UNIVERSIDADE FEDERAL FLUMINENSE

JUAN RIBEIRO REIS

An Approach for Assessing Metadata Completeness in Open Data Portals

NITERÓI

2019

UNIVERSIDADE FEDERAL FLUMINENSE

JUAN RIBEIRO REIS

An Approach for Assessing Metadata Completeness in Open Data Portals

Master thesis presented to the Graduate Program in Computer Science of the Universidade Federal Fluminense, in partial fulfillment of the requirements for the degree of Master of Science. Concentration area: Systems and Information Engineering.

Supervisor:
JOSÉ VITERBO

Co-Supervisor:
FLAVIA BERNARDINI

NITERÓI

2019

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

R375a Reis, Juan Ribeiro
An Approach for Assessing Metadata Completeness in Open Data Portals / Juan Ribeiro Reis ; José Viterbo Filho, orientador ; Flavia Cristina Bernardini, coorientador. Niterói, 2019.
84 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2019.

DOI: http://dx.doi.org/10.22409/PGC.2019.m.12597680746

1. Metadado. 2. Dados Governamentais Abertos. 3. Portal de Dados Abertos. 4. Produção intelectual. I. Viterbo Filho, José, orientador. II. Bernardini, Flavia Cristina, coorientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

JUAN RIBEIRO REIS

An Approach for Assessing Metadata Completeness in Open Data Portals

Master thesis presented to the Graduate Program in Computer Science of the Universidade Federal Fluminense, in partial fulfillment of the requirements for the degree of Master of Science. Concentration area: Systems and Information Engineering.

Approved in April of 2019.

APPROVED BY

DSc. José Viterbo Filho, UFF

DSc. Flavia Cristina Bernardini, UFF

DSc. Daniel Cardoso Moraes de Oliveira, UFF

DSc. Cristiano Maciel, UFMT

Niterói

2019

Epigraph: I dedicate this work to my parents, Lina and Altair for having supported my journey so far, for always believing in me and always has helped me in all that they could. My grandmother (in memoriam), for inspiring me to follow my path and encourage me to this day. To my brother, for showing me different perspectives in my career and in my life. My wife, for having supported me in all that work.

Acknowledgements

I thank Professor José Viterbo, who at all times helped me in several ways, and also transmitted to me the spirit of research that was lacking for my insertion in the academic environment.

I thank Professor Flavia Bernardini, who with all her critical eye made me see several points of view that were not visible to me.

I thank UFF and his staff of teachers and staff, who trained and supported me in every way with great care and effort.

I thank UERJ and its staff of professors and staff, who have given me the academic foundation necessary to take on new flights.

Resumo

Cidadãos e desenvolvedores estão obtendo amplo acesso a fontes públicas de dados, disponibilizadas em portais de dados abertos. Esses conjuntos de dados legíveis por máquina permitem a criação de aplicativos que ajudam a população de várias maneiras, dando-lhes a oportunidade de participar ativamente dos processos de governança, como a tomada de decisões e a formulação de políticas. Embora o número de portais de dados abertos cresça ao longo dos anos, os pesquisadores identificaram problemas recorrentes com os dados que eles fornecem. Um dos problemas recorrentes citados na literatura está relacionado aos metadados associados a cada conjunto de dados. Os metadados são vitais durante todo o ciclo de vida dos dados abertos. A má qualidade dos metadados leva a descrições ou classificações inadequadas de conjuntos de dados, o que afeta diretamente a usabilidade e a capacidade de pesquisa dos recursos. Uma importante métrica de metadados de qualidade que é abordada várias vezes na literatura diz respeito à completude. Completude é o grau em que uma instância de metadados contém todas as informações necessárias para ter uma representação abrangente do recurso descrito. Neste trabalho, propomos uma abordagem baseada no alinhamento de esquemas de metadados para avaliar o grau de completude dos metadados de conjuntos de dados disponíveis em portais de dados abertos. Para isso, criamos um processo dividido em dois subprocessos, onde o primeiro subprocesso alinha os esquemas escolhidos, pondera seus campos e, assim, gera um esquema padrão de completude de metadados. O segundo subprocesso alinha os campos de metadados dos conjuntos de dados com o esquema gerado no primeiro subprocesso, gerando uma avaliação de completude de metadados. Para avaliar a abordagem proposta, realizamos dois estudos de caso, nos quais aplicamos a abordagem para avaliar os registros de metadados de conjuntos de dados reais disponíveis em diferentes portais de dados abertos, incluindo o Portal Europeu de Dados Abertos, que reúne conjuntos de dados de vários países da União Européia, e o Portal de Dados Abertos do Estado de Nova York, que reúne conjuntos de dados de um dos estados mais críticos dos Estados Unidos e o centro financeiro e comercial mais significativo do país. Com o experimento, concluímos que a abordagem é consistente, sendo capaz de avaliar na prática a completude de metadados de portais de dados abertos referentes a um esquema de metadados de referência. Além disso, a abordagem é adaptavel e pode ser aplicada considerando diferentes esquemas para criar o esquema de referência. Em cada caso, esses esquemas podem ser escolhidos para melhor atender ao objetivo do processo de avaliação.

Palavras-chave: Completude de Metadados, Dados Governamentais Abertos, Portal de Dados Abertos.

Abstract

Citizens and developers are gaining full access to public data sources made available in open data portals. These machine-readable datasets allow the creation of applications that help the citizens in a variety of ways, allowing them to actively participate in government processes such as decision-making and policy-making. While the number of open data portals has grown over the years, researchers have been able to identify several drawbacks regarding the data they provide. One of the recurring problems cited in the literature is related to metadata associated with each dataset. Metadata is vital throughout the open data life cycle. Poor metadata quality leads to inadequate descriptions or classifications of datasets, which directly affects the usability and searchability of resources. An important quality metadata metric that is addressed several times in the literature concerns completeness. Completeness is the degree to which a metadata instance contains all the information needed to have a comprehensive representation of the described resource. In this work, we propose an approach based on the alignment of metadata schemas to assess the degree of completeness of the metadata of datasets available in open data portals. For this, we create a process divided into two subprocesses, where the first subprocess aligns the chosen schemas, ponders its fields and thus generates a metadata completeness standard schema. The second subprocess aligns the metadata fields of the datasets with the schema generated in the first subprocess, thus making a metadata completeness assessment. To evaluate the proposed approach, we conducted two case studies, in which we applied the approach to assess the metadata records of real datasets available in different open data portals, including the European Open Data Portal, that gathers datasets from various countries of the European Union, and the New York State Open Data Portal,that gathers datasets from one of the most critical states in the United States and the most significant financial and commercial center in the country. With the experiment, we concluded that the approach is consistent, being able to assess in practice the completeness of metadata of open data portals concerning a reference metadata scheme. Also, the approach is adaptive and can be applied considering different schemes to create the reference scheme. In each case, these schemes may be chosen to best suit the primary purpose of the evaluation process.

Keywords: Metadata Completeness, Open Government Data, Open Data Portals.

List of Figures

List of Tables

List of Abbreviations and Acronyms

API	:	Application Programming Interface
CKAN	:	Comprehensive Knowledge Archive Network
CSV	:	Comma-separated values
ID	:	Identification
JSON	:	Javascript Object Notation
NYS	:	New York State
OGD	:	Open Government Data
PSI	:	Public Sector Information
RDF	:	Resource Description Framework
RESTful API	:	Representational State Transfer Application Programming Interface
URL	:	Uniform Resource Locator
W3C	:	World Wide Web Consortium

Contents

Chapter 1

Introduction

Nowadays, the publication of open government data is widely disseminated among several countries, comprising all the different administrative levels (HUIJBOOM; BROEK, 2011). In such a scenario, the population is gaining access to data from the various sectors of public activity, such as security, health, transportation, among others. As an outcome, governments are moving towards a more transparent administration, in which citizens can have access to government-produced data, discovering various information relevant to their daily life (RIBEIRO; ALMEIDA, 2011).

Also, developers have straight access to public data sources made available in open data portals. These machine-readable datasets available for re-use, enable the creation of applications that help the population in several ways, allowing them to participate in governance processes actively, such as decision and policy-making (ATTARD et al., 2015), as seen in Colombia (ROJAS et al., 2014), Brazil (BEGHIN et al., 2014), India (CHATTAPADHYAY, 2014), among others. Furthermore, the provision of public information on a variety of themes, like health, income, human development, and others — brings greater visibility also through various media such as newspapers, websites or television programs that amplify the dissemination of open government data (STARKE et al., 2016).

Hence, the number of open data portals and the volume of data they provide are growing at a rapid pace worldwide (RIBEIRO et al., 2015; TYGEL et al., 2016). Authors of (TYGEL et al., 2016) point out that the number of data portals grows fast over the years. In 2012, there were already 115 Portals of this nature available, offering about 710,000 data sets (HENDLER et al., 2012). According to (INTERNATIONAL, 2011), there are currently 551 open government data portals available in all continents in 2019, proving the accelerated rise of this paradigm.

With this increase, problems that affect the use of open data, such as lack of data standards, difficulty to access the data, poor data understandability, datasets with irrelevant information, and others, are becoming more evident. One of the recurring problems cited by the literature is related to dataset metadata (REIS et al., 2018). They are of extreme importance throughout the data life cycle. This type of problem occurs in several domains, such as digital libraries (BEALL, 2005), in open data portals (MARGARITOPOULOS et al., 2012), among others.

Discoverability is an essential factor for datasets (BRAUNSCHWEIG et al., 2012). Without sufficient metadata, such as descriptions or tags, neither manual nor automatic search can find the dataset. Metadata help to create order in datasets by describing, classifying and organizing information(ZUIDERWIJK et al., 2012). Also, metadata improves the accessibility of data by helping to describe, locate and retrieve the data efficiently. According to (WEBFOUNDATION, 2017), data is hard to use because there is no metadata or guidance documentation available. Less than a third (31%) of the published datasets have some supporting basic metadata or companion guidance documentation. Complete, high-quality government data and metadata is still challenging to find. Neumaier et al. (2016) concludes that there are metadata quality issues that could disrupt the success of Open Data: inadequate descriptions or classifications of datasets directly affect the usability and searchability of resources. Several quality metrics demonstrate how useful metadata describes datasets. The author (OCHOA; DUVAL, 2009) defines, for example, metadata quality metrics as accuracy, conformance to expectations, logical consistency and coherence, among others.

Literature address several times the completeness of the metadata as one important metric of metadata quality. Completeness is the degree to which the metadata instance contains all the information needed to have a comprehensive representation of the described resource, as defined by (OCHOA; DUVAL, 2009). It is measured based on the presence or absence of values in metadata fields [1] defined in different metadata standards (MARGARITOPOULOS et al., 2012). In some research efforts, the fields to be considered when completeness is measured are selected by the application that uses the metadata based on their importance for the specific process or activity handling the metadata.

According to (MARGARITOPOULOS et al., 2012), an incomplete metadata record is a record of degraded quality. There are several metadata standards in an attempt to cover the maximum of features that describe the data itself efficiently. In theory, the dataset

[1] The terms “metadata field” and “metadata element” are used interchangeably throughout this article.

would have a complete and useful description if all its metadata were filled correctly. However, as pointed out in several surveys, reality shows that the effort to fill several metadata fields is expensive and time-consuming. Therefore, metadata is often sparsely populated.

Relevant surveys (FRIESEN, 2004) (GUINCHARD, 2002) (NAJJAR et al., 2003) have shown that data publisher tend to fill out only particular metadata elements that could be considered "popular", while they ignore other elements of less popularity. The creation of metadata is a task requiring significant labor and financial cost and, most importantly, the involvement of knowledgeable and experienced people (BARTON et al., 2003)(LIDDY et al., 2002).

Some methods attempt to improve the quality of datasets by placing metadata with relevance when attempting to achieve a higher standard of data value. One example is data governance (NWABUDE et al., 2014). In (KHATRI; BROWN, 2010), one of the Khatri and Brown's data governance decision domains is the domain of metadata, where they classify as having an essential role in the discovery, retrieval, collation and analysis of data. It be can mention (REIS et al., 2018), where the authors trace a parallel between data governance and open data and affirm that the administrator must perform a search to identify the best set of characteristics that describe their datasets, making it easier for users to retrieve information.

1.1 Problem Setting

One of the main problems related to metadata concerns the completeness of its fields. Several studies point to this problem, using it as an indicator of the quality of its metadata. Several studies point out this problem by using it as a quality indicator of their metadata (BRÜMMER et al., 2014) (REICHE; HÖFIG, 2013) (DUVAL et al., 2002). The lack of metadata filling causes the searchability of the datasets present in the portals to be compromised and becomes a problem for users because they can not access the information they need. So, Portal administrators need to evaluate how much filled are the metadata of their datasets. Existing approaches to measuring metadata completeness limit their scope in counting the existence of values in fields, regardless of the metadata field importance. It is necessary to consider several issues that a traditional approach overlooks. To this end, an approach is needed to enable them to assess what action to take.

1.2 Goals

This research aims to propose an approach for measure the completeness of metadata fields for publication of open government data, contributing to improving automated access by human consumers and external systems, allowing better use of this data for the development of computational models of knowledge to be applied in several areas. The specific objectives of this work are:

1. Carry out a literature review to look for the main problems with the metadata.
2. Do a literature review concerning metadata completeness quality measurement.
3. Propose an approach for creating a quality measurement of completeness of metadata.
4. Validate the proposed approach.

1.3 Organization

Besides this chapter, we present a contextualization, the problem context, and the enumeration of the goals, this work contains other chapters organized and distributed as follows:

- Chapter 2: This chapter presents the theoretical foundation that addresses the main knowledge used in this dissertation, emphasizing the main themes of this research. Also, this chapter presents the main works related to the proposed approach.
- Chapter 3: This chapter presents the proposed approach, which assesses metadata records from open data portals.
- Chapter 4: This chapter presents the experimental analysis done with the proposed approach applied to metadata data from the open European Union and NYS (New York State) data portals and their results.
- Chapter 5: Finally, this chapter presents the conclusions about the work presented, as well as some suggestions for future work.

Chapter 2

Background and Literature Review

This chapter discusses the theoretical basis of this dissertation, necessary for the analysis and understanding of the elements of the presented approach, as well as the understanding of its implementation and own validation process. The sections were distributed as follows.

First, in Section 2.1, the concept of open data and its importance to society are explained. After that, we define open government data, and we present its relevance to all other productive segments of society in Section 2.2. Next, we elucidate in Section 2.3 what open data frameworks are, and some of the leading most used. Later, Section 2.4 presents metadata and its nuances for open data. After that, Section 2.5 defines the common terms related to metadata. Finally, in Section 2.6 explains what metadata schema is, and its quality assessments are defined in Section 2.7 and so its main issues in Section 2.8.

2.1 Open Data

According to (INTERNATIONAL, 2005), open data are data that can be freely used, re-used, and redistributed by anyone - subject, at most, to the requirement of source allocation and sharing by the same rules. Open data initiatives aim to open all non-personal and non-commercial data, especially (but not exclusively) all data collected and processed by government organizations. It is very similar in spirit to the open source or open access movements(BRAUNSCHWEIG et al., 2012).

Initially, the term gained popularity in the academic circles as a movement aimed at the development of open scientific communities by ensuring free access to academic data published in special digital depositories (MURRAY-RUST, 2008).

Later, the idea gained a political meaning, especially, with the launch of open-data government projects such as data.gov in the United States. In January 2009, President Barack Obama announced that his administration would start a transparency strategy which would imply an unprecedented level of openness in government. In a memorandum for the Heads of Executive Departments and Agencies, he stated that (OBAMA, 2009) "[...] We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in government.".

For data to be opened, it must be available as a whole and at no higher cost than a reasonable cost of reproduction, preferably possible to be downloaded through the internet (MOLLOY, 2011). The data must also be available in a convenient and modifiable way by anyone. They also need to be provided in a way that allows re-use and redistribution, including combining with other data sets. Besides, any individual group or areas of action should not be discriminated against and have free access to use, re-use, and redistribute.

For open data to utilize their full potential, they need an essential feature: interoperability, i.e., the ability of various systems and organizations to work together to interoperate - in this case, the possibility of interoperating different sets of data. For this, linked open data is necessary (BAUER; KALTENBÖCK, 2011). To fully benefit from open data, it is crucial to put information and data into a context that creates new knowledge and enables robust services and applications. As Linked Open Data facilitates innovation and knowledge creation from interlinked data, it is an essential mechanism for information management and integration. Linked Open Data is becoming increasingly important in the fields of state-of-the-art information and data management. It is already being used by many well-known organizations, products, and services to create portals, platforms, internet-based services, and applications.

In (BAUER; KALTENBÖCK, 2011), the author shows that data is still locked up in specific applications. The technical problem with today's most common information architecture is that metadata and schema information is not separated well from application logic. It is not possible to re-use data as rapidly as it should be. When a database is designed, it is often known that specific applications are built on top. If there is no longer an emphasis on which applications will use the data and focus on meaningfully describing the data itself, there will be long-term gains.

Another crucial point is the data files format provided. These should have free formats so that they can be used and implemented by anyone. Moreover, for published datasets

to be re-used and shared by all, they must be in the public domain or provided under an open license (INTERNATIONAL, 2005). The license should elucidate aspects such as commercial use, sharing, among others.

2.2 Open Government Data

For the purposes of (OECD, 2014), PSI (Public Sector Information) is broadly defined as "information, including information products and services, generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for a government or public institution", taking into account the legal requirements and restrictions referred to in the last paragraph of the preamble of the recommendation.

More than a decade ago, the potential for a significant part of PSI to be re-used outside the public sector for various commercial and non-commercial social purposes was recognized. In 2003, the European Union adopted the 'Directive on the Re-use of Public Sector Information' (COX; ALEMANNO, 2003), which encourages the member states to make as much of the information they possess available for re-use as possible. It establishes a minimum set of rules as well as practical means for facilitating this re-use, focusing mainly on its economic aspects.

The term OGD (Open Government Data) has come into prominence relatively recently, becoming popular in 2008 after the publication of a set of open government data principles by advocates in the United States. According to (OECD, 2017), OGD is a philosophy, and increasingly a set of policies, that promotes transparency, accountability, and value creation by making government data available to all. Public bodies produce and commission vast quantities of data and information of different types to perform their tasks. The extraordinary quantity and centrality of data collected by governments make these data particularly significant as a resource for increased public transparency. By encouraging the use, re-use, and free distribution of datasets, governments promote business creation and innovative, citizen-centric services. OGD can be used to help the public better understand what the government does and how well it performs, and to hold it accountable for wrongdoing or unachieved results (UBALDI, 2013).

Theoretically, the primary value of open data as a concept is that in providing free public access to various official files the government not only becomes presumably more transparent but also more efficient as it potentially could promote civic engagement by enabling citizens to participate in various discussions on how to better address their needs

(KASSEN, 2013).

As the open data movement grows, and even more governments and organizations sign up to open data, it becomes even more critical that there is a clear and agreed definition for what "open data" means if we are to realize the full benefits of openness and avoid the risks of creating incompatibility between projects and splintering the community (WESSELS et al., 2017).

The new Directive highlights the importance of PSI as a vast, diverse, and valuable pool of resources that can benefit the knowledge economy and encourages the proliferation of OGD portals (ZIJLSTRA; JANSSEN, 2013). It includes policies for 'encouraging the wide availability and re-use of PSI for private or commercial purposes, with minimal or no legal, technical or financial constraints, and promoting the circulation of information not only for economic operators but also for the public, which can play an important role in kick-starting the development of new services based on novel ways to combine and make use of such information, stimulate economic growth and promote social engagement'.

2.3 Open Data Frameworks

Open data frameworks are web-based interfaces designed to make it easier to find re-usable information. Like library catalogs, they contain metadata records of datasets published for re-use, i.e., mostly relating to information in the form of raw, numerical data and not to textual documents. In combination with specific search functionalities, they facilitate finding datasets of interest. APIs (Application Programming Interface) are also often available, offering direct and automated access to data for software applications.

These portal frameworks provide ecosystems to describe, publish, and consume datasets, i.e., metadata descriptions along with pointers to data resources. Portal software frameworks typically consist of a content management system, some query and search features as well as RESTful API (Representational State Transfer Application Programming Interface) to allow agents to interact with the platform. The metadata usually can be retrieved in a structured format via the API (JSON (Javascript Object Notation) data). However, the metadata schemas are heterogeneous concerning the underlying software framework.

There exists three prominent software frameworks for publishing Open Data, the commercial Socrata Open Data portal; the open source framework CKAN (Comprehensive Knowledge Archive Network), developed by the Open Knowledge Foundation; and the

data publishing platform OpenDataSoft, deployed mainly in French Open Data portals (NEUMAIER et al., 2016).

CKAN (NETWORK, 2017) is the world's leading open-source data management system. It helps users from different domains (national and regional governments, companies, and organizations) to easily publish their data through a set of workflows to publish, share, search, and manage datasets. CKAN is a complete catalog system with an integrated data storage and powerful RESTful JSON API. It offers a rich set of visualization tools (e.g., maps, tables, charts) as well as an administration dashboard to monitor datasets usage and statistics. CKAN allows publishing datasets either via an import feature or through a web interface. Relevant metadata describing the dataset and its resources, as well as organization related information, can be added.

Socrata (SOCRATA, 2005) is a commercial platform to streamline data publishing, management, analysis, and re-using. It empowers users to review, compare, visualize, and analyze data in real time. Datasets hosted in Socrata can be accessed using RESTful API that facilitates search and data filtering. Socrata allows flexible data management by implementing various data governance models and ensuring compliance with metadata schema standards. It also enables administrators to track data usage and consumption through dashboards with real-time reporting. Socrata is very flexible when it comes to customization. It has a consumer-friendly experience allowing users to tell their story with data. Socrata's data model is designed to represent tabular data: it covers a basic set of metadata properties and has excellent support for geospatial data.

OpenDataSoft (OPENDATASOFT, 2018) is a cloud-based platform aiming to act as a resource for firms to publish and retrieve open data. The OpenDataSoft platform uses an API for search based retrieval and analysis of published data sets. Parameters allow for complex queries using full-text or geolocation search strategies, and for exporting a flow of records in JSON or CSV (Comma-separated values).

2.4 Metadata

Metadata is commonly defined as "data about data", according to its etymology. Metadata summarizes necessary information about data, which can make finding and work with particular instances of data more accessible. In (GREENBERG, 2003), the author defines metadata as "structured data about an object that supports functions associated with the designated object". Metadata structure involves the systematic organization of

data, and this is now accomplished mainly through the use of metadata schema. The functions enabled can be diverse, but they are in many cases related to facilitating discovery or search, or to restrict access (e.g., in the case of licensing information) or to combine meta-information to relate resources described separately. Metadata establishes the semantics or "content" of the data to be interpreted by users (KHATRI; BROWN, 2010). It explains what the data is and provides the mechanism that consistently describes the data representation, thus helping to interpret the meaning or semantics of data. Metadata helps to connect humans and machines and enables knowledge sharing across domains (GREENBERG; GAROUFALLOU, 2013). Accurate, consistent, sufficient, and thus, reliable metadata is a powerful tool that enables the user to discover and retrieve relevant materials quickly and easily and to assess whether they may be suitable for re-use (GREENBERG; ROBERTSON, 2002). Metadata can be created manually or by automated information processing. Manual creation tends to be more accurate, allowing the user to input any information they feel is relevant or needed to help describe the file. Automatic metadata creation can be much more elementary, usually only displaying information such as file size, file extension, when the file was created and who created the file.

```
:dataset-001
    a dcat:Dataset ;
    dct:title "Imaginary dataset" ;
    dcat:keyword "accountability","transparency" ,"payments" ;
    dct:issued "2011-12-05"^^xsd:date ;
    dct:modified "2011-12-05"^^xsd:date ;
    dcat:contactPoint <http://example.org/transparency-office/contact> ;
    dct:temporal <http://reference.data.gov.uk/id/quarter/2006-Q1> ;
    dct:spatial <http://www.geonames.org/6695072> ;
    dct:publisher :finance-ministry ;
    dct:language <http://id.loc.gov/vocabulary/iso639-1/en>  ;
    dct:accrualPeriodicity <http://purl.org/linked-data/sdmx/2009/code#freq-W>  ;
    dcat:distribution :dataset-001-csv ;
    .
```

Figure 2.1: Example of metadata dataset

Data published dictates how the open data portals generate metadata, such as statistical, geographical, or financial, as these are linked to the metadata standards that they employ (LISOWSKA, 2016). These standards are a direct response to the need for descriptive, structured information on the data published on any given subject. Metadata schemas rule them.

2.5 Metadata Definitions

In this section, we present some definitions of several terms related to metadata usually quoted used in this dissertation, according to author (GREENBERG, 2003).

Metadata schema – A unified and structured set of rules developed for object documentation and functional activities. A schema is a conceptualization that in a specification is represented or formalized. Section 2.6 explains thoroughly.

Metadata specification – A proper representation for human and/or machine processing of a schema conceptualization. Specifications provide metadata element semantics and often syntactic and schema application rules. Different specifications produced over time and distinguished by the version or release numbers can represent a single schema.

Data dictionary (metadata dictionary) – A subsystem of a database that records the definitions (semantics) for all the metadata elements used in a database. A data dictionary may also include detailed documentation about the relationships among metadata elements, as well as syntax and schema application rules. The term data dictionary comes from the relational database community and may be viewed as a type of metadata specification.

Metadata elements – Properties of the object that are defined in a specification. 'Author/creator', 'title' and 'subject' are properties commonly identified as metadata elements. Metadata elements may also be defined as object attributes.

Metadata semantics – Definitions of metadata elements delineated in a specification, data dictionary, or other resources. A comment or examples may support the semantic definition of a metadata element and may reference metadata qualifiers, including attribute value schemas.

Metadata vocabulary – The term metadata vocabulary is used in two distinct ways: for metadata schemas and metadata specifications.

Metadata label – The public name for a metadata element. The label identifies the metadata for the end user and supports searching, administrative activities, and other functions that involve user interaction.

Metadata records – An organized collection of metadata elements with content values that represent an object.

2.6 Metadata Schema

According to (STANDARDIZATION, 2006), a schema is a logical plan showing the relationships between metadata elements, usually through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionally (obligation level) of values. A metadata schema provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that assists the identification, discovery, and use of information within that discipline (CHAN; ZENG, 2006). In the literature, the words 'schema', 'scheme', and 'element set' have been used interchangeably to refer to metadata standards. In practice, the word 'schema' usually refers to an entire entity including the semantic and content components (which are usually regarded as an 'element set') as well as the encoding of the elements with a markup language. A metadata element set has two primary components:

- Semantics – Definitions of the meanings of the elements and their refinements.
- Content – Declarations or instructions of what and how values should be assigned to the elements.

For each element defined, a metadata standard usually provides content rules for how content should be included (e.g., how to identify the main title), representation rules for content (e.g., capitalization rules or standards for representing time), and allowable content values (e.g., whether values must be taken from a specified controlled vocabulary or can be author-supplied, derived from text, or added by metadata creators working without a controlled term list.).

In this work, in addition to using the Schemas related to the frameworks CKAN, Socrata and Opendatasoft, DCAT-AP schema is used as well. The DCAT is a W3C (World Wide Web Consortium) metadata recommendation for publishing data on the Web (MAALI; ERICKSON, 2014). DCAT is defined in RDF (Resource Description Framework) and re-uses the Dublin Core Metadata vocabulary. The recent DCAT-AP (COMMISSION, 2018) for data portals in European Union extends the DCAT core vocabulary and aims towards the integration of datasets from different European data portals. It extends the existing DCAT schema by a set of additional properties. The European Union Data Portal, which currently harvests 68 European data portals, supports DCAT-AP metadata. Because of this, it is a schema that facilitates interoperability between data portals published on the Web.

2.7 Metadata Assessment

Several researchers have created metrics to measure metadata quality by computing indicators of quality and, among them, the completeness indicator of a metadata record.

In (MOREIRA et al., 2009), the authors present a tool to perform an automatic evaluation of a digital library. For the evaluation of metadata specifications, it is considered two quality indicators. The metadata completeness, which reflects how many of the attributes specified in the standard metadata have their values defined in a metadata specification. The metadata conformance, which indicates whether the metadata attributes and their respective values in the metadata specification, follow the rules defined in a given metadata standard.

In (KUBLER et al., 2018) develops an Open Data Portal Quality framework that enables end-users to quickly and in real-time assess/rank open data portals. For this work, the authors classified into five main categories of quality indicators for metadata:

- Existence – The existence of essential metadata keys.
- Conformance – the adherence of metadata information to a particular format if exist.
- Retrievability – The availability and retrievability of the metadata and data.
- Accuracy – If the information accurately describes the underlying resources.
- Open Data – If the specified format and license information suitable to classify a dataset as open.

Several studies use with frequency the quality metric that refers to completeness. It is one of the most critical metrics, and for this reason, we choose to analyze it in this study.

2.8 Metadata Issues in Open Data Portals

There are several open data portals around the world, with a broad set of datasets with all kind of information to be accessed by anyone from anywhere. However, users and researchers have reported problems in some aspects that difficult the use of the data published by governments. Reis et al. (2018) identified several problems that frequently occur in several open data portals of various countries. One of the problems repeated in

several works that the authors have identified that damages data portals open around the globe says about poor metadata quality, which the author defined as 'the absence or poor quality of metadata (information about datasets).'.

The lack of standard data, data formats issues, and poor metadata are pointed out in (BRAUNSCHWEIG et al., 2012), where the authors show the importance of standardizing data in several aspects and making datasets available in standard formats in order to facilitate their use. They concluded that most platforms lack adequate standards and APIs, and have released much information that is not machine-readable or in a proprietary format that prohibits automated tools from re-using, which implies that many open datasets are not open at all. The surveyed repositories also provide varying degrees of metadata, which causes problems in integrating open datasets from different platforms.

The authors (ZUIDERWIJK et al., 2014) presents examples of barriers to open data processes. They identified some issues like lack of data completeness, difficult data access and availability, Poor understandability, among others. In addition to these issues, they recognized problems with metadata, in an open data operational perspective, such as unfindable metadata, metadata interoperability difficulties, lack of metadata about data quality, among others.

According to (MARGARITOPOULOS et al., 2012), an incomplete metadata record is a record of degraded quality. There are several metadata standards in an attempt to cover the maximum of features that describe the data itself efficiently. In theory, if all metadata were filled correctly, it would have a complete and useful description of the data set. However, as pointed out in several surveys, reality shows that the effort to fill several metadata fields is expensive and time-consuming. Therefore, metadata is often sparsely populated.

In the library domain, the authors of (BEALL, 2005) describes the major types of data quality errors that occur both in full-text objects and in metadata in digital libraries. Some types of common errors include typographic errors, errors in scanning and data conversion, and errors in finding and replacing. Metadata errors can also hamper digital library access.

2.9 Literature Review

In the literature, the authors have identified several problems, including filling in the metadata fields, trying to measure them to qualify the datasets or portals that contain

them or even correcting them. In order to measure the completeness of the metadata, the approach proposed was based on work from two different domains: digital libraries and open data. The work had a more significant reference concerning the domain of open data.

2.9.1 Digital Libraries Domain

Király and Büchler (2018) analyzed Europeana Metadata quality. Europeana, the European Cultural Heritage Digital Platform, has a varied collection of metadata records from over 3200 data providers. This paper proposes an open source method and implementation to measure some of these data's structural features, such as completeness, multilingualism, uniqueness, record patterns, to reveal quality issues. In the research, the relationship between functionality and the metadata schema are rethought and implements a framework, which proved to be successful in measuring structural features which correlate with metadata issues. The user of the framework can select low and high-quality records. According to the hypothesis, structural features such as existence and cardinality of fields correlate with metadata quality, and it proved to be true. The research extended the volume of the analyzed records by introducing big data tools that were not mentioned previously in the literature. For this work, we based the completeness analysis on the two metrics present in the paper (simple completeness and completeness of sub-dimensions), as well as how the authors gave the relative weights for each indicator, giving us the basis to create our weight assignment.

Gavrilis et al. (2015) proposes a robust multidimensional metadata quality evaluation model that measures metadata quality based on five metrics and by taking into account contextual parameters concerning metadata generation and use. An implementation of this metadata quality evaluation model is presented and tested against a large number of real metadata records from the humanities domain and for different applications. The criteria introduced in this paper are the following: (i) Completeness, (ii) Accuracy, (iii) Consistency, (iv) Appropriateness, and (v) Auditability. This criterion is a composite measure, analyzed into three partial measures: (i) completeness of the mandatory set of elements, (ii) completeness of the 'recommended' element set and (iii) completeness of optional elements. As in the related work cited above, (GAVRILIS et al., 2015) uses a way of giving relevance to the most prominent fields, giving weight to them. The proposed approach presented in this dissertation also uses the same artifice.

2.9.2 Open Data Domain

Neumaier et al. (2016) propose a set of objective quality metrics (based on the W3C metadata schema DCAT) to monitor the quality of Open Data portals in a generic and automated way. They have introduced a generic abstraction of web-based data portals to integrate a large amount of existing data portals in an extensible manner: Based on prior metadata homogenization approaches, they have mapped the metadata occurring on the leading Open Data publishing systems(i.e., CKAN, Socrata, and OpenDataSoft) to DCAT. Based on this mapping, it was implemented and deployed an Open Data portal monitoring and quality assessment framework—the "Open Data Portal Watch" platform24—that monitors their metrics in weekly snapshots of the metadata from over 261 Open Data portals. With this work, we noticed how the authors mapped the fields of the main metadata schemes, and, similarly, we used the same technique.

Assaf et al. (2015) surveyed the landscape of various models and vocabularies that described datasets on the web. Since the key to communication is to establish a common vocabulary or model, the need for a harmonized data set metadata model with adequate data was recognized so that customers can readily comprehend and process data sets. They included four main sections in the recognized model: resources, groups, tags, and organizations. Furthermore, they classified information to be included into eight types. The main contribution is a set of mappings between each property of those models and has lead to the design of HDL, a harmonized dataset model, that takes the best out of these models and extends them to ensure complete metadata coverage to enable data discovery, exploration, and re-use. In the current work, we get inspired by the harmonization proposal to align the metadata fields that the approach employs.

Kubler et al. (2016) points out the lack of frameworks and tools to dynamically assess Open Data portal quality and compare those portals with one another. To address this lack, along with the multi-criteria decision-making nature of the comparison process, the research applies the Analytic Hierarchy Process technique. The methodology of this technique is turned into an Open Data Portal Quality Web dashboard, that enables any Open Data stakeholder to identify, at any point in time, the quality and ranking of one or a group of Open Data portals. A use case, which monitored 146 CKAN portals (and over 900K datasets), is presented showing how end-user preferences can be taken into consideration in the AHP-based comparison process. This study evaluates five dimensions of quality, whose completeness, which is studied in the current work, is between them. They based this quality measure on the CKAN metadata schemas. The current study

is inspired by the alignment of several metadata schemes, focusing on the metrics of metadata completeness.

The present work creates a new approach related to filling the data present in the dataset metadata of the open data portals. The approach is based on the harmonization of the fields of the main metadata schemes, giving weight to the fields that occur most among the schemas. Thus, creating a form to measure the completeness quality of the metadata. This approach is presented in Chapter 3.

Chapter 3

Proposed Approach

As seen in Chapters 1 and 2, the metric of completeness when evaluating the quality of metadata is vital in several respects in an open data portal. Hence, we can see that an approach that succeeds in demonstrating concisely, and that can produce satisfactory results, is necessary. According to Chapter 2, several works were done in this area, which this dissertation was based.

Therefore, This chapter elucidates an approach which has bases on an alignment of metadata schemas, which evaluate the degree of completeness of dataset metadata in comparison to the new schema. For this, it was associated weights for each field, according to which field appeared more across schemas, creating levels of importance for more relevant fields and generating metrics of completeness of the metadata. Finally, the metadata to be analyzed are aligned with the new metadata schema and their weights, presenting a metadata completeness assessment report.

3.1 Overview

The approach presented in this dissertation aims to help the open data portals administrators to measure how complete the metadata is present concerning any set of metadata schemes. This same approach is not limited to the open data domain and can be applied to any set of metadata. The current work, however, focuses on the open domain of data. We present a scenario to create an efficient parallel in order to exemplify the goals of the approach better.

One country, which we called 'Omega', plans to launch its open data portal on the internet. This portal would disseminate data and information that is generated by the

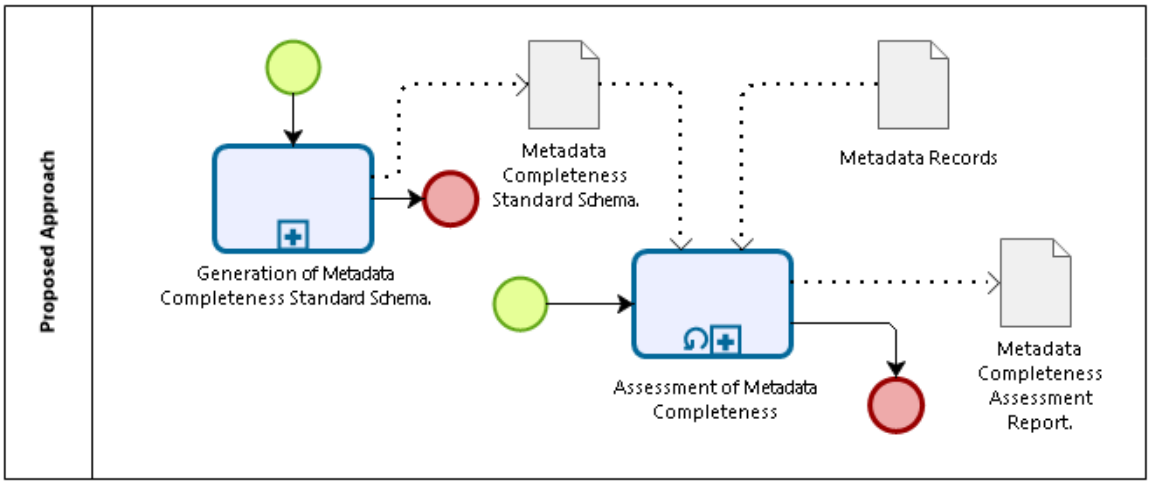

Figure 3.1: Overview of The Proposed Approach

government and its public administration and related services that are of interest to its population in order to increase its degree of transparency. The portal is managed by an administrator, who must choose the best framework for the access of its users. The administrator must consider several features and options between frameworks. One of these characteristics concerns the existing metadata of its datasets. It needs to know how well they are concerning the main schemas existing in the frameworks most used in open data portals. For this, the administrator could use the approach that is present in this work, which is explained next.

The proposed approach in the current work is divided into two sub-processes, as shown in Figure 3.1.

The first sub-process focuses on the generation of a metadata completeness standard schema. In the end, this process generates a metadata completeness standard schema, which will assist in evaluating any load of metadata records.

The second sub-process focuses on the assessment of metadata completeness in the open data portal, which uses the metadata completeness standard schema created in the first sub-process to evaluate any number of metadata records from any open data portals.

3.2 Generation of Metadata Completeness Standard Schema

The first sub-process is divided into three sub-processes, as seen in Figure 3.2. In Subsection 3.2.1, we explain the metadata schema selection sub-process, which covers the beginning of the approach. In this step, we select the schemes.

In Subsection 3.2.2, we explain the Field Alignment sub-process, in which it draws a parallel between the fields of each metadata schema, aligning them coherently and logically.

In Subsection 3.2.3 explains the sub-process of Field weighting, where we give weight to fields about greater relevance based on the relation of occurrence of the same in each metadata schema. In the end, the sub-process generates a metadata completeness standard schema.

3.2.1 Metadata Schema Selection

Figure 3.3 shows the first sub-process of the proposed approach. At this stage, it must be taken into account the objectives of using this approach. If the goal is to analyze whether the metadata filled adheres to similar metadata schemas, the choice of schemas should be made based on schemas with fields that are equivalent to the schema fields to be compared. If the objective is to adhere to the most significant number of existing portals, we should perform a search for the most commonly used open data portal schemas. The approach provides liberty to choose the amount and schemes.

According to the scenario previously described in Subsection 3.1, the Omega country open data portal administrator decided to choose three of the most widely used frameworks for publishing data that are open around the world. The administrator does this to determine if the metadata records are well-filled with the chosen schemas, in order to choose a framework that best fits the situation or even better fill the metadata in order to improve the user experience. The administrator could also decide to choose schemes whose central feature interoperability, in case of the open data portal the country Omega was a centralizing dataset from various districts.

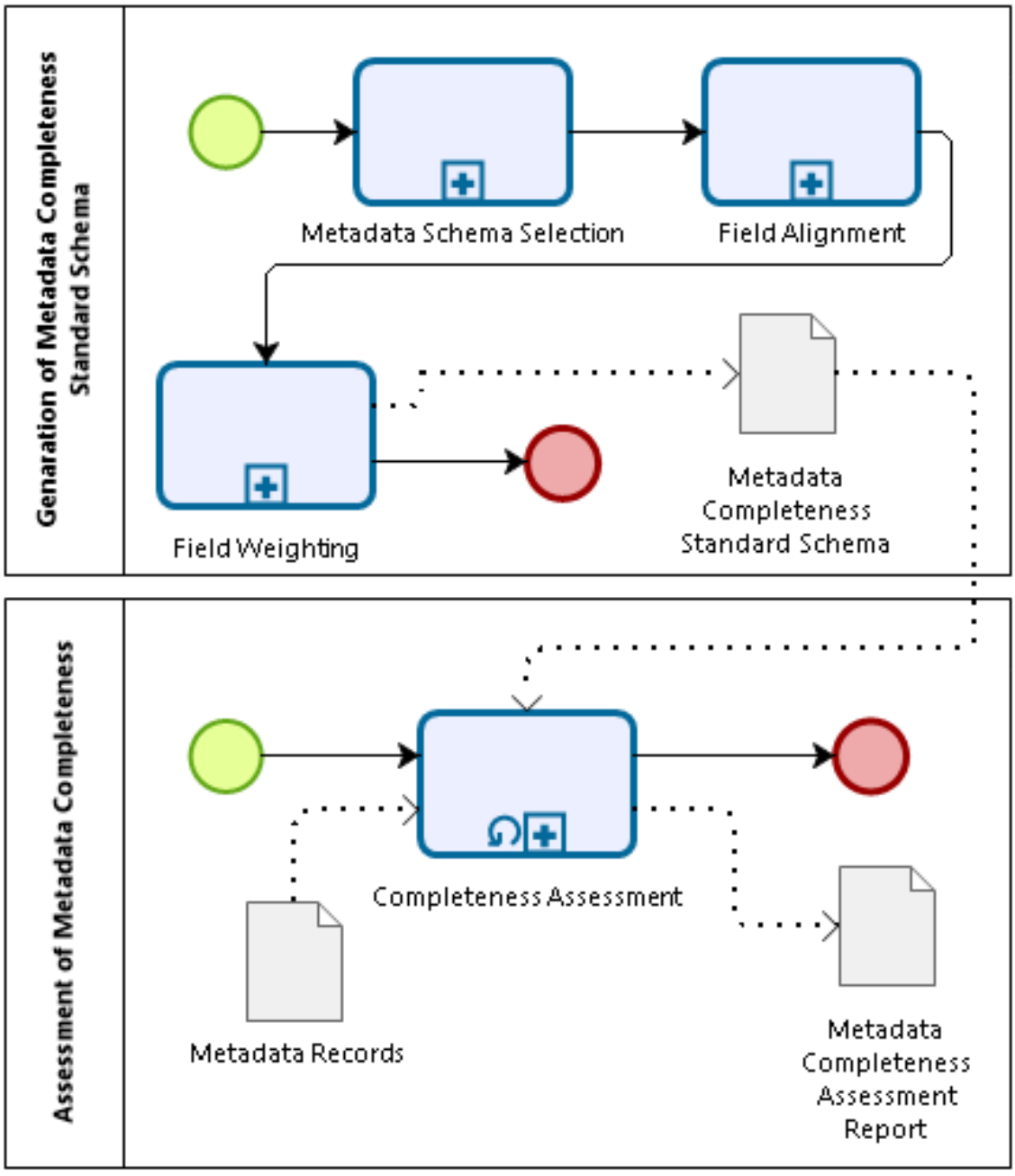

Figure 3.2: Second level of the proposed approach

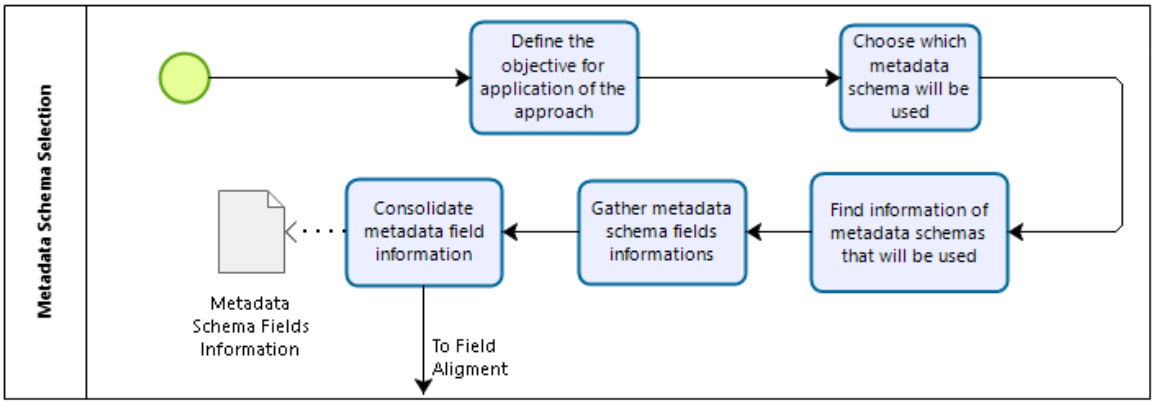

Figure 3.3: Metadata Schema Selection

After defining the schemas used in the approach, it is needed to get the information of the metadata schemas. It is recommended to collect information about the metadata field in each schema from official websites. With the schema information, it is necessary to gather the metadata elements, which play an essential role throughout the sub-process. Finally, all the information is consolidated clearly, in order to feed the next stage. After that, the approach moves on to the next stage.

We also create the heuristic that describes the sub-process, described in Algorithm 1. First, the objective is defined (Line 1) and based on it are chosen which schemas will be used in a list(Line 2). With the list of schemas defined, for each scheme, the algorithm searches for the location of the information of the same (Line 4). When found, each information is gathered in a list (Line 6) and finally formatted in the best way to be analyzed (Line 9).

Algorithm 1: MetadataSchemaSelection

Result: Metadata Schema Fields Information[ ]

```
1 Define The Objective For Application of the Approach;
2 SetOfChosenSchemas[ ] = Choose Which Metadata Schema Will Be Used;
3 forall Schema in SetOfChosenSchemas do
4 |   Find Information;
5 |   while Metadata Schema Fields Informations not gathered do
6 |   |   MetadataFieldInformation[ ] = Get Metadata Schema Field Information;
7 |   end
8 end
9 Return Format(MetadataFieldInformation);
```

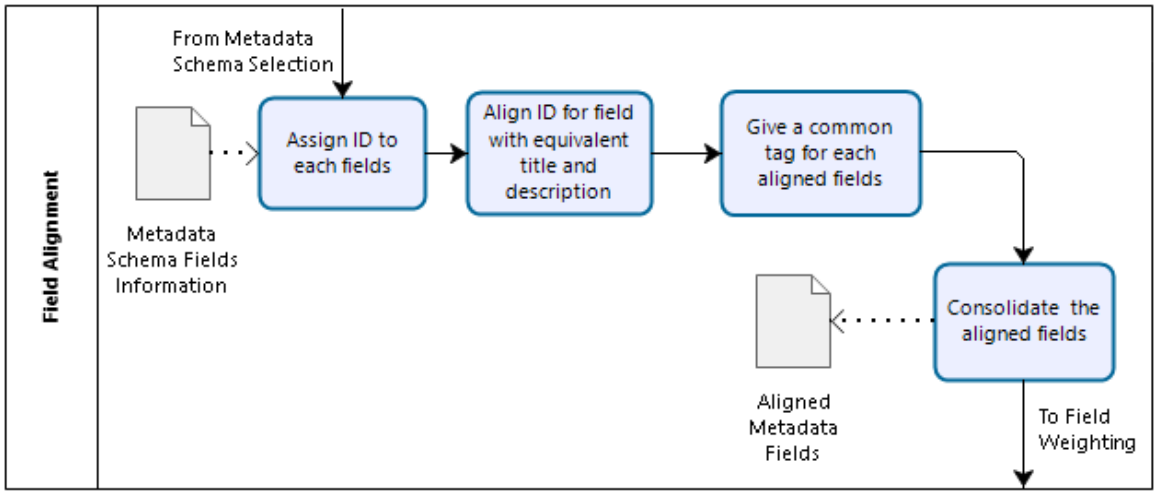

Figure 3.4: Field Alignment

3.2.2 Field Alignment

After consolidate the metadata field information of each framework, the next step is align the fields, as can be seen in Figure 3.4.

Accordingly, we analyze each field concerning the title and its descriptions. It is assigned an ID (identification) to each field. Then, IDs for fields with similar title and description are aligned. We evaluate the titles and descriptions of each field in order to create a match between them and to align repeated or equivalent fields between the schemes. If one field is related to the other, it is assigned the same ID for both fields. So, it is given a common tag for each aligned field, representing them clearly and logically. Subsequently, all the aligned fields are consolidated clearly, in order to feed the next stage. The consolidation of these fields can be done in any way, either in a table or even in JSON format, as long as it is made in order to facilitate the analysis of this data.

The heuristic described by Algorithm 2 specifies the Field Alignment sub-process. The algorithm receives as a parameter the list of Metadata Schema Fields Information. The algorithm then crosses the list by analyzing whether the field already has an ID (Line 2) and if it does not, assigns one to it (Line 3). Then the list is researched for fields similar to the previous one, and if it finds, the field IDs are aligned (Line 9). Finally, after all the fields have an assigned ID, a Tag is associated with each ID that best describes the fields to that associated (Line 14). Thus, the algorithm returns a consolidated list of aligned fields formatted in a way that best describes them (Line 16).

Algorithm 2: FieldAlignment(Metadata Schema Fields Information[ ])

```
Result: Aligned Metadata Fields[ ]
foreach CurrentField in Metadata Schema Fields Information do
    if CurrentField Has Not An ID then
        Assign ID;
    else
        Next;
    end
    foreach CompareField in Metadata Schema Fields Information do
        if CurrentField and CompareField are Equivalent Title and Description
         then
            Align ID's;
        end
    end
end
while All Aligned Fields do not have a tag do
    Give a logical Common Tag;
end
Return Format(Consolidate the Aligned Fields[ ]);
```

3.2.3 Field Weighting

After aligning the metadata fields of each framework, the next step is to weight the fields, with purpose decide which fields are most prominent among frameworks and have the most need to be filled, as can be seen in Figure 3.5.

After aligning the metadata fields, some of them appear in more than one schema. Therefore, these fields may have more importance in some aspects than others. Thus, to provide further relevance to the most occurring fields, it is necessary to weight each field differently. Hence, the occurrence of each field in each tag is counted and consolidate them. These occurrences can be consolidated in any manner, either in a table or even in JSON format, as long as it is taken to promote the evaluation of this information.

$$TagWei = FieldOccurrence\Big(\frac{100}{TotalOfOccurrence}\Big) \tag{3.1}$$

Afterward, we use a weighted average between the fields for the weighting calculation of each tag. As seen in equation 3.1, we calculate the weight of each tag (TagWei) by the number of occurrences of the fields in the tag (FieldOccurrence) times one hundred divided by the total sum of occurrences (TotalOfOccurrence).

Subsequent using equation 3.1 in each tag, they are grouped according to equal

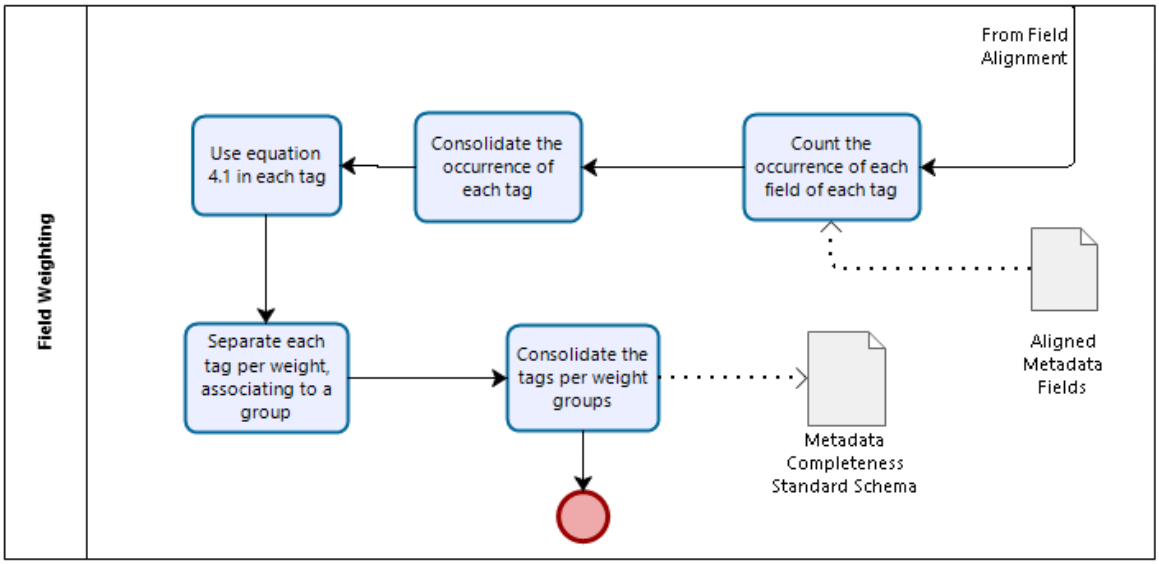

Figure 3.5: Field Weighting

weights, thus forming groups of relevance. Finally, with all the tags separated per groups of relevance consolidated, the metadata completeness standard schema is created. With this scheme, We can evaluate the metadata records concerning completeness.

We create Algorithm X to explain the Field Weighting sub-process, which receives by parameter the set of aligned metadata fields. In the beginning, the algorithm counts how many occurrences of fields are grouped in the same tag, using the same ID (Line 14), generating a list of occurrences for each tag, which is then formatted to analyze best (Line 14). With this, for each tag, the algorithm applies the equation 3.1, where the occurrence of the tag is multiplied by one hundred, and divided by the sum of all occurrences of the list (Line 20). Therefore we have a weighting of each tag, being able to perceive which of them would have higher significance concerning all the chosen schemes. Finally, the algorithm separates the tags by the weighting given to it and thus delivering a metadata completeness standard schema (Line 23).

Going back to the scenario described in Subsection 3.1, the Omega country open data portal administrator will now be able to compare any number of metadata present in the datasets to be published in the portal, concerning the schemas of the main frameworks. He could, for example, analyze all the metadata, or only the dataset metadata for the transport sector. We discuss the analysis of completeness with the schema created in Section 3.3.

Algorithm 3: FieldWeighting(Aligned Metadata Fields[ ])

```
   Result: Metadata Completeness Standard Schema[ ]
 1 Count = 1;
 2 Index = 1;
 3 SortAscending(Aligned Metadata Fields);
 4 foreach CurrentID in Aligned Metadata Fields do
 5 |   if ComparedID is Null then
 6 |   |   ComparedID = CurrentID;
 7 |   end
 8 |   if ComparedID = CurrentID then
 9 |   |   Count = Count + 1;
10 |   |   Next;
11 |   else
12 |   |   ComparedID = CurrentID;
13 |   end
14 |   Occurrence of Aligned Metadata Fields[ Index ] = Count;
15 |   Count = 1;
16 |   Index = Index + 1;
17 end
18 Format(Consolidate the Occurrence of each Tag[ ]);
19 foreach Consolidated Occurrence of Tag do
20 |   TagWei = Field Occurrence * 100 / Sum of Occurrence;
21 end
22 Separate Groups of Tag per weight;
23 Return Format(Consolidate the Tags per Weight Groups[ ]);
```

3.3 Assessment of Metadata Completeness

Figure 3.2 show the second sub-process. We explain the sub-process of completeness assessment, the final stage of the proposed approach in Subsection 3.3.1 where it is analyzed the adherence of the records of metadata datasets to the metadata completeness standard schema created in the 3.2.3 section.

3.3.1 Completeness Assessment

Now, after the metadata completeness standard schema has been created, according to Section 3.2, any metadata records can be evaluated to its completeness. Figure 3.6 describes this sub-process. First, for evaluation, it is necessary to gather all the metadata records. The formats in which they meet can be varied. Both the metadata gathering and the extraction of its content can be done in the best way, either through programming, through a script, through software or even manually.

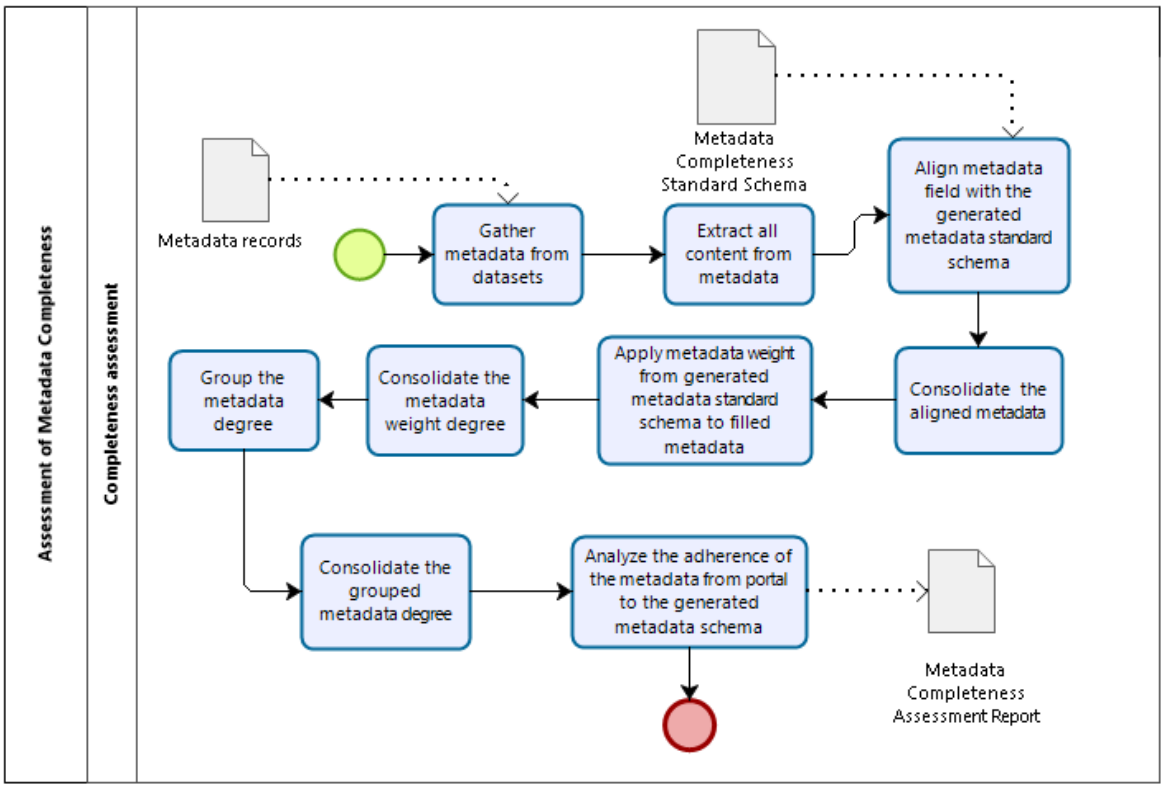

Figure 3.6: Completeness Assessment

Thereby, the metadata fields get aligned with the generated metadata completeness schema fields from Section 3.2. For this, the titles and contents are assessed in order to

check if there are equivalent fields. After alignment, we consolidate this information in the best-chosen format.

With fields aligned, it is assigned the same weights as the metadata completeness standard scheme, generated in Section 3.2, for aligning the fields. For non-aligned fields, we specify zero weight. Thereat, we consolidate the weight of metadata in the best-chosen format.The fields are then grouped according to their weight, to assess the essential group of fields and to consolidate this data in the best format selected.

Lastly, we assess the metadata records fields filled in by applying the weights aligned with each of them. The total weight of the metadata record is given by adding all the weights of the filled fields in order to weight each metadata record. After that, we consolidate the evaluation. The scores for each metadata record then group the results.

Finally, the metadata records adherence is analyzed concerning the standard metadata completeness scheme, resulting in a report in which one realizes how complete they are to other metadata schemas.

The last algorithm (Algorithm 4) created represents the evaluation of the metadata records compared to the metadata completeness standard schema, where it receives both as a parameter. First, all fields and their information are assembled from each record metadata in a list by the algorithm (Line 1). Thus, the algorithm compares each field present in the metadata completeness standard schema with the field list of metadata records, taking into account title, description, and content, and then aligns similar fields (Line 5). Hence, the list of aligned metadata record field with metadata completeness standard schema fields is formatted for better analysis (Line 9).

Accordingly, now the algorithm can weigh each field about its completeness. For this, the algorithm runs through each metadata record and checks whether the field is filled or not. If it filled, the weight relative to the metadata completeness standard schema aligned is applied (Line 13). After that, we have a consolidated list of weighted metadata, which is formatted (Line 17). Then the algorithm separates, consolidates and formats each record metadata by weights (Line 18 and Line 19). Thus, finally, we have a report, in which we can analyze the completeness of metadata records with the chosen schemas (Line 20).

Figure 3.7 shows the proposed approach as a whole, showing everything that chapter 3 presented.

Algorithm 4: AssessmentOfMetadataCompleteness(Metadata Completeness Standard Schema, Metadata Records)

```
   Result: Metadata Completeness Assessment Report
 1 MetadataInfo[ ] = Gather all Metadata from Datasets(Metadata Records);
 2 foreach StandardMetadataField in Metadata Completeness Standard Schema do
 3 |   foreach RecordMetadataField in MetadataInfo[ ] do
 4 |   |   if CompareTitleDescriptionAndContent(StandardMetadataField,
   |   |    RecordMetadataField) is True then
 5 |   |   |   Align Field;
 6 |   |   end
 7 |   end
 8 end
 9 Format(Aligned Metadata Record Field with Metadata Completeness Standard
    Schema Fields[ ]);
10 foreach Metadata Records do
11 |   foreach MetadataContent do
12 |   |   if MetadataContent is not Null then
13 |   |   |   Apply Weight;
14 |   |   end
15 |   end
16 end
17 Format(Consolidate Metadata Weight Degree[ ]);
18 Separate Groups of Metadata Degree;
19 Format(Consolidate Metadata Degree[ ]);
20 Return Metadata Completeness Assessment Report[ ];
```

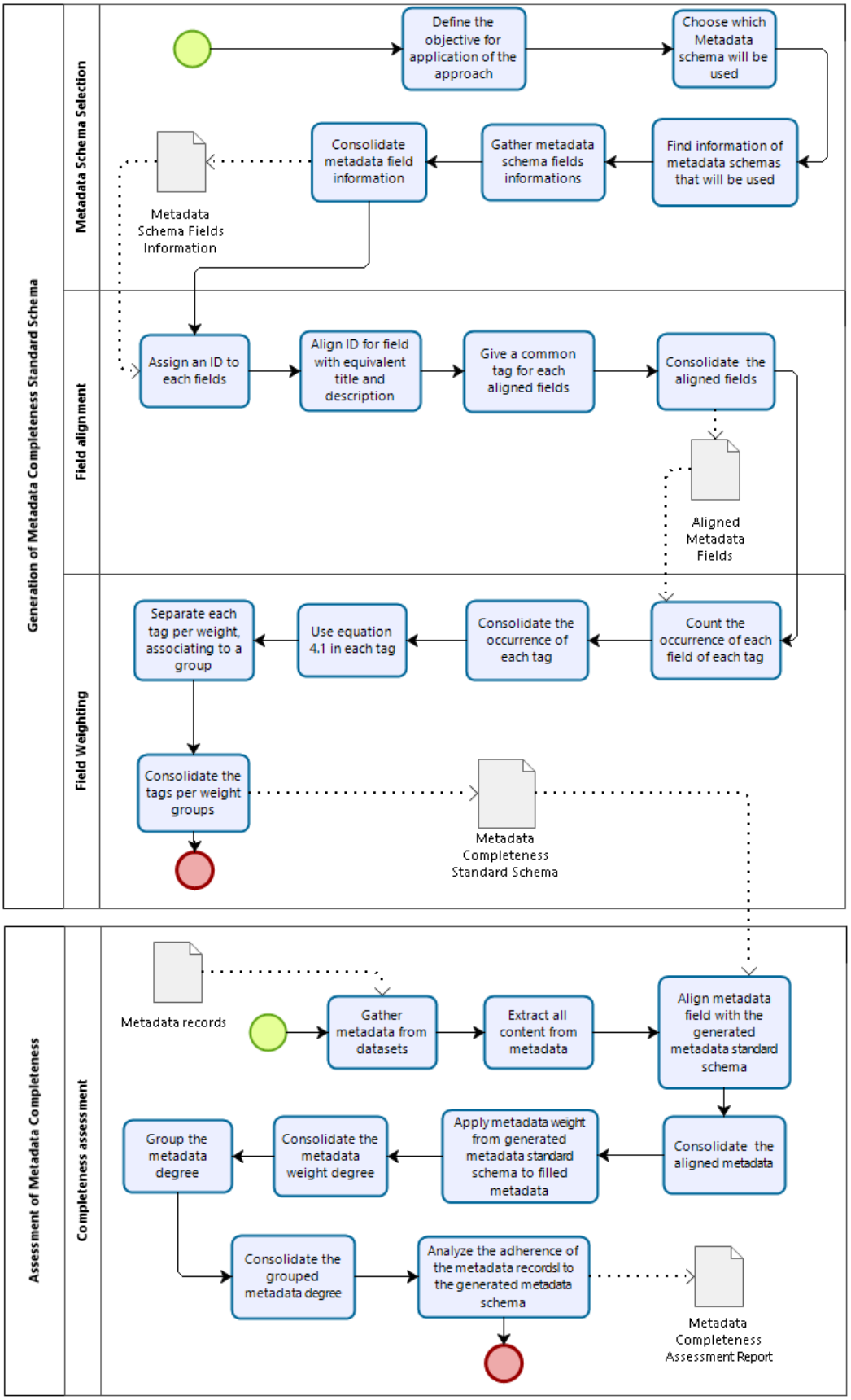

Figure 3.7: Proposed Approach

Chapter 4

Case Studies

In this chapter, we evaluate the proposed approach by creating two case studies. The first case consisted of aligning three metadata schemes selected from the most commonly used frameworks to create open data portals around the world, in this case, as stated by the authors (NEUMAIER et al., 2016), Socrata, CKAN, and OpenDataSoft, to create a standard metadata completeness scheme that we call 'Alpha'. After the creation of the Alpha schema, we evaluated the metadata datasets from the Europa portal and the NYS portal concerning the Alpha schema. The second case consisted of aligning the DCAT-AP scheme with the three systems mentioned in the earlier case. DCAT-AP was developed to describe data sets in the European public sector. After the alignment, we created a metadata completeness standard schema, which we call 'Beta'. After the creation of the Beta schema, we also evaluated the metadata datasets from the Europa portal and the NYS portal, but now concerning the Beta schema.

4.1 Case 1

For the application of the proposed approach, we followed the steps presented in Chapter 3. As seen in Subsection 3.2.1, first, it is necessary to define the objective for the application of the approach, which we choose for assess for adherence to the most significant number of existing portal metadata. The next step is to decide which metadata schemas to use, so we select the three main frameworks of open data (CKAN, Socrata, and Opendatasoft). Hence, to find information of metadata schemas that will be used and gather metadata schema fields information, we searched for the metadata of each framework on the official sites CKAN (2017), Socrata (2017), Opendatasoft (2018). With this, we consolidate all fields in Appendix A.1. In this work, we identified this first metadata scheme generated

by the approach as ‘Alpha’.

After consolidated metadata fields information, which we consolidated in a table format, the next step was to align the fields, as seen in Subsection 3.2.2. For this, we evaluated each field in terms of title and description using Table A.1. We assigned IDs to each field, and equivalent fields received the same ID. In some cases, we grouped more than one field from the same schema into the same ID, because they were classified as possible sublevels of the same field (e.g. ‘License’ and ‘Rights’ fields from Socrata metadata schema). For each ID, we assigned a tag in a logical manner, which best expressed the aligned fields. For Alpha schema, Table 4.1 shows the result.

Table 4.1: Alpha Metadata Fields Mapping

ID	Tag	Ckan Field	Socrata Field	Opendatasoft Field
1	Title	Title	Title	Title
2	Description	Description	Description	Description
3	Tags	Tags	Tags	Keywords
4	Last Updated	-	Last Updated	Last modification
5	Publisher	-	-	Publisher
6	Contact Email	-	Contact Email	-
7	Unique identifier	Unique identifier	Unique Identifer	Identifier
8	Public Access Level	-	Public Access Level	-
9	Agency / Department	-	Agency / Department	-
10	License	License	License / Rights	License
11	Spatial Geographical Area	-	Geographic Unit	Geographic area
12	Temporal Coverage	-	Temporal Coverage	-
13	Data Dictionary	-	Data Dictionary	-
14	Language	-	-	Language
15	Permalink / Identifier	-	Permalink/Identifier	-
16	Related Documents	-	Related Documents	References

17	Theme / Category / Groups	Groups	Category	Theme
18	API key	API key	API Endpoint	-
19	Frequency	-	Frequency of Data Change | Frequency of Publishing	-
20	Format	Multiple formats (if provided)	-	-
21	Public Access Level Comment	-	Public Access Level Comment	-
22	Data preview	Data preview	-	-
23	Revision history	Revision history	-	-
24	Extra fields	Extra fields	-	-
25	Data Steward	-	Data Steward	-
26	Row Count	-	Row Count	-
27	Download URL	-	Download URL	-
28	Source	-	Link	Attributions
29	Timezone	-	-	Timezone

Comparing the created tags with each schema, the number of fields increased. Compared to the CKAN Schema, there was a 163,64% increase in the number of fields (18 more fields). In comparison with the Socrata Schema, there was a 31,82% increase in the number of fields (7 more fields). Compared to the Opendatasoft Schema, there was a 152,63% increase in the number of fields (10 more fields). As can be seen, with this approach, we noticed that the most dominant scheme might be what has more metadata fields.

According to Section 3.2.3, after consolidating all the tags, the occurrence of each field present in each tag was counted, as can be seen in the results in Figure 4.1.

When counting the number of fields in tags, we noticed that in Table 4.1 the field 'Frequency' is formed by the union of two fields of Socrata schema (Frequency of Data Change | Frequency of Publishing). In this case, we counted the occurrence only be counted one time, because even if it occurred in two different fields, they are from the same schema.

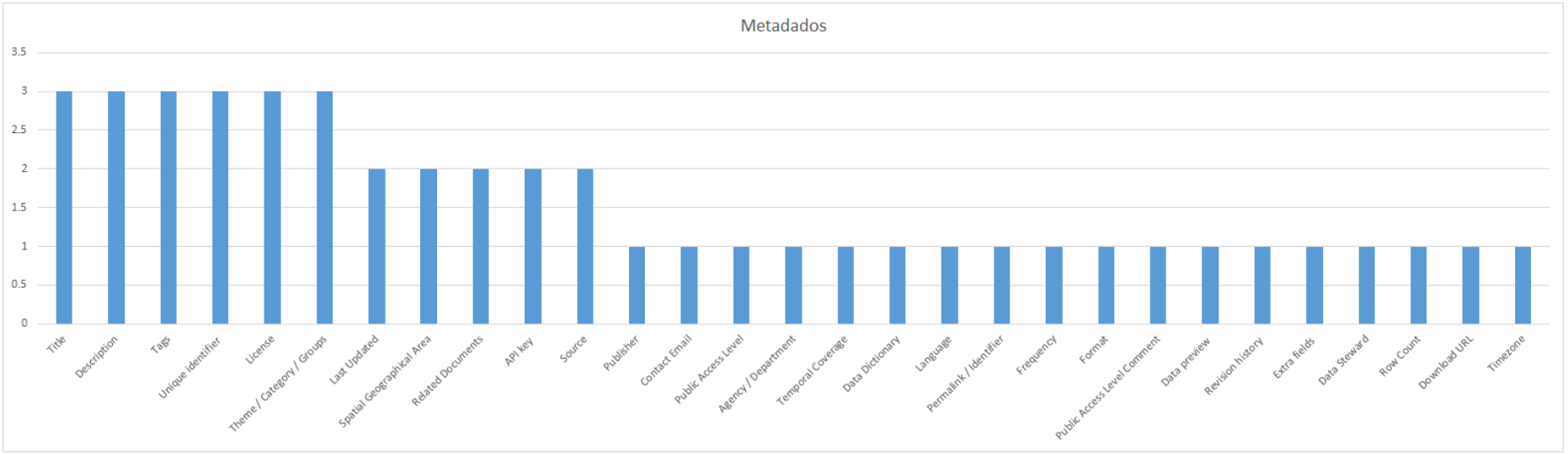

Figure 4.1: Alpha Metadata Field Occurrence Graph

For the weight of each field calculation, we applied the Equation 3.1. Then, we separated each result per groups and consolidated in Table 4.2.

Table 4.2: Alpha Metadata Weight in Percents

Alpha Metadata field	Weight	Groups
Title	6.52%	Group 1
Description	6.52%	Group 1
Tags	6.52%	Group 1
Unique identifier	6.52%	Group 1
License	6.52%	Group 1
Theme / Category / Groups	6.52%	Group 1
Last Updated	4.35%	Group 2
Spatial Geographical Area	4.35%	Group 2
Related Documents	4.35%	Group 2
API key	4.35%	Group 2
Source	4.35%	Group 2
Publisher	2.17%	Group 3
Contact Email	2.17%	Group 3
Public Access Level	2.17%	Group 3
Agency / Department	2.17%	Group 3
Temporal Coverage	2.17%	Group 3
Data Dictionary	2.17%	Group 3
Language	2.17%	Group 3
Permalink / Identifier	2.17%	Group 3
Frequency	2.17%	Group 3
Format	2.17%	Group 3
Public Access Level Comment	2.17%	Group 3
Data preview	2.17%	Group 3
Revision history	2.17%	Group 3
Extra fields	2.17%	Group 3
Data Steward	2.17%	Group 3
Row Count	2.17%	Group 3
Download URL	2.17%	Group 3
Timezone	2.17%	Group 3

From this analysis, we recognized that there are three distinct metadata groups. We sorted these groups according to the number of times they appear in the metadata schemas. In Figure 4.2, we can see the amount of weight of each group.

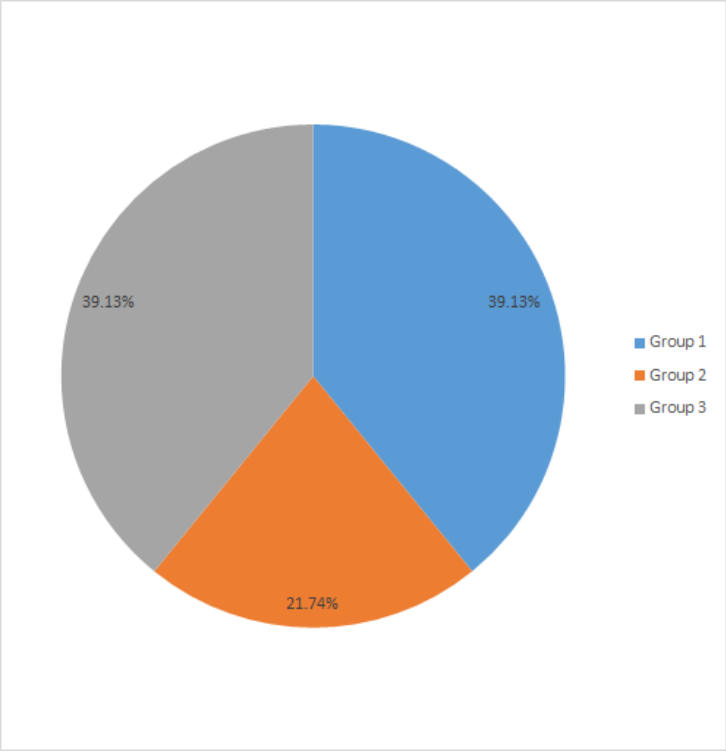

Figure 4.2: Alpha Metadata Weight in Percentual per Groups

The group of metadata fields that appears in the three schemas (Group 1), thus being more relevant, amounts to 39.13% of the total weight (Title, Description, Tags, Unique identifier, License, Theme / Category / Groups). These fields can be classified as primary information in the identification of a dataset. This supports the approach because it gives importance to the fields that have more prominence.

The group of metadata fields that appears in the three schemas (Group 1), thus being more relevant, amounts to 39.13% of the total weight (Title, Description, Tags, Unique identifier, License, Theme / Category / Groups). We classified these fields as primary information in the identification of a dataset. This classification promotes the approach as it provides more importance to the fields.

The group of metadata fields that appear in only two schemas (Group 2) amounts to 21.74% of the total weight (Last Updated, Spatial Geographical Area, Related Documents, API key, Source). These fields are essential, but not enough to identify the dataset that ordinary users seek.

The last group of metadata fields that appear in only 1 schema (Group 3), amounts to 39.13% of the total weight (Publisher, Contact Email, Public Access Level, Agency / Department, Temporal Coverage, Data Dictionary, Language, Permalink / Identifier, Frequency, Format, Public Access Level Comment, Data preview, Revision history, Extra fields, Data Steward, Row Count, Download URL (Uniform Resource Locator), Timezone). Most of these fields add more detail to metadata and are very relevant to solution developers, which demonstrates that much of the metadata does not cover all the information that is needed when creating tools for various industries. Besides, we can observe that the interoperability between frameworks is not yet well established concerning metadata. With this, the Alpha metadata completeness standard schema was created, being able to evaluate any metadata records.

In order to evaluate the proposed approach, we applied the metadata completeness standard schema Alpha to metadata records from open data portals that are well recognized and have relevant metadata volume so that could be analyzed statistically how the fields were being filled about them. So, we searched for a dataset of metadata of datasets. The European Union Open Data Portal that gathers datasets from various countries of the European Union (PORTAL, 2016) has a dataset that contains 12642 metadata assemblies from various datasets. The portal has as characteristic the meeting of several European portals, consolidating diverse datasets in a single repository. For this, the site has created a metadata schema of its (COMMISSION, 2018), to be able to gather and link them. The dataset was divided into 13 parts, in the JSON format.

To analyze the entire data load, we used a script to automate the process. Next, as seen in Section 3.3.1, all fields were aligned with each field in Alpha schema. For this, we filtered all fields that occurred in all datasets. To align the fields, we analyzed the titles and description of the fields of the dataset of metadata about the Alpha schema fields. Also, we analyzed the content of the dataset of metadata punctually, in order to to have a more accurate alignment.

It was noticed that some of the fields appeared in more than one dataset field. This was the cases of the ‘Frequency’, ‘License’, ‘Theme/Category/Groups’ and ‘Temporal Coverage’ fields. For ‘Frequency’ there are two fields: ‘temporal_granularity’ and ‘accrual_periodicity’. For ‘License’ there are three fields: ‘license_id’, ‘license_title’ and ‘license_url’. For ‘Theme/Category/Groups’ there are two fields: ‘concepts_eurovoc’ and ‘groups’. For ‘Temporal Coverage’ there are two fields: ‘temporal_coverage_to’ and ‘temporal_coverage_from’. For these fields, the weight will be proportionally divided

according to the number of correlated fields.

We could not align some fields, such as 'Data preview', 'Public Access Level Comment', 'Row Count' and 'Timezone'. That is due to the title and description did not match with any fields of the Alpha metadata schema. With this, we can observe that no dataset could cover 100% of Alpha schema, since the sum of the weights of the cited fields reaches 8.70%, causing them to reach a maximum of 91.3%. Table 4.3 shows the metadata fields correspondences and their respective weights.

Table 4.3: Alpha Metadata Fields Correlation with European Portal Metadata Fields

Alpha Metadata Title	Europe Dataset Metadata Title	Weight
Description	Key: description	6.52%
Tags	Key: keywords	6.52%
Title	Key: title	6.52%
Unique identifier	Key: identifier	6.52%
API key	Key: owner_org	4.35%
Theme/Category/Groups	Key: concepts_eurovoc	3.26%
Theme/Category/Groups	Key: groups	3.26%
Last Updated	Key: modified_date	4.35%
Source	Key: resources(link)	4.35%
Related Documents	Key: resources	4.35%
Spatial Geographical Area	Key: geographical_coverage	4.35%
Contact Email	Key: contact_email	2.17%
Extra fields	Key: extras	2.17%
Agency/Department	Key: organization	2.17%
Data Dictionary	Key: resources(name)	2.17%
Data preview	-	2.17%
Download URL	Key: resources(download)	2.17%
Frequency	Key: temporal_granularity	1.09%
Frequency	Key: accrual_periodicity	1.09%
Permalink/Identifier	Key: url	2.17%
Language	Key: language	2.17%
Format	Key: resources(format)	2.17%
Public Access Level	Key: private	2.17%
Public Access Level Comment	-	2.17%
Revision history	Key: revision_timestamp	2.17%

Row Count	-	2.17%
Timezone	-	2.17%
License	Key: license_id	2.17%
License	Key: license_title	2.17%
License	Key: license_url	2.17%
Data Steward	Key: maintainer	2.17%
Publisher	Key: author	2.17%
Temporal Coverage	Key: temporal_coverage_to	1.09%
Temporal Coverage	Key: temporal_coverage_from	1.09%
Notcorrelated	Key: contact_name	0.00%
Notcorrelated	Key: contact_webpage	0.00%
Notcorrelated	Key: contact_address	0.00%
Notcorrelated	Key: alternative_title	0.00%
Notcorrelated	Key: capacity	0.00%
Notcorrelated	Key: contact_telephone	0.00%
Notcorrelated	Key: creator_user_id	0.00%
Notcorrelated	Key: interoperability_level	0.00%
Notcorrelated	Key: isopen	0.00%
Notcorrelated	Key: metadata_created	0.00%
Notcorrelated	Key: metadata_language	0.00%
Notcorrelated	Key: metadata_modified	0.00%
Notcorrelated	Key: name	0.00%
Notcorrelated	Key: num_resources	0.00%
Notcorrelated	Key: num_tags	0.00%
Notcorrelated	Key: owner_org	0.00%
Notcorrelated	Key: rdf	0.00%
Notcorrelated	Key: relationships_as_object	0.00%
Notcorrelated	Key: relationships_as_subject	0.00%
Notcorrelated	Key: release_date	0.00%
Notcorrelated	Key: revision_id	0.00%
Notcorrelated	Key: state	0.00%
Notcorrelated	Key: status	0.00%
Notcorrelated	Key: tracking_summary	0.00%
Notcorrelated	Key: type	0.00%

Notcorrelated	Key: type_of_dataset	0.00%
Notcorrelated	Key: version	0.00%
Notcorrelated	Key: version_description	0.00%
Notcorrelated	Key: maintainer_email	0.00%
Notcorrelated	Key: author_email	0.00%

After consolidating the alignment, we applied the weights to each record metadata and thus having the degree of completeness each of them. We can see the results in Figure 4.3.

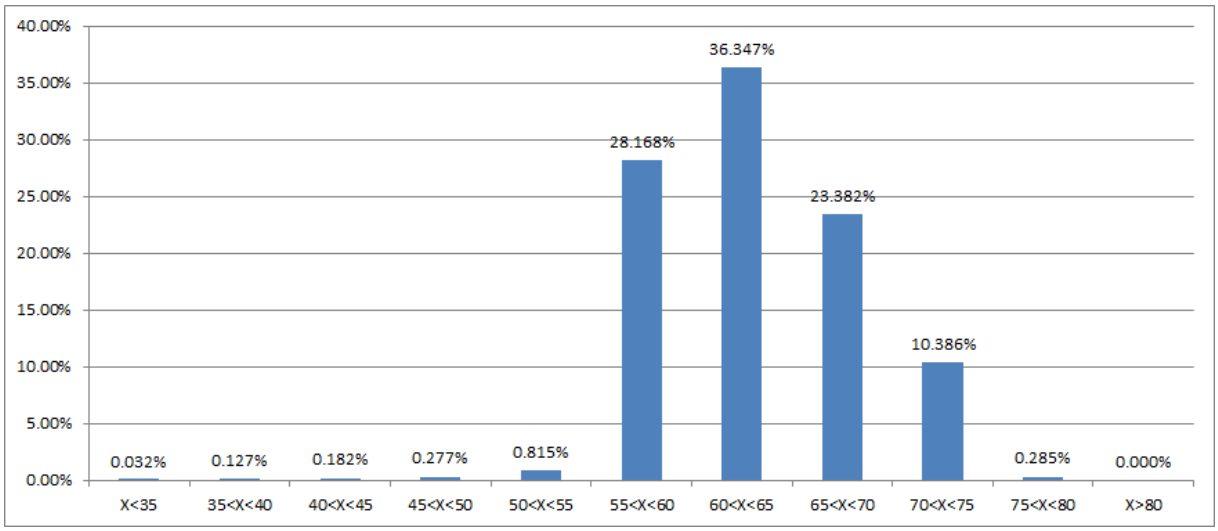

Figure 4.3: Europe Metadata Dataset Weight in Percentual Ranges of Alpha Schema

As we can see, most datasets have more than 55% weighting compared to Alpha metadata schema. Most datasets are in the range of 60 to 65 percent of the weights. It only has 10.67% of the datasets reaching more than 70 percent of weight. At least half of the main fields are filled, but the result is far from expected since they do not have any dataset with more than 80% of weighting.

In addition to presenting the metadata records degree of completeness, this assessment makes it possible to evaluate the amount of metadata filled in each metadata field. We can see it in Figure 4.4.

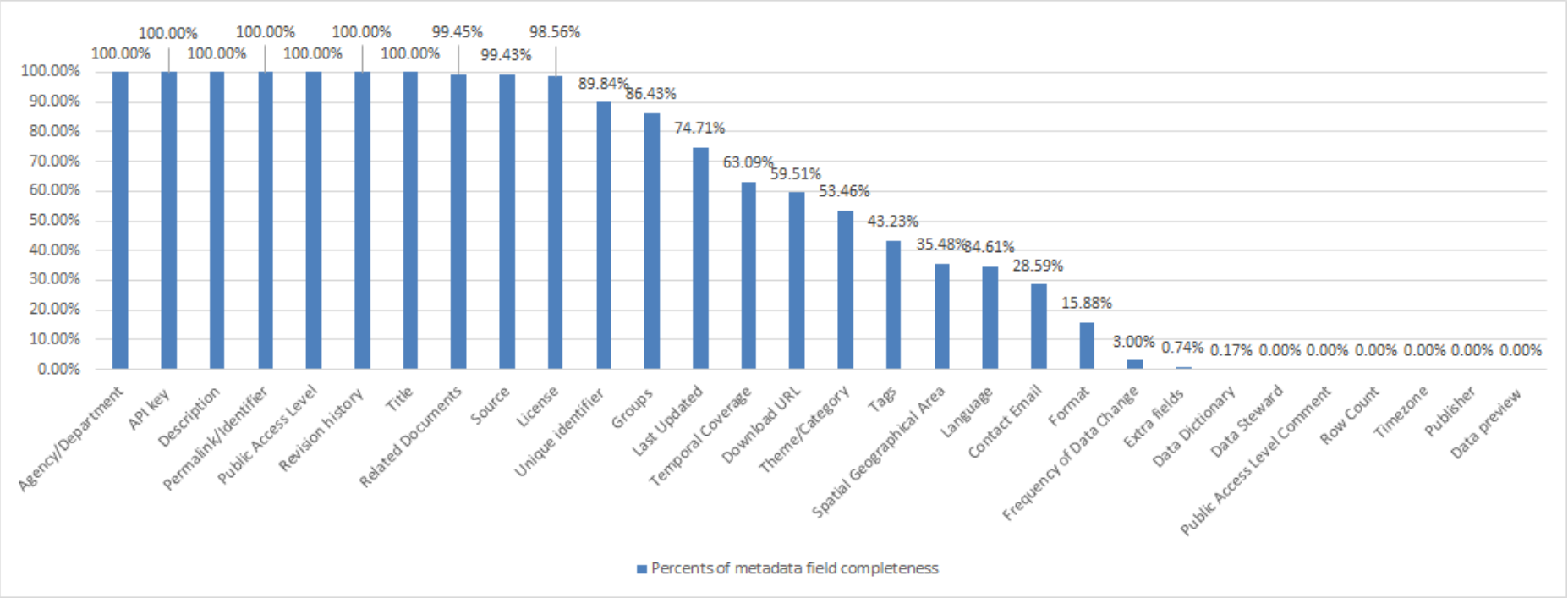

Figure 4.4: Percents of Europe Metadata Dataset Field Completeness in Comparison to Alpha Metadata Schema

The most filled fields were ‘Agency / Department’, ‘API key’, ‘Description’, ‘Permalink / Identifier’, ‘Public Access Level’, ‘Revision history’ and ‘Title’, having 100% of datasets with these fields filled, which is equivalent to 26.09% by weight in relation to Alpha metadata schema.

However, those that have had no occurrence are: ‘Data Steward’, ‘Public Access Level Comment’, ‘Row Count’, ‘Time zone’, ‘Publisher’ and ‘Data Preview’. They are equivalent to 13.04% of the weight about the metadata set.

Afterward, we analyzed the metadata dataset for the state of NY, which is one of the most critical states in the United States and the most significant financial and commercial center in the country(PILBEAM, 2018), and the fourth largest industrial center in the United States. We found a dataset containing 1650 metadata of datasets from the NYS open data portal (YORK, 2005), available in several formats. It was used for this work the CSV format file.

We do not need to perform the steps of creating the Alpha metadata completeness standard schema, because the schema is already created. So, we need only to apply the process described in section 3.3.1. This demonstrates the adaptability of the approach, avoiding rework in creating the completeness standard schema. Next, as seen in Section 3.3.1, all fields were aligned with each field in Alpha schema. For this, we filtered all fields that occurred in all datasets. To align the fields, we analyzed the titles and description of the fields of the dataset of metadata about the Alpha schema fields. Also, we analyzed the content of the dataset of metadata punctually, in order to to have a more accurate alignment.

We regarded that some of our fields appeared in more than one dataset field. This was the case of the ‘Spatial Geographical Area’ field. For ‘Spatial Geographical Area’ there are two fields: ‘Coverage’ and ‘Localities’. For this field, we proportionally divided the weight according to the number of correlated fields.

Other fields could not be aligned, such as ‘License’, ‘Related Documents’, ‘Data Dictionary’, ‘Data Steward’, ‘Download URL’, ‘Extra fields’, ‘Format’, ‘Language’, ‘Public Access Level’, ‘Public Access Level Comment’, ‘Revision history’, ‘Row Count’, ‘Temporal Coverage’ and ‘Timezone’. With this, we observe that no dataset could cover 100% of our set of metadata, since the sum of the weights of the cited fields reaches 36.96%, causing them to reach a maximum of 63.04%.

Table 4.4 shows the metadata fields correspondences and their respective weights.

Table 4.4: Alpha Metadata Fields Correlation with NYS Metadata Fields

Metadata Title	Dataset Metadata	Weight
Description	Description	6.52%
License	-	6.52%
Tags	Keywords	6.52%
Theme/Category/Groups	Category	6.52%
Title	Name	6.52%
Unique identifier	U ID	6.52%
API key	api_endpoint	4.35%
Last Updated	Last Update Date (data)	4.35%
Related Documents	-	4.35%
Source	Source Link	4.35%
Agency/Department	Agency	2.17%
Contact Email	Contact Information	2.17%
Data Dictionary	-	2.17%
Data preview	Derived View	2.17%
Data Steward	-	2.17%
Download URL	-	2.17%
Extra fields	-	2.17%
Format	-	2.17%
Frequency	Posting Frequency	2.17%
Language	-	2.17%
Permalink/Identifier	URL	2.17%
Public Access Level	-	2.17%
Public Access Level Comment	-	2.17%
Publisher	Data Provided By	2.17%
Revision history	-	2.17%
Row Count	-	2.17%
Spatial Geographical Area	Coverage	2.17%
Spatial Geographical Area	Localities	2.17%
Temporal Coverage	-	2.17%
Timezone	-	2.17%
Notcorrelated	Type	0.00%
Notcorrelated	Domain	0.00%

Notcorrelated	Organization	0.00%
Notcorrelated	See Also	0.00%
Notcorrelated	Granularity	0.00%
Notcorrelated	Limitations	0.00%
Notcorrelated	Notes	0.00%
Notcorrelated	Owner	0.00%
Notcorrelated	Visits	0.00%
Notcorrelated	Downloads	0.00%
Notcorrelated	Creation Date	0.00%
Notcorrelated	Parent UID	0.00%
Notcorrelated	County Filter	0.00%
Notcorrelated	County Column	0.00%
Notcorrelated	Municipality Filter	0.00%
Notcorrelated	Municipality_Column	0.00%

After the consolidated alignment, we applied the weights to each record metadata and thus having the degree of completeness each of them. We can see the results in Figure 4.5.

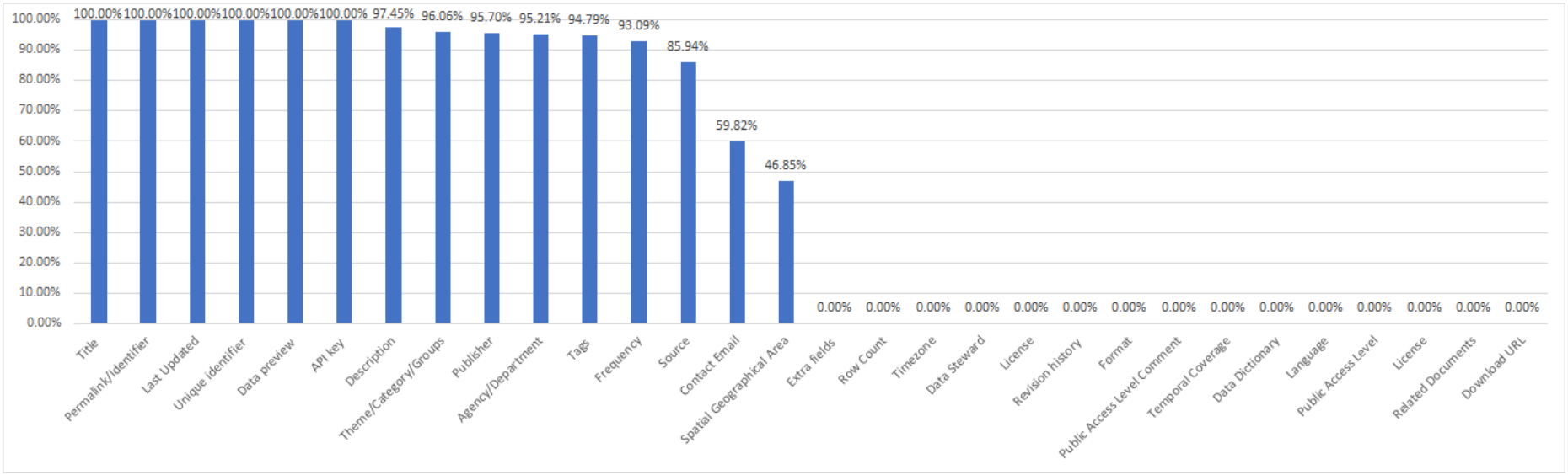

Figure 4.5: Percents of NYS Metadata Dataset Field Completeness in Comparison to Alpha Metadata Schema

As can be observed, the most filled fields were 'Title', 'Permalink/Identifier', 'Last Updated', 'Unique identifier', 'Data preview' and 'API key', having 100% of datasets with these fields filled, which is equivalent to 26,09% by weight in relation to Beta metadata schema.

However, those that have had no occurrence are: 'Extra fields', 'Row Count', 'Timezone', 'Data Steward', 'License', 'Revision history', 'Format', 'Public Access Level Comment', 'Temporal Coverage', 'Data Dictionary', 'Language', 'Public Access Level', 'License', 'Related Documents' and 'Download URL'. They are equivalent to 36,96% of the weight in relation to the metadata set.

After the consolidated alignment, we applied the weights to each record metadata and thus having the degree of completeness each of them. See Figure 4.6 for the results.

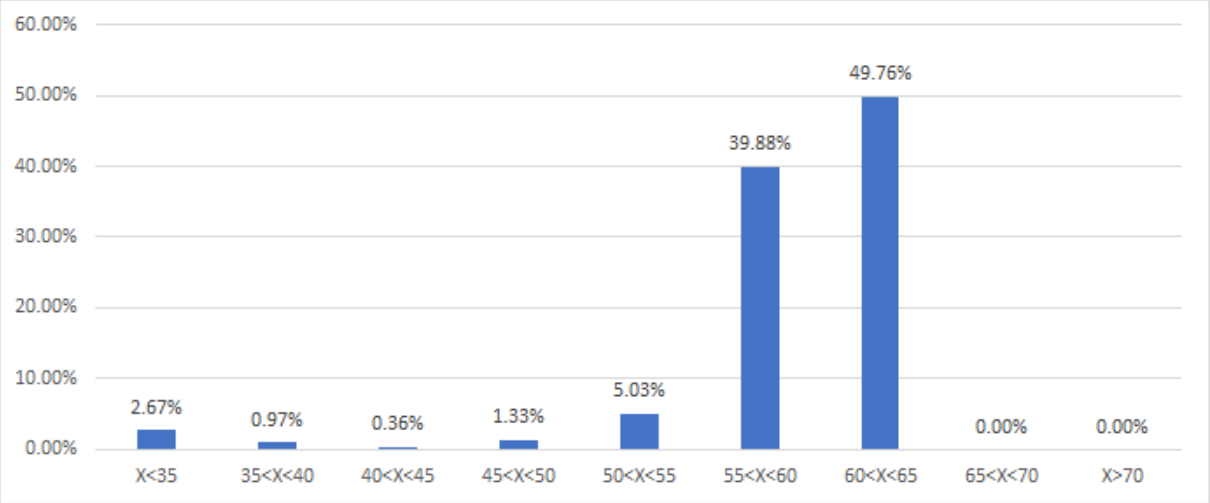

Figure 4.6: NYS Metadata Dataset Weight in Percentual Ranges of Alpha Schema

As observed, most data sets have more than 55% weighting compared to Beta metadata schema. Most datasets are in the range of 60 to 65 percent of the weights. Therefore, at least half of the main fields are filled, but the result is far from expected since they do not have any dataset with more than 65%.

4.2 Case 2

Now, to demonstrate the adaptability of the proposed approach, a new metadata completeness standard schema was created, which is called in this work as 'Beta', whose difference between it and Alpha is the addition of the DCAT-AP schema metadata fields. For this, we gather the fields that composed the DCAT-AP schema referenced in the

official site of schema (COMMISSION, 2018), as can be seen in Annex A.1.

After that, according to Section 3.2.2, the next step is to relate the fields by the title and description and may categorize and attach more than one schema of one field in one field. The result is shown in Table 4.5.

Table 4.5: Beta Metadata Fields Mapping

ID	Tag	Ckan Field	Socrata Field	Opendatasoft Field	DCAT-AP Field
1	Title	Title	Title	Title	title
2	Description	Description	Description	Description	description
3	Tags	Tags	Tags	Keywords	keyword / tag
4	Last Updated	-	Last Updated	Last modification	update / modification date
5	Publisher	-	-	Publisher	publisher
6	Contact Email	-	Contact Email	-	contact point
7	Unique identifier	Unique identifier	Unique Identifier	Identifier	identifier
8	Public Access Level	-	Public Access Level	-	-
9	Agency / Department	-	Agency / Department	-	-
10	License	License	License / Rights	License	access rights | license | licence type | rights
11	Spatial Geographical Area	-	Geographic Unit	Geographic area	spatial / geographical coverage
12	Temporal Coverage	-	Temporal Coverage	-	temporal coverage
13	Data Dictionary	-	Data Dictionary	-	-
14	Language	-	-	Language	language
15	Permalink / Identifier	-	Permalink / Identifier	-	-

16	Related Documents	-	Related Documents	References	related resource
17	Theme/Category/Groups	Groups	Category	Theme	theme / category
18	API key	API key	API Endpoint	-	-
19	Frequency	-	Frequency of Data Change / Frequency of Publishing	-	frequency
20	Format	Multiple formats (if provided)	-	-	format
21	Public Access Level Comment	-	Public Access Level Comment	-	-
22	Data preview	Data preview	-	-	-
23	Revision history	Revision history	-	-	-
24	Extra fields	Extra fields	-	-	-
25	Data Steward	-	Data Steward	-	-
26	Row Count	-	Row Count	-	-
27	Download URL	-	Download URL	-	download URL
28	Source	-	Link	Attributions	source
29	Timezone	-	-	Timezone	-
30	Type	-	-	-	type
31	Dataset distribution	-	-	-	dataset distribution
32	Release date	-	-	-	release date
33	Documentation	-	-	-	documentation
34	Access URL	-	-	-	access URL
35	Byte size	-	-	-	byte size
36	Checksum	-	-	-	checksum

37	Conforms to	-	-	-	conforms to
38	End date/time	-	-	-	end date/time
39	Provenance	-	-	-	provenance
40	Version	-	-	-	version | version notes | has version | is version of
41	Linked schemas	-	-	-	linked schemas
42	Start date/time	-	-	-	start date/time
43	Media type	-	-	-	media type

Comparing the created tags with each schema, the number of fields increased. Compared to the CKAN Schema, there was a 290,90% increase in the number of fields (32 more fields). In comparison with the Socrata Schema, there was a 95,45% increase in the number of fields (30 more fields). Compared to the Opendatasoft Schema, there was a 230,76% increase in the number of fields (17 more fields). Compared to the DCAT-AP schema, there was a 38,70% increase in the number of fields (12 more fields).

As seen in Section 3.2.3, IDs were grouped in order to count which ones were present in more tools. We can see the results in Figure 4.7. To calculate the weight of each field in order to give more relevance to fields that appear in more than one schema, we used the Equation 3.1 and Table 4.6 contains the result.

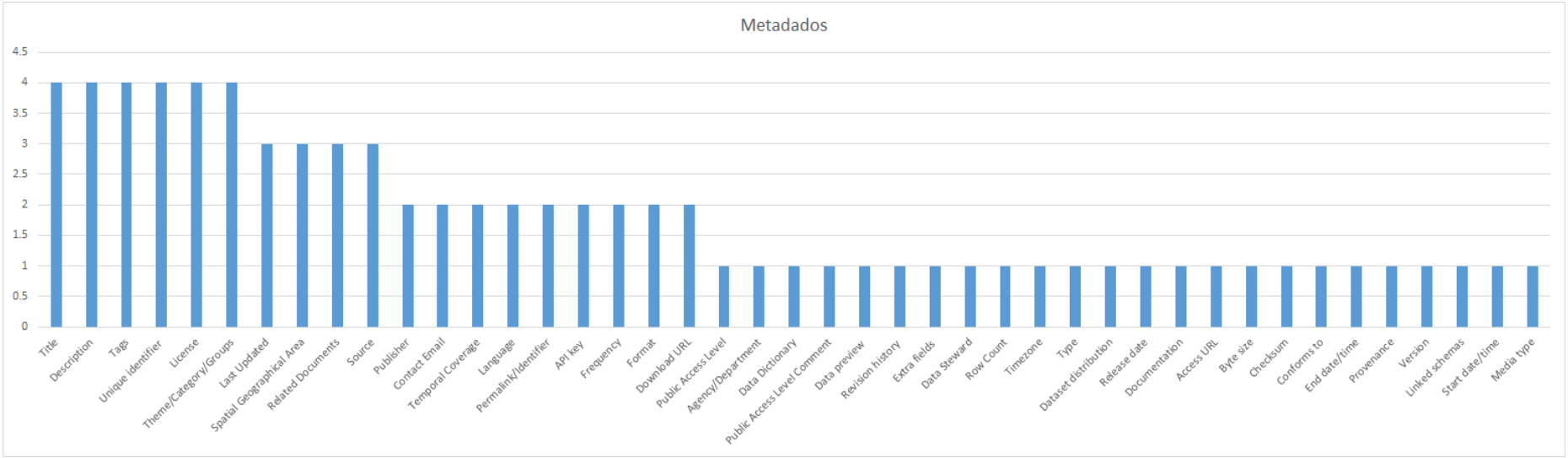

Figure 4.7: Beta Metadata Field Occurrence Graph

Table 4.6: Beta Metadata Weight in Percents

Beta Metadata Field	Weight	Groups
Title	5.13%	Group 1
Description	5.13%	Group 1
Tags	5.13%	Group 1
Unique identifier	5.13%	Group 1
License	5.13%	Group 1
Theme/Category/Groups	5.13%	Group 1
Last Updated	3.85%	Group 2
Spatial Geographical Area	3.85%	Group 2
Related Documents	3.85%	Group 2
Source	3.85%	Group 2
Publisher	2.56%	Group 3
Contact Email	2.56%	Group 3
Temporal Coverage	2.56%	Group 3
Language	2.56%	Group 3
Permalink/Identifier	2.56%	Group 3
API key	2.56%	Group 3
Frequency	2.56%	Group 3
Format	2.56%	Group 3
Download URL	2.56%	Group 3
Public Access Level	1.28%	Group 4
Agency/Department	1.28%	Group 4
Data Dictionary	1.28%	Group 4
Public Access Level Comment	1.28%	Group 4
Data preview	1.28%	Group 4
Revision history	1.28%	Group 4
Extra fields	1.28%	Group 4
Data Steward	1.28%	Group 4
Row Count	1.28%	Group 4
Timezone	1.28%	Group 4
Type	1.28%	Group 4
Dataset distribution	1.28%	Group 4
Release date	1.28%	Group 4

Documentation	1.28%	Group 4
Access URL	1.28%	Group 4
Byte size	1.28%	Group 4
Checksum	1.28%	Group 4
Conforms to	1.28%	Group 4
End date/time	1.28%	Group 4
Provenance	1.28%	Group 4
Version	1.28%	Group 4
Linked schemas	1.28%	Group 4
Start date/time	1.28%	Group 4
Media type	1.28%	Group 4

From this analysis, we verified that there are four distinct metadata groups. We sorted according to the number of times they appear in the metadata schemas. Figure 4.8 shows the amount of weight of each group.

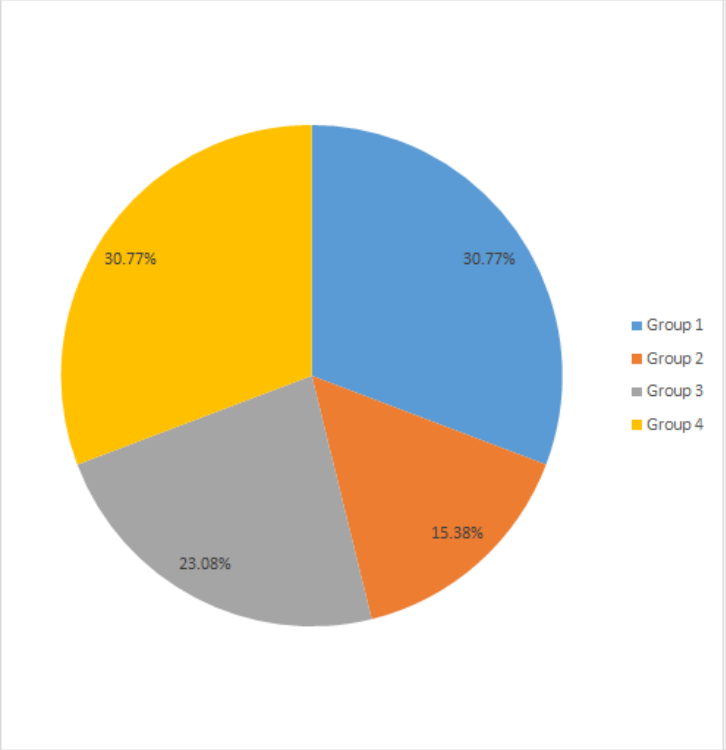

Figure 4.8: Beta Metadata Weight in Percentual per Groups

The group of metadata fields that appears in the four schemas (Group 1), thus being more relevant, amounts to 30.77% of the total weight (Title, Description, Tags, Last

Updated, Publisher, Contact Email). We classified these fields as primary information in the identification of a dataset. This supports the approach because it gives importance to the fields that have more prominence.

The group of metadata fields that appear in 3 schemas (Group 2) amounts to 15.38 % of the total weight (Last Updated, Spatial Geographical Area, Related Documents, Source). The group of metadata fields that appear in 2 schemas (Group 3) amounts to 23.08 % of the total weight (Publisher, Contact Email, Temporal Coverage, Language, Permalink/Identifier, API key, Frequency, Format, Download URL). These fields are essential, but not enough to identify the dataset that ordinary users seek.

The last group of metadata fields that appear in only 1 schema (Group 3), amounts to 36.17 % of the total weight (Publisher, Contact Email, Public Access Level, Agency / Department, Temporary Coverage, Data Dictionary, Language, Permalink / Identifier, Format, Public Access Level Comment, Data preview, Revision history, Extra fields, Data Steward, Row Count, Download URL, Timezone). Most of these fields add more detail to metadata and are very relevant to solution developers and demonstrates that much of the metadata does not cover all the information that is needed when creating tools for various industries. We noticed that the interoperability between frameworks is not yet well established concerning metadata.

As done in Section 4.1, in order to evaluate the proposed approach, we applied the metadata completeness standard schema, in this case, Beta schema, to metadata records from the European Union Open Data Portal.

Next, as seen in Section 3.3.1, all fields were aligned with each field in Beta schema. For this, we filtered all fields that occurred in all datasets. To align the fields, we analyzed the titles and description of the fields of the dataset of metadata about the Alpha schema fields. Also, we analyzed the content of the dataset of metadata punctually, in order to to have a more accurate alignment.

We noticed that some of our fields appeared in more than one dataset field. This was the cases of the 'Frequency', 'License', 'Theme/Category/Groups', 'Temporal Coverage', 'Type' and 'Version' fields. For 'Frequency' there are two fields: 'temporal_granularity' and 'accrual_periodicity'. For 'License' there are three fields: 'license_id', 'license_title' and 'license_url'. For 'Theme/Category/Groups' there are two fields: 'concepts_eurovoc' and 'groups'. For 'Temporal Coverage' there are two fields: 'temporal_coverage_to' and 'temporal_coverage_from'. For 'Type' there are two fields: 'type' and 'type_of_dataset'. For 'version' there are two fields: 'version' and 'version_description'. For these fields, we

proportionally divided the weight according to the number of correlated fields.

Other fields could not be aligned, such as 'Access URL', 'Byte size', 'Checksum', 'Conforms to', 'Data preview', 'Documentation', 'End date/time', 'Linked schemas', 'Media type', 'Public Access Level Comment', 'Row Count', 'Dataset distribution', 'Start date/time' and 'Timezone'. With this, we observed that no dataset could cover 100% of our set of metadata, since the sum of the weights of the cited fields reaches 17.95%, causing them to reach a maximum of 82.05%. Table 4.7 shows the metadata field correspondences and their respective weights.

Table 4.7: Beta Metadata Fields Correlation With European Portal Metadata Fields

Beta Metadata Title	Europe Dataset Metadata Title	Weight
Description	Key: description	5.13%
Tags	Key: keywords	5.13%
Title	Key: title	5.13%
Unique identifier	Key: identifier	5.13%
Last Updated	Key: modified_date	3.85%
Related Documents	Key: resources	3.85%
Source	Key: resources(link)	3.85%
Spatial Geographical Area	Key: geographical_coverage	3.85%
Theme/Category/Groups	Key: concepts_eurovoc	2.56%
Theme/Category/Groups	Key: groups	2.56%
API key	Key: owner_org	2.56%
Contact Email	Key: contact_email	2.56%
Download URL	Key: resources(download)	2.56%
Format	Key: resources(format)	2.56%
Language	Key: language	2.56%
Permalink/Identifier	Key: url	2.56%
Publisher	Key: author	2.56%
License	Key: license_id	1.71%
License	Key: license_title	1.71%
License	Key: license_url	1.71%
Access URL	-	1.28%
Agency/Department	Key: organization	1.28%
Byte size	-	1.28%
Checksum	-	1.28%

Conforms to	-	1.28%
Data Dictionary	Key: resources(name)	1.28%
Data preview	-	1.28%
Data Steward	Key: maintainer	1.28%
Documentation	-	1.28%
End date/time	-	1.28%
Extra fields	Key: extras	1.28%
Frequency	Key: temporal_granularity	1.28%
Frequency	Key: accrual_periodicity	1.28%
Linked schemas	-	1.28%
Media type	-	1.28%
Provenance	Key: owner_org	1.28%
Public Access Level	Key: private	1.28%
Public Access Level Comment	-	1.28%
Release date	Key: release_date	1.28%
Revision history	Key: revision_timestamp	1.28%
Row Count	-	1.28%
Source	-	1.28%
Start date/time	-	1.28%
Temporal Coverage	Key: temporal_coverage_to	1.28%
Temporal Coverage	Key: temporal_coverage_from	1.28%
Timezone	-	1.28%
Type	Key: type	0.64%
Type	Key: type_of_dataset	0.64%
Version	Key: version	0.64%
Version	Key: version_description	0.64%
Notcorrelated	Key: contact_name	0.00%
Notcorrelated	Key: contact_webpage	0.00%
Notcorrelated	Key: contact_address	0.00%
Notcorrelated	Key: alternative_title	0.00%
Notcorrelated	Key: capacity	0.00%
Notcorrelated	Key: contact_telephone	0.00%
Notcorrelated	Key: creator_user_id	0.00%
Notcorrelated	Key: interoperability_level	0.00%

Notcorrelated	Key: isopen	0.00%
Notcorrelated	Key: metadata_created	0.00%
Notcorrelated	Key: metadata_language	0.00%
Notcorrelated	Key: metadata_modified	0.00%
Notcorrelated	Key: name	0.00%
Notcorrelated	Key: num_resources	0.00%
Notcorrelated	Key: num_tags	0.00%
Notcorrelated	Key: rdf	0.00%
Notcorrelated	Key: relationships_as_object	0.00%
Notcorrelated	Key: relationships_as_subject	0.00%
Notcorrelated	Key: revision_id	0.00%
Notcorrelated	Key: state	0.00%
Notcorrelated	Key: status	0.00%
Notcorrelated	Key: tracking_summary	0.00%
Notcorrelated	Key: maintainer_email	0.00%
Notcorrelated	Key: author_email	0.00%

After the consolidated alignment, we applied the weights to each record metadata and thus having the degree of completeness each of them. The results can be seen in Figure 4.9.

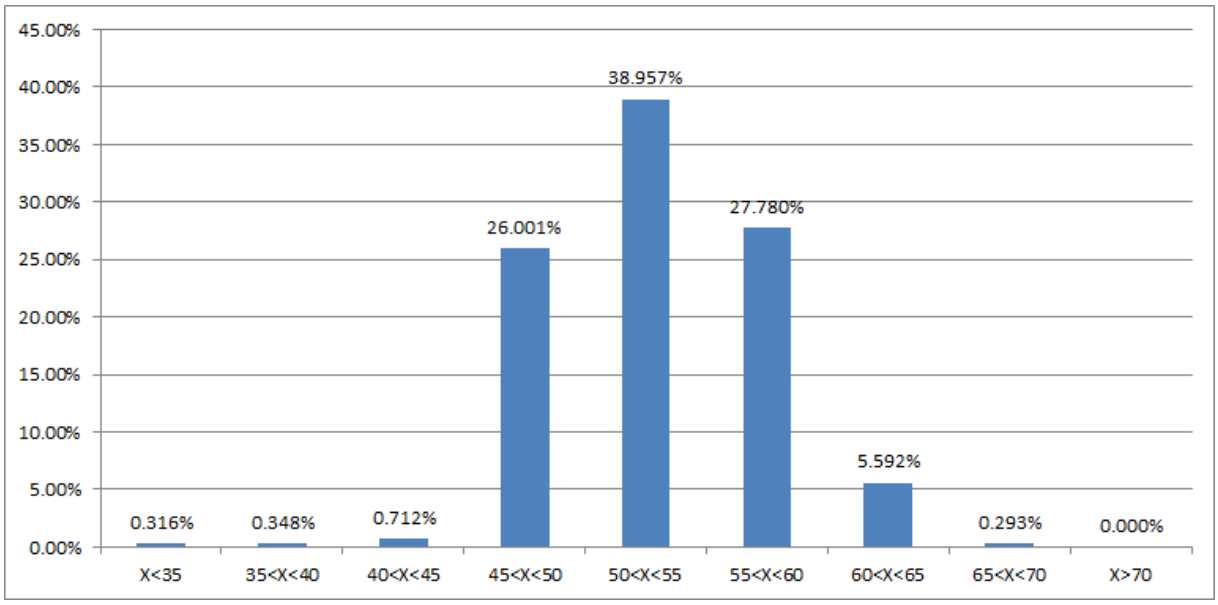

Figure 4.9: Europe Metadata Dataset Weight in Percentual Ranges of Beta Schema

As can be seen, most data sets have more than 45% weighting compared to Beta

metadata schema. Most datasets are in the range of 50 to 55 percent of the weights. Only 0.29% of the datasets reaches more than 65 percent of weight. At least half of the main fields are filled, but the result is far from expected since they do not have any dataset with more than 70%.

In addition to presenting the metadata records degree of completeness, this assessment makes it possible to evaluate the amount of metadata filled in each metadata field, as we can see in Figure 4.10.

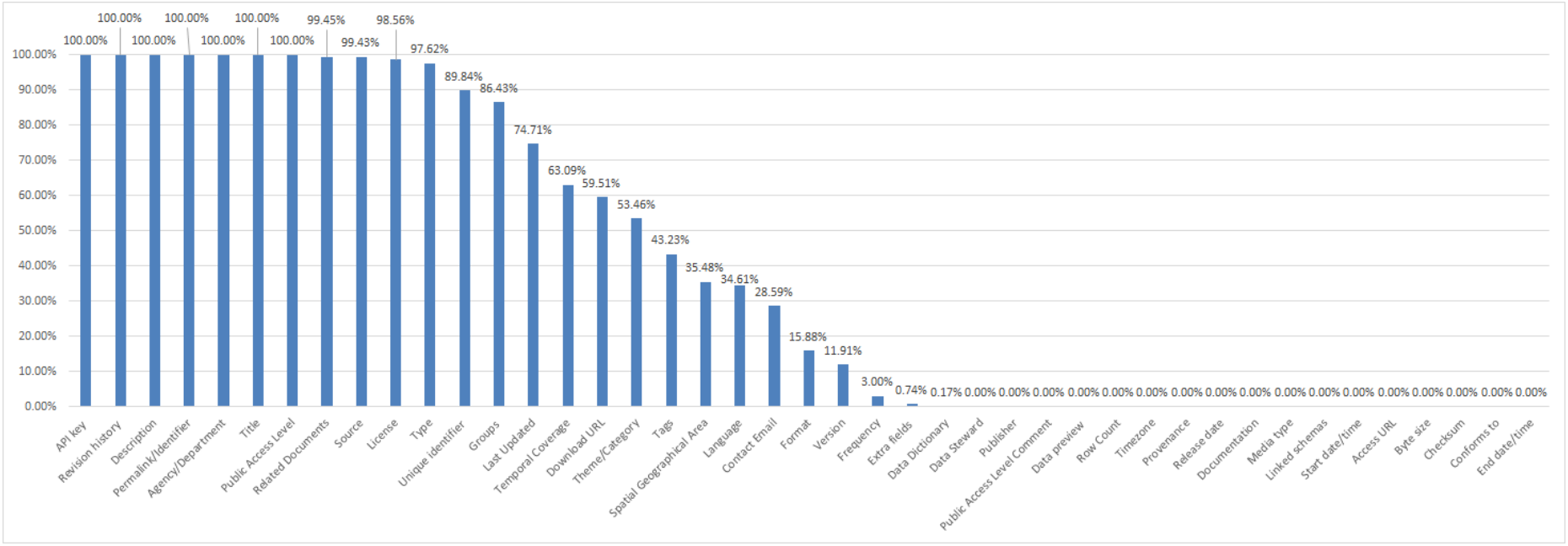

Figure 4.10: Percents of Europe Metadata Dataset Field Completeness in Comparison to Beta Metadata Schema

The most filled fields were ‘API key’, ‘Revision history’, ‘Description’, ‘Permalink/Identifier’, ‘Agency/Department’, ‘Title’, ‘Public Access Level’, having 100% of datasets with these fields filled, which is equivalent to 19.23% by weight in relation to Beta metadata schema.

However, those that have had no occurrence are: ‘Data Steward’, ‘Publisher’, ‘Public Access Level Comment’, ‘Data preview’, ‘Row Count’, ‘Timezone’, ‘Provenance’, ‘Release date’, ‘Documentation’, ‘Media type’, ‘Linked schemas’, ‘Start date/time’, ‘Access URL’, ‘Byte size’, ‘Checksum’, ‘Conforms to’ and ‘End date/time’. They are equivalent to 23.08% of the weight in relation to the metadata set.

After that, we also applied the process to NYS open data portal metadata records. So, we aligned all fields with each field in Beta schema. For this, we filtered all fields that occurred in all datasets. To align the fields, we analyzed the titles and description of the fields of the dataset of metadata about the Beta schema fields. Also, we analyzed the content of the dataset of metadata punctually, in order to to have a more accurate alignment.

We noticed that some of our fields appeared in more than one dataset field. This was the case of the ‘Spatial Geographical Area’ field. For ‘Spatial Geographical Area’ there are two fields: ‘Coverage’ and ‘Localities’. For this field, we proportionally divided the weight according to the number of correlated fields.

Other fields could not be aligned, such as ‘Access URL’, ‘Byte size’, ‘Checksum’, ‘Conforms to’, ‘Data Dictionary’, ‘Data Steward’, ‘Dataset distribution’, ‘Documentation’, ‘Download URL’, ‘End date/time’, ‘Extra fields’, ‘Format’, ‘Language’, ‘License’, ‘Linked schemas’, ‘Media type’, ‘Provenance’, ‘Public Access Level’, ‘Public Access Level Comment’, ‘Related Documents’, ‘Revision history’, ‘Row Count’, ‘Start date/time’, ‘Temporal Coverage’, ‘Timezone’ and ‘Version’. With this, it can be observed that no dataset could cover 100% of our set of metadata, since the sum of the weights of the cited fields reaches 44.87%, causing them to reach a maximum of 55.13%. The metadata and all correspondences and their respective weights can be seen in Table 4.8.

Table 4.8: Beta Metadata Fields Correlation with NYS Metadata Fields

Metadata Title	Dataset Metadata	Weight
Description	Description	5.13%
License	-	5.13%
Tags	Keywords	5.13%

Theme/Category/Groups	Category	5.13%
Title	Name	5.13%
Unique identifier	U ID	5.13%
Last Updated	Last Update Date (data)	3.85%
Related Documents	-	3.85%
Source	Source Link	3.85%
Spatial Geographical Area	Coverage	1.92%
Spatial Geographical Area	Localities	1.92%
API key	api_endpoint	2.56%
Contact Email	Contact Information	2.56%
Download URL	-	2.56%
Format	-	2.56%
Frequency	Posting Frequency	2.56%
Language	-	2.56%
Permalink/Identifier	URL	2.56%
Publisher	Data Provided By	2.56%
Temporal Coverage	-	2.56%
Access URL	-	1.28%
Agency/Department	Agency	1.28%
Byte size	-	1.28%
Checksum	-	1.28%
Conforms to	-	1.28%
Data Dictionary	-	1.28%
Data preview	Derived View	1.28%
Data Steward	-	1.28%
Dataset distribution	-	1.28%
Documentation	-	1.28%
End date/time	-	1.28%
Extra fields	-	1.28%
Linked schemas	-	1.28%
Media type	-	1.28%
Provenance	-	1.28%
Public Access Level	-	1.28%
Public Access Level Comment	-	1.28%

Release date	Creation Date	1.28%
Revision history	-	1.28%
Row Count	-	1.28%
Start date/time	-	1.28%
Timezone	-	1.28%
Type	Type	1.28%
Version	-	1.28%
Notcorrelated	Domain	0.00%
Notcorrelated	Organization	0.00%
Notcorrelated	See Also	0.00%
Notcorrelated	Granularity	0.00%
Notcorrelated	Limitations	0.00%
Notcorrelated	Notes	0.00%
Notcorrelated	Owner	0.00%
Notcorrelated	Visits	0.00%
Notcorrelated	Downloads	0.00%
Notcorrelated	Parent UID	0.00%
Notcorrelated	County Filter	0.00%
Notcorrelated	County Column	0.00%
Notcorrelated	Municipality Filter	0.00%
Notcorrelated	Municipality_Column	0.00%

In addition to presenting the metadata records degree of completeness, this assessment makes it possible to evaluate the amount of metadata filled in each metadata field, as we can see in Figure 4.11.

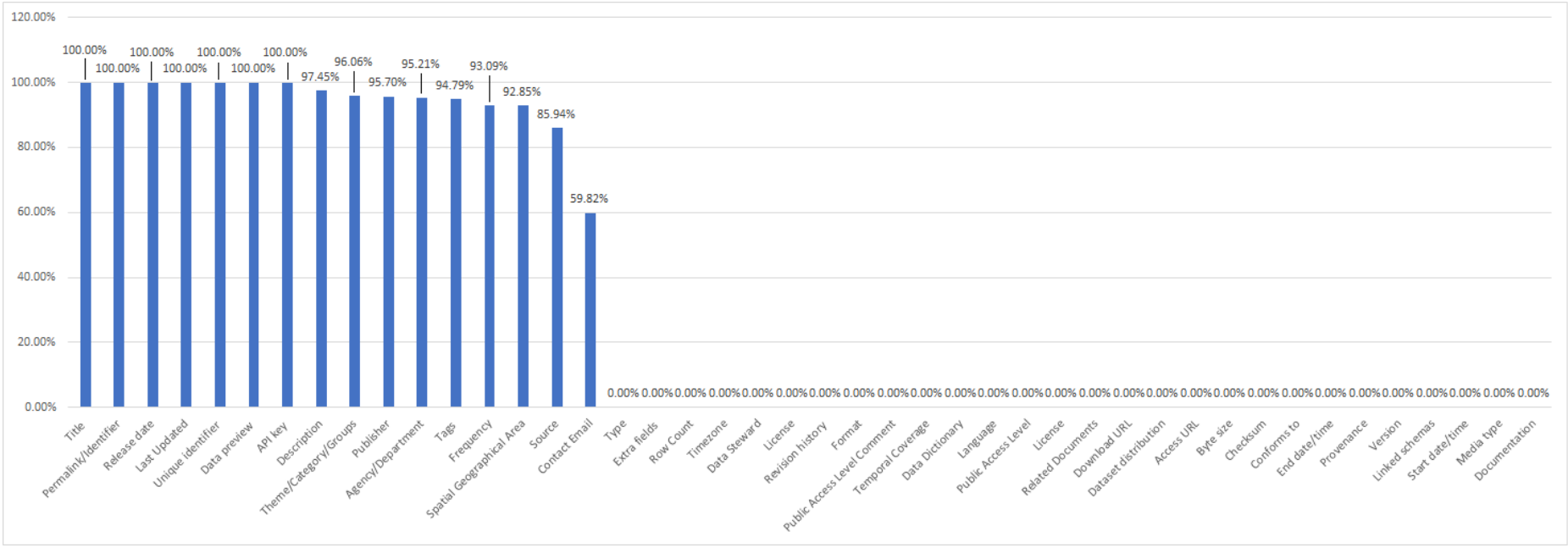

Figure 4.11: Percents of NYS Metadata Dataset Field Completeness in Comparison to Beta Metadata Schema

The most filled fields were ‘Title’, ‘Permalink/Identifier’, ‘Last Updated’, ‘Unique identifier’, ‘Data preview’ and ‘API key’, having 100% of datasets with these fields filled, which is equivalent to 26,09% by weight in relation to Beta metadata schema.

However, those that have had no occurrence are: ‘Extra fields’, ‘Row Count’, ‘Timezone’, ‘Data Steward’, ‘License’, ‘Revision history’, ‘Format’, ‘Public Access Level Comment’, ‘Temporal Coverage’, ‘Data Dictionary’, ‘Language’, ‘Public Access Level’, ‘License’, ‘Related Documents’ and ‘Download URL’. They are equivalent to 36,96% of the weight in relation to the metadata set.

After the consolidated alignment, we applied the weights to each record metadata and thus having the degree of completeness each of them. The results can be seen in Figure 4.12.

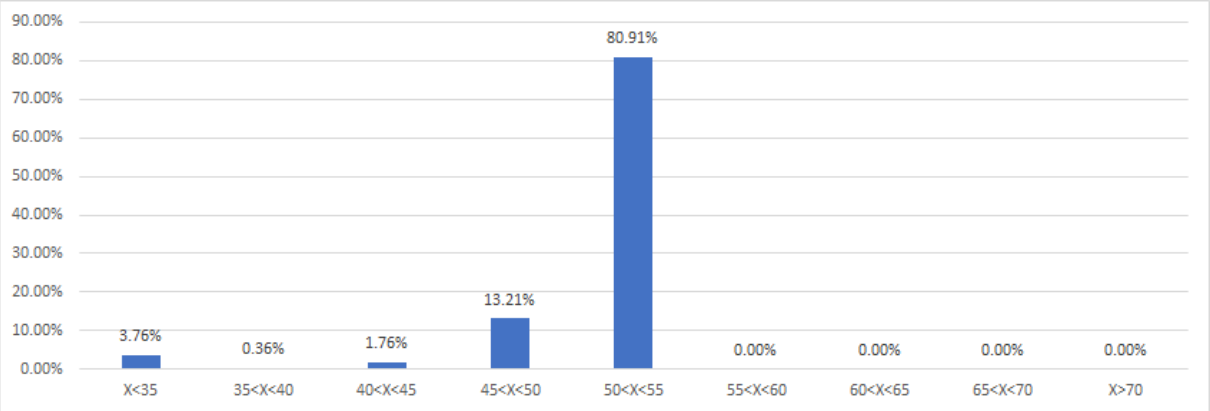

Figure 4.12: NY Metadata Dataset Weight in Percentual Ranges of Beta Schema

As can be noticed, most data sets have more than 55% weighting compared to Beta metadata schema. Most datasets are in the range of 60 to 65 percent of the weights.

Hence, at least half of the main fields are filled, but the result is far from expected since they do not have any dataset with more than 65%.

4.3 Comparison Between Metadata Schemas

After applying the proposed approach, we analyzed the Alpha and Beta schemas. The first apparent difference is seen in the number of metadata fields. Alpha presents 29 Metadata fields, while Beta presents 43 metadata fields, as shown in Figure 4.13.

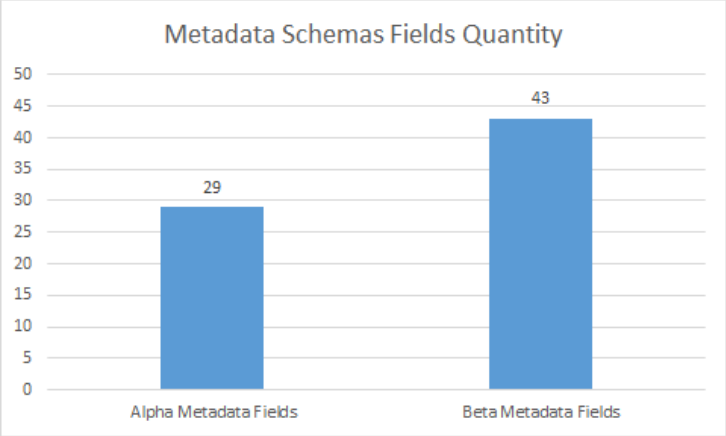

Figure 4.13: Quantity Comparison Between Alpha and Beta Schemas

The difference can be because the DCAT-AP metadata schema was created with the objective of interoperability between European portals, to enable a connection and links between datasets with the open data portal of Europe (COMMISSION, 2018). With this, the schema holds a higher amount of metadata, with more detail. When we applied the approach is applied with the DCAT-AP schema, it ends up introducing several fields without correlation with the other schemas (CKAN, Socrata, and Opendatasoft).

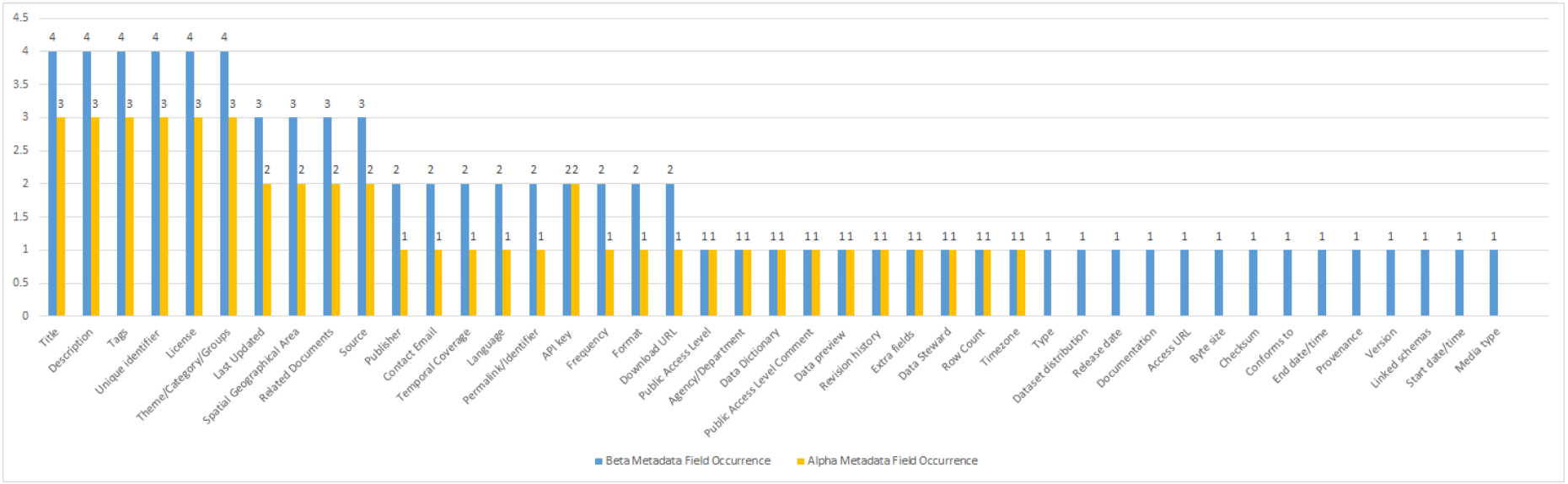

Figure 4.14: Occurrences Comparison Between Alpha and Beta Metadata Fields

Comparing field to field of each schema, as seen in Figure 4.14, some appear in all related schemas (Title, Description, Tags, Unique identifier, License, Theme / Category / Groups). By analyzing this information, these fields can be considered as the key fields of the metadata schemas, and therefore, the most relevant fields of the Alpha and Beta schemas. It is also noted that the Last Updated, Spatial Geographical Area, Related Documents, Source fields do not occur in one of the schemas (CKAN).

4.4 Comparison Between Analyses

After the analyses made in Section 4.1 and Section 4.1, we performed a comparison between the analyses, in order to investigate the nuances and characteristics that differ the analyses.

First, we compared the analysis of European metadata dataset to Alpha scheme and analysis of European metadata dataset to the Beta scheme, as shown in Figure 4.15.

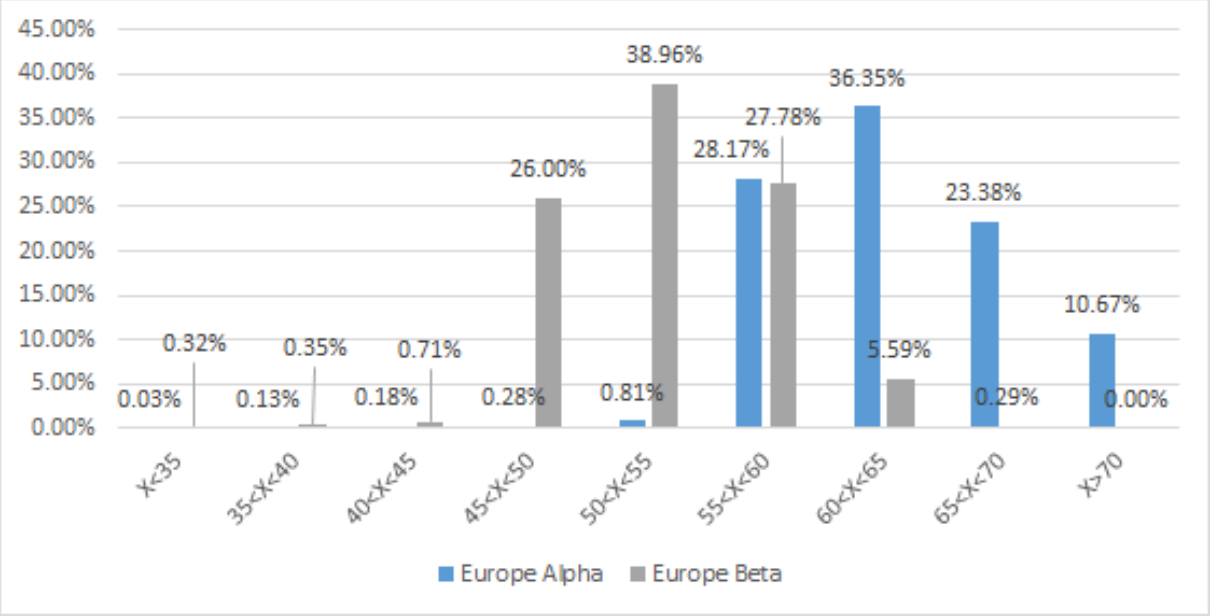

Figure 4.15: Comparison Between Europe Dataset Alpha X Beta Schema

The metadata present in the dataset of the European Union open data portal, as can be noticed, has better adherence to the Alpha schema than to the Beta schema. Note that in the alpha scheme, most of the metadata was classified in the region above 55%. In the beta scheme, most of the metadata was classified in the region between 45% and 55%. This is because the Beta schema is more detailed because it contains more metadata fields to fill.

This result is surprising since the Beta scheme incorporates the fields of the DCAT-AP scheme, a scheme created by the open data portal of the European Union so that there is greater integration between the open data portals of other countries, and even then the metadata adhered to the Alpha schema better. One possibility of this detachment from the Beta schema may be the time difference between the creation of the analyzed dataset (PORTAL, 2016) and the most current revision of the DCAT-AP metadata scheme (COMMISSION, 2018).

Next, it is compared to the analysis of NYS dataset with relation to the Alpha scheme and analysis of NYS metadata dataset with relation to the Beta scheme, as shown in Figure 4.16.

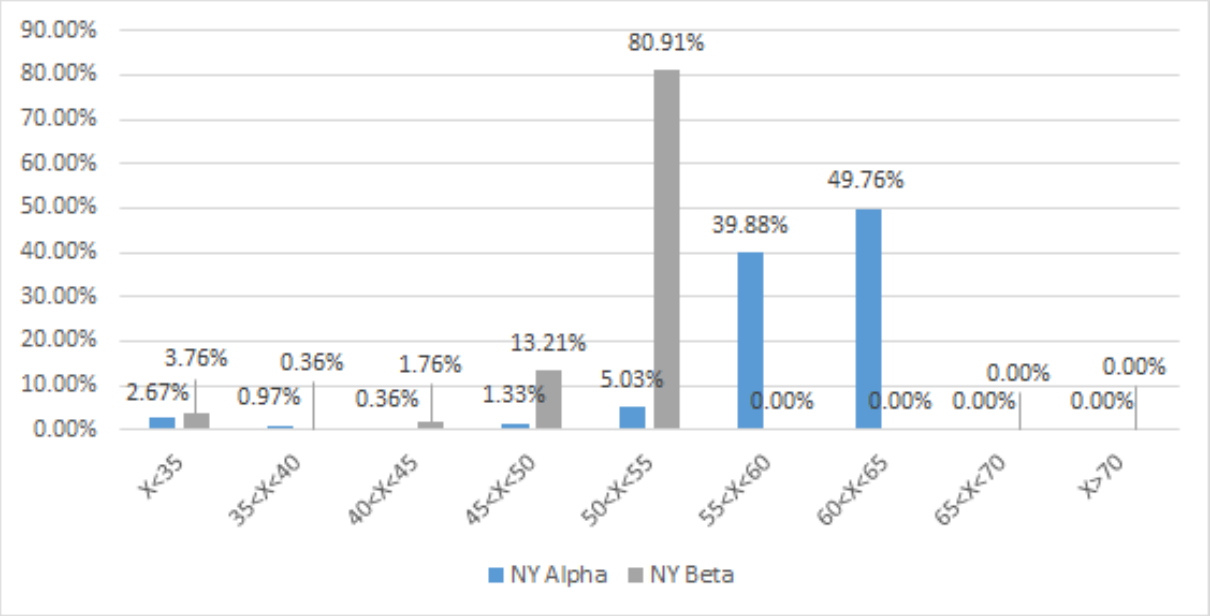

Figure 4.16: Comparison Between NYS Dataset Alpha X Beta Schema

As can be identified, the metadata of the NYS dataset has more adherence to the alpha schema than to the beta schema. Its metadata is in the region between 55% and 65% about the Alpha scheme. Already in the Beta scheme, it focuses on the region from 50% to 55%.

The NYS dataset In addition to having fewer metadata fields, it is also not very well populated. Hence, in the Alpha schema, it does not get a good correlation.

Then, it was compared to the analysis of Europe dataset with relation to the Alpha scheme and analysis of NYS metadata dataset with relation to the Alpha scheme, as shown in Figure 4.17.

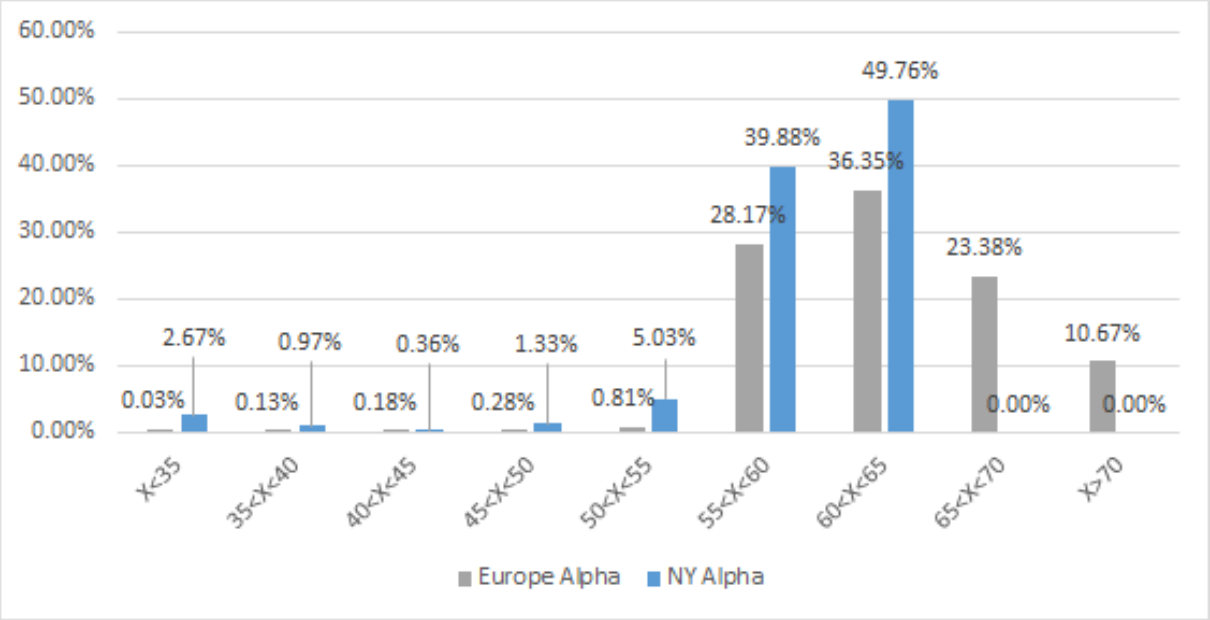

Figure 4.17: Comparison Between Europe X NYS Dataset Alpha Schema

Comparing both datasets, both of them are above 55% about the Alpha scheme. However, the dataset of Europe can perform better, as can 34.05% of its datasets are in the region above 65%. Also, only 1.43% of Europe's metadata is in the region below 55%, compared to 10.36% of the NYS metadata located in the same region.

Finally, it was compared to the analysis of European metadata dataset with the Beta scheme and analysis of NYS metadata dataset with the Beta scheme, as shown in Figure 4.17.

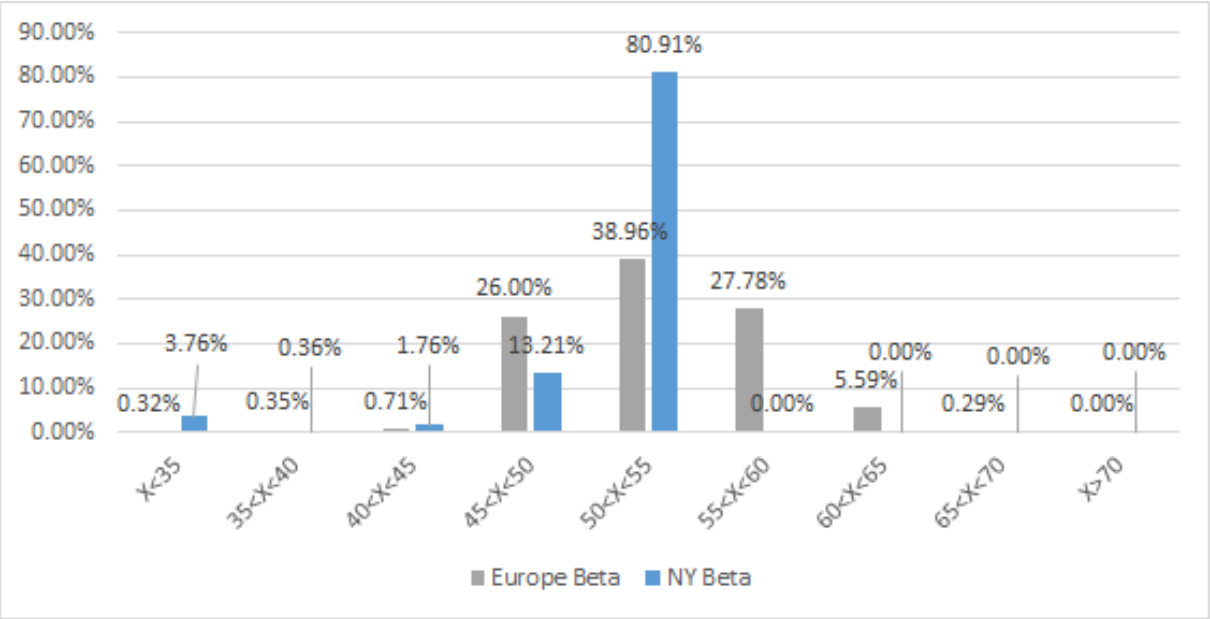

Figure 4.18: Comparison Between Europe X NYS Dataset Beta Schema

Comparing the datasets to the beta schema, the NYS metadata dataset performs better than European metadata dataset, with 80.91% of its metadata in the region between 50% and 55%. The metadata of the dataset in Europe, however, accounted for 72.33% in the region above 50%, with 27.78% being in the region between 55% and 60%.

Chapter 5

Conclusion

Metadata is a very significant component of open data portals datasets. Without them, data can be confusing, disconnected, irrelevant, and untraceable to the target audience, whether they be ordinary people with data access needs, developers to power their applications with data loads, or researchers to support their research.

With this, comes the need to have consistent metadata and to be able to describe the datasets assigned to them. The metadata needs to be filled out, and their content clear and understandable. Thus, it is necessary to standardize frameworks concerning the metadata issue so that we can make a possible connection between portals or between datasets.

In terms of data management, several works recognize metadata having relevance in several methodologies, such as in data governance(NWABUDE et al., 2014). Several authors point out how important this topic is when it comes to efficient data management (BAUER; KALTENBÖCK, 2011).

With this, the study proposed covers an approach for creating a metadata schema, aligning other metadata schemas into a single one, in order to assess the completeness of metadata of other datasets. Thus, to better illustrate, a case of study was made with more than 12,000 metadata dataset from European Union open data portal and with more than 1,500 metadata dataset from NYS'S open data portal.

After applying the approach, it was created completeness quality indicators. The more a field occurs, the more relevant it is and thus can classify how much the metadata of a dataset has been filled and somehow identified better than others. For this two metadata completeness standard schemas were created: The first (Alpha) covering the metadata seen in the three main frameworks (CKAN, 2017)(SOCRATA, 2017)(OPENDATASOFT,

2018) and the second (Beta) covering the metadata of the frameworks of the previous scheme with the addition of the DCAT-AP metadata scheme (COMMISSION, 2018). It was noted that the frameworks do not have uniform and standardized metadata fields, which generated a set of twenty-nine fields for the alpha scheme, where only six of them were common to all three tools and generated a set of forty-three fields for the Beta scheme, where only five of them were common to all four tools. This fact leads us to believe that metadata schemas are not standardized, which makes it challenging to connect datasets and interoperability between portals becomes difficult, as well as making it harder for both developers and ordinary users to search for the requested data. Also, more detailed metadata schemes, that is, with more fields, prevailed in the final scheme, surpassing those in less detail.

Thus, it was compared both datasets, from the portal of the European Union and the portal of NYS, with the two new schemes. The conclusion was that the datasets were more responsive to the Alpha schema because they had fewer fields and were less detailed than the Beta schema. The European Union metadata dataset had more adherence to the Alpha scheme, while the NYS dataset had more adherence to the beta scheme and this points us to conclude that the two portals do not load the metadata of their datasets correctly, their schemes have no standard and their communication still depends on the homogenization of their fields.

Thus, in terms of data completeness, the results obtained shows that the approach is consistent, being able to measure in a practical way the completeness of metadata of open data portals is about other metadata schemas. Besides, the approach is scalable and can be done with any quantity schemas. The schemas can be chosen to best fit in the main objective of the assessment of the metadata.

5.1 Future Work

After analyzing both the proposed approach and its implementation, some improvements are identified.

In work in question, we approached only one metric of the correlation of title and description of the metadata fields. For the correlation between being somewhat more consistent, semantic methods must be applied in field correlation to make this connection more reliable. It is necessary to investigate the different current methodologies and to analyze which approach is best adapted.

Also, the approach only aims the quality the completeness of the fields, not taking into account if they are with their relevant content, understandable, and valid for their objectives. For this, a method to read the metadata that can understand its content and classify it according to metrics that qualify must be created.

In order to have more consistency with the experiment, real-time metadata from a more significant number of portals could be collected. With this, the behavior of the created schema could be analyzed to see if it is compatible with most portals.

Finally, it would be to create a tool that would read batch datasets and create metrics reports, and convert the metadata to different frameworks, sparing the work of publishers from having to adapt their metadata.

References

HUIJBOOM, N.; BROEK, T. Van den. Open data: an international comparison of strategies. *European journal of ePractice*, v. 12, n. 1, p. 4–16, 2011.

RIBEIRO, C. J. S.; ALMEIDA, R. F. d. Dados abertos governamentais (open government data): instrumento para exercício de cidadania pela sociedade. *XII Enancib-Políticas de Informação para a Sociedade-Anais. Brasília: Thesaurus*, p. 2568–2580, 2011.

ATTARD, J.; ORLANDI, F.; SCERRI, S.; AUER, S. A systematic review of open government data initiatives. *Government Information Quarterly*, v. 32, n. 4, p. 399–418, 2015.

ROJAS, L. A. R.; BERMÚDEZ, G. M. T.; LOVELLE, J. M. C. Open data and big data: A perspective from colombia. In: SPRINGER. *International Conference on Knowledge Management in Organizations*. [S.l.], 2014. p. 35–41.

BEGHIN, N.; ZIGONI, C. et al. Avaliando os websites de transparência orçamentária nacionais e subnacionais e medindo impactos de dados abertos sobre direitos humanos no brasil. Instituto de Estudos Socioeconômicos, 2014.

CHATTAPADHYAY, S. Access and use of government data by research and advocacy organisations in india: A survey of (potential) open data ecosystem. In: ACM. *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*. [S.l.], 2014. p. 361–364.

STARKE, C.; NAAB, T. K.; SCHERER, H. Free to expose corruption: The impact of media freedom, internet access and governmental online service delivery on corruption. *International Journal of Communication*, v. 10, p. 21, 2016.

RIBEIRO, D. C.; VO, H. T.; FREIRE, J.; SILVA, C. T. An urban data profiler. In: ACM. *Proceedings of the 24th International Conference on World Wide Web*. [S.l.], 2015. p. 1389–1394.

TYGEL, A.; AUER, S.; DEBATTISTA, J.; ORLANDI, F.; CAMPOS, M. L. M. Towards cleaning-up open data portals: A metadata reconciliation approach. In: IEEE. *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on*. [S.l.], 2016. p. 71–78.

HENDLER, J.; HOLM, J.; MUSIALEK, C.; THOMAS, G. Us government linked open data: semantic. data. gov. *IEEE Intelligent Systems*, v. 27, n. 3, p. 25–31, 2012.

INTERNATIONAL, O. K. *Data Portals*. 2011. Http://dataportals.org/. Last accessed in February 12, 2019.

REIS, J. R.; VITERBO, J.; BERNARDINI, F. A rationale for data governance as an approach to tackle recurrent drawbacks in open data portals. In: ACM. *Proceedings of*

the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. [S.l.], 2018. p. 73.

BEALL, J. Metadata and data quality problems in the digital library. *Journal of Digital Information*, v. 6, n. 3, 2005.

MARGARITOPOULOS, M.; MARGARITOPOULOS, T.; MAVRIDIS, I.; MANITSARIS, A. Quantifying and measuring metadata completeness. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 63, n. 4, p. 724–737, 2012.

BRAUNSCHWEIG, K.; EBERIUS, J.; THIELE, M.; LEHNER, W. The state of open data: Limits of current open data platforms. *In the International World Wide Web Conference*, Lyon, France, 2012.

ZUIDERWIJK, A.; JEFFERY, K.; JANSSEN, M. The necessity of metadata for linked open data and its contribution to policy analyses. In: *Proceedings CeDEM2012 Conference, Donau-Universitat, Krems*. [S.l.: s.n.], 2012. p. 281–94.

WEBFOUNDATION, W. W. *Open Data Barometer 4th edition Data*. 2017. Https://opendatabarometer.org/4thedition/report/. Last accessed in March 8, 2019.

NEUMAIER, S.; UMBRICH, J.; POLLERES, A. Automated quality assessment of metadata across open data portals. *J. Data and Information Quality*, ACM, New York, NY, USA, v. 8, n. 1, p. 2:1–2:29, out. 2016. ISSN 1936-1955. Available from Internet: <http://doi.acm.org/10.1145/2964909>.

OCHOA, X.; DUVAL, E. Automatic evaluation of metadata quality in digital repositories. *International journal on digital libraries*, Springer, v. 10, n. 2-3, p. 67–91, 2009.

FRIESEN, N. International lom survey report. *ISO/IEC JTC1/SC36 sub-committee*, 2004.

GUINCHARD, C. Dublin core use in libraries: a survey. *OCLC Systems & Services: International digital library perspectives*, MCB UP Ltd, v. 18, n. 1, p. 40–50, 2002.

NAJJAR, J.; TERNIER, S.; DUVAL, E. The actual use of metadata in ariadne: an empirical analysis. In: CITESEER. *Proceedings of the 3rd Annual ARIADNE Conference*. [S.l.], 2003. p. 1–6.

BARTON, J.; CURRIER, S.; HEY, J. M. 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*. [S.l.: s.n.], 2003. p. 39–48.

LIDDY, E. D.; ALLEN, E.; HARWELL, S.; CORIERI, S.; YILMAZEL, O.; OZGENCIL, N. E.; DIEKEMA, A.; MCCRACKEN, N.; SILVERSTEIN, J.; SUTTON, S. Automatic metadata generation & evaluation. In: ACM. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 2002. p. 401–402.

NWABUDE, C.; BEGG, C.; MCROBBIE, G. Data governance in small businesses-why small business framework should be different. *International Proceedings of Economics Development and Research*, IACSIT Press, v. 82, p. 101, 2014.

KHATRI, V.; BROWN, C. V. Designing data governance. *Commun. ACM*, ACM, New York, NY, USA, v. 53, n. 1, p. 148–152, jan. 2010. ISSN 0001-0782. Available from Internet: <http://doi.acm.org/10.1145/1629175.1629210>.

BRÜMMER, M.; BARON, C.; ERMILOV, I.; FREUDENBERG, M.; KONTOKOSTAS, D.; HELLMANN, S. Dataid: Towards semantically rich metadata for complex datasets. In: ACM. *Proceedings of the 10th International Conference on Semantic Systems*. [S.l.], 2014. p. 84–91.

REICHE, K. J.; HÖFIG, E. Implementation of metadata quality metrics and application on public government data. In: IEEE. *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*. [S.l.], 2013. p. 236–241.

DUVAL, E.; HODGINS, W.; SUTTON, S.; WEIBEL, S. L. Metadata principles and practicalities. *D-lib Magazine*, Citeