subject was the 1^{st} suggestion by the system), and P@3 (i.e. if the manually annotated subject was one of the three suggestions by the system).

With respect to RQ3, the methodology we presented, has demonstrated that was partially possible to predict the correct metadata subject, in this case from the textual description of a resource. Overall the methodology still needs further improvements. More specifically, we identified the following main issues that also could be used as a base for a further improvement of the proposed methodology:

- The usage of the only textual fields to predict metadata information may be non enough. To have better and more reliable predictions also features extracted from the resource images should be integrated into the training model. An example is represented by the resource A548_6 which we also reported in Table 5.12. The resource description is: *Half-length bust in the round on a rock column pedestal and rectangular base, inserted in an ovoid masonry niche. Represents Pope Pius VII in religious clothing.* We manually annotated this resource as "Human Being, Man in General" but we obtained as predicted subject "Religion and Magic". The model we trained was not able to recognize that the textual description was not pointing to a specific iconography (e.g."Deposition of Christ") but instead, was focusing on describing the figure represented by the bust (Pope Pius VII) and the clothing (religious clothing).

- The annotated dataset we used to train the model has demonstrated two main limitations:
  - The Inconclass iconography descriptions only refer to the domain of visual artworks, so it might be useful to extend the dataset also to the Archaeology domain. This is to have a more precise identification of the iconography related to the Classic Mythology, which is the dominant theme in the Greek and Roman vase painting.
  - The Iconclass categories are unbalanced. For example, "Religion and Magic" consists of 12.246 descriptions while "Bible" has 5.961. To obtain better performance, it might be useful to integrate the less represented categories in the Iconclass dataset with more iconography descriptions. This could help, for example, to have a better distinction of the subjects which are semantically close as "Religion and Magic" and "Bible".

6.2 Reusability of the methodology

The methodology we presented aims to develop a generic and machine-processable pipeline to evaluate metadata quality of cultural heritage resources. To achieve this goal we used as main reference the metadata standard from Dublin Core and the guidelines provided by the Italian Istituto Centrale per il Catalogo e la Documentazione (ICCD). As a result, we were able to reproduce this method on a national scale. To apply our methodology to a more global level, such as a digital infrastructure, a few changes are needed. An example would be Europeana, where the metadata are harvested from other national aggregators on a European scale. Applying our methodology in a non-Italian context would have the following limitations:

- Guidelines: To determine high-quality and low-quality metadata, we used the Dublin Core metadata schema and the ICCD guidelines. In particular, we rely on such recommendations for the definition of mandatory, optional, and recommended elements for the Completeness dimension, and for the annotation of the resources in the gold standard dataset to assess the Accuracy and Coherence dimensions. A significant limitation, in this case, is that the guidelines are only applicable at the Italian national scale;
- Language: For the assessment of the Accuracy and Coherence dimensions, the models were trained solely on the Italian language.

To fill these gaps, one possible solution would be to make a survey at the European level about the guidelines adopted for the implementation of metadata at different cultural heritage institutions, to integrate them in the methodology we have proposed and to create training data at least in English.

Bibliography

[1] Mathias M. Adankon and Mohamed Cheriet. *Support Vector Machine*, pages 1303–1308. Springer US, Boston, MA, 2009.

[2] Alessio Palmero Aprosio and Giovanni Moretti. Tint 2.0: an all-inclusive suite for NLP in italian. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[3] Jane Barton, Sarah Currier, and Jessie MN Hey. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In *International Conference on Dublin Core and Metadata Applications*, pages 39–48, 2003.

[4] Abdelhak Belhi, Abdelaziz Bouras, and Sebti Foufou. Leveraging known data for missing label prediction in cultural heritage context. *Applied Sciences*, 8(10), 2018.

[5] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Web Semant.*, 7(1):1–10, January 2009.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Trans-*

actions of the Association for Computational Linguistics, 5:135–146, 2017.

[7] Thomas R Bruce and Diane I Hillmann. The continuum of metadata quality: defining, expressing, exploiting. ALA editions, 2004.

[8] Irene Buonazia and M Emilia Masci. Il pico application profile. un dublin core application profile per il portale della cultura italiana. 2007.

[9] Irene Buonazia, Maria Emilia Masci, and Davide Merlitti. The project of the italian culture portal and its development. a case study: Designing a dublin core application profile for interoperability and open distribution of cultural contents. pages 393–404, 2007.

[10] Toby Burrows, Antoine Brix, Douglas Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Synnove Myking, Lynn Ransom, et al. Modelling the history of medieval and renaissance manuscripts for the mapping manuscript migrations portal. In *Proceedings: Data for History 2020. Modelling Time, Places, Agents.* Humboldt-Universität zu Berlin, 2020.

[11] Lois Mai Chan and Marcia Lei Zeng. Metadata interoperability and standardization–a study of methodology part i. *D-Lib magazine*, 12(6):1082–9873, 2006.

[12] Christian Chiarcos. Interoperability of corpora and annotations. In *Linked Data in Linguistics*, pages 161–179. Springer, 2012.

[13] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12(null):2493–2537, November 2011.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273â–297, September 1995.

[15] Leendert D Couprie. Iconclass: an iconographic classification system. *Art Libraries Journal*, 8(2):32–49, 1983.

[16] Michael Day, Marieke Guy, and A Powell. Improving the quality of metadata in eprint archives. *Ariadne*, 38, 2004.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, 2019.

[18] Sara Di Giorgio. Culturaitalia, the italian national content aggregator in europeana. *Procedia Computer Science*, 38:40–43, 2014.

[19] Sara Di Giorgio, Achille Felicetti, Patrizia Martini, and Emilia Masci. Dati. culturaitalia: a use case of publishing linked open data based on cidoc-crm. In *EMF-CRM@ TPDL*, pages 44–54, 2015.

[20] Martin Doerr. The cidoc crm, an ontological approach to schema heterogeneity. In Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors, *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2005.

[21] Mauro Dragoni, Sara Tonelli, and Giovanni Moretti. A knowledge management architecture for digital cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(3):1–18, 2017.

[22] Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108, 2020.

[23] Dimitris Gavrilis, Dimitra-Nefeli Makri, Leonidas Papachristopoulos, Stavros Angelis, Konstantinos Kravvaritis, Christos Papatheodorou, and Panos Constantopoulos. Measuring quality in metadata repositories. In *International Conference on Theory and Practice of Digital Libraries*, pages 56–67. Springer, 2015.

[24] Peter S Graham. Quality in cataloging: Making distinctions. *Journal of academic librarianship*, 16(4):213–18, 1990.

[25] Angelika Grund. Iconclass. on subject analysis of iconographic representations of works of art. *KO KNOWLEDGE ORGANIZATION*, 20(1):20–29, 1993.

[26] Carolyn Guinchard. Dublin core use in libraries: a survey. *OCLC Systems & Services: International digital library perspectives*, 2002.

[27] Sarah Higgins. The dcc curation lifecycle model. *International journal of digital curation*, 3(1), 2008.

[28] Youichi Ishida, Toshiyuki Shimizu, and Masatoshi Yoshikawa. An analysis and comparison of keyword recommendation methods for scientific data. *International Journal on Digital Libraries*, pages 1–21, 2020.

[29] Amy S Jackson, Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W Cole. Dublin core metadata harvested through oai-pmh. *Journal of Library Metadata*, 8(1):5–21, 2008.

[30] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[31] Anyla Kabashi. *ICONCLASS-classification system for art and iconography*. PhD thesis, University of Zagreb. University of Zagreb, Faculty of Humanities and Social ..., 2019.

[32] Péter Király. A metadata quality assurance framework, 2015. (Research project plan).

[33] Péter Király. A metadata quality assurance framework. *Göttingen: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. Dostopno na: http://pkiraly. github. io/metadata-quality-proiect-plan. pdf [16. 6. 2019]*, 2015.

[34] Péter Király and Marco Büchler. Measuring completeness as metadata quality metric in europeana. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2711–2720. IEEE, 2018.

[35] P. Király and M. Büchler. Measuring completeness as metadata quality metric in europeana. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2711–2720, 2018.

[36] Alexandros Koulouris, Vangelis Banos, and Emmanouel Garoufallou. Aggregating metadata for europeana: the greek paradigm. 2015.

[37] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[38] Elizabeth D Liddy, Eileen Allen, Sarah Harwell, Susan Corieri, Ozgur Yilmazel, N Ercan Ozgencil, Anne Diekema, Nancy McCracken, Joanne Silverstein, and Stuart Sutton. Automatic metadata generation & evaluation. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 401–402, 2002.

[39] Matteo Lorenzini. Automatic metadata curation of the cultural heritage resources. *Proceedings of AIxIA Doctoral Consortium*, (2249):33–37, 2018.

[40] Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Annotated dataset to assess the accuracy of the textual description of cultural heritage records.

[41] Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Computer Assisted Curation of Digital Cultural Heritage Repositories. In *Proceedings of DH2019*. DataverseNL, 2019.

[42] Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Proposta per una valutazione automatica della completeness dei metadati nel contesto delle biblioteche digitali. *DigItalia*, 2:159–167, 2020.

[43] Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. Automatically evaluating the quality of textual descriptions in cultural heritage records. *International Journal on Digital Libraries*, 22(2):217–231, 2021.

[44] Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. On assessing metadata completeness in digital cultural heritage repositories. *Digital Scholarship in the Humanities*, 36:182–188, 11 2021.

[45] Merkourios Margaritopoulos, Thomas Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. Quantifying and measuring metadata completeness. *Journal of the American Society for Information Science and Technology*, 63(4):724–737, 2012.

[46] Thomas Margaritopoulos, Merkourios Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. A fine-grained metric system for the completeness of metadata. In *Research Conference on Metadata and Semantic Research*, pages 83–94. Springer, 2009.

[47] Davide Merlitti. Linee guida per lo sviluppo di sistemi informatici interoperabili con CulturaItalia, 2005.

[48] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[49] William E Moen, Erin L Stewart, and Charles R McClure. Assessing metadata quality: Findings and methodological considerations from an evaluation of the us government information locator service (gils). In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98-*, pages 246–255. IEEE, 1998.

[50] Jehad Najjar, Stefaan Ternier, and Erik Duval. The actual use of metadata in ariadne: an empirical analysis. In *Proceedings of the 3rd Annual ARIADNE Conference*, pages 1–6. Citeseer, 2003.

[51] David Newman, Kat Hagedorn, Chaitanya Chemudugunta, and Padhraic Smyth. Subject metadata enrichment using statistical topic

models. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, page 366–375, New York, NY, USA, 2007. Association for Computing Machinery.

[52] NISO Framework Working Group (with support from the Institute of Museum and Library Services). A framework of guidance for building good digital collections. *Baltimore, MD: National Information Standards Organization (NISO)*, 2007.

[53] Xavier Ochoa and Erik Duval. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10(2-3):67–91, 2009.

[54] Davor Ostojic, Go Sugimoto, and Matej Ďurčo. Curation module in action-preliminary findings on vlo metadata quality. In *Proceedings of the CLARIN Annual Conference 2016*, Aix-en-Provence, 2017.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[56] Audrey Romero Pelaez and Pedro P Alarcon. Metadata quality assessment metrics into ocw repositories. In *Proceedings of the 2017 9th International Conference on Education Technology and Computers*, pages 253–257, 2017.

[57] Maureen Pennock. Digital curation: a life-cycle approach to managing and preserving usable digital information. *Library & Archives*, 1:34–45, 2007.

[58] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly, 2012.

[59] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

[60] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[61] Besiki Stvilia, Les Gasser, Michael B Twidale, and Linda C Smith. A framework for information quality assessment. *Journal of the American society for information science and technology*, 58(12):1720–1733, 2007.

[62] Go Sugimoto. Number game - experience of a european research infrastructure (clarin) for the analysis of web traffic. In *Selected papers from the CLARIN Annual Conference 2016 Aix-en-Provence 26–28 October 2016*, Aix-en-Provence, 2017. HAL.

[63] Alice Tani, Leonardo Candela, and Donatella Castelli. Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, 49(6):1194–1205, 2013.

[64] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[65] Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

[66] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Appendix A

Appendix

In the following tables we report the most relevant metadata schema and ontologies used to describe the digital resources.

Name	Focus	Description
DDI	Archiving and Social Science	The Data Documentation Initiative is an international effort to establish a standard for technical documentation describing social science data. A membership-based Alliance is developing the DDI specification, which is written in XML.
EBUCore	The EBUCore metadata set for audiovisual content	EBUCore is a set of descriptive and technical metadata based on the Dublin Core and adapted to media. EBUCore is the flagship metadata specification of EBU [1], the largest professional association of broadcasters around the world. It is developed and maintained by EBU's Technical Department [2]. EBU has a long history in the definition of metadata solutions for broadcasters [3]. EBUCore is largely used as shown is this report [4]. EBUCore is registered in SMPTE. It is also available in RDF [5] and the documentation from [6].
EBU CCDM	The EBU Class Conceptual Data Model - CCDM	The EBU Class Conceptual Data Model (CCDM) is an ontology defining a basic set of Classes and properties as a common vocabulary to describe programmes in their different phases of creation from commissioning to delivery. CCDM is a common framework and users are invited to further enrich the model with Classes and properties fitting more specifically their needs.
FOAF	Friend of a Friend (FOAF)	The Friend of a Friend (FOAF) project is about creating a Web of machine-readable homepages describing people, the links between them and the things they create and do.
EAD	Archiving	Encoded Archival Description. A standard for encoding archival finding aids using XML in archival and manuscript repositories.
CDWA	Arts	Categories for the Description of Works of Art is a conceptual framework for describing and accessing information about works of art, architecture, and other material culture.
VRA Core	Arts	Visual Resources Association. The standard provides a categorical organization for the description of works of visual culture as well as the images that document them.
ONIX	Book industry	Online Information Exchange. International standard for representing and communicating book industry product information in electronic form.

Table A.1: Metadata Standard Schema Table 1

Name	Focus	Description
TEI	Humanities, social sciences and linguistics	Text Encoding Initiative - a standard for the representation of texts in digital form, chiefly in the humanities, social sciences and linguistics.
NISO MIX	Images	Z39.87 Data dictionary - technical metadata for digital still images (MIX) - NISO Metadata for Images in XML is an XML schema for a set of technical data elements required to manage digital image collections.
MARC	Librarianship	MARC - MAchine Readable Cataloging - standards for the representation and communication of bibliographic and related information in machine-readable form.
METS	Librarianship	Metadata Encoding and Transmission Standard - an XML schema for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.
MODS	Librarianship	Metadata Object Description Schema - is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications.
XOBIS	Librarianship	XML Organic Bibliographic Information Schema - a XML schema for modeling MARC data.
MPEG-7	Multimedia	The Multimedia Content Description Interface MPEG-7 is an ISO/IEC standard and specifies a set of descriptors to describe various types of multimedia information and is developed by the Moving Picture Experts Group.
MEI	Music notation	Music Encoding Initiative is a community-driven effort to create a commonly accepted, digital, symbolic representation of music notation documents.
Dublin Core	Networked resources	Dublin Core - interoperable online metadata standard focused on networked resources.
DOI	Networked resources	Digital Object Identifier - provides a system for the identification and hence management of information ("content") on digital networks, providing persistence and semantic interoperability.
DIF	Scientific data sets	Directory Interchange Format - a descriptive and standardized format for exchanging information about scientific data sets.
RAD	Librarianship and archiving	The Rules for Archival Description (RAD) is the Canadian archival descriptive standard. It is overseen by the Canadian Committee on Archival Description of the Canadian Council of Archives.[10] Similar in structure to AACR2, it was last revised in 2008.[11]
RDF	Web resources	General method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats.
CMDI	Linguistic resources	It provides a framework to describe and reuse metadata blueprints.
EDM	Digital library	Formal specification of the classes and properties that could be used in Europeana.
PICO-AP	Digital library	Formal specification of the classes and properties that could be used in Cultura Italia.
CIDOC-CRM	Web resources	Theoretical and practical tool for information integration in the field of cultural heritage.
FRBRoo	Web resources	Formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information.

Table A.2: Metadata Standard Schema Table 2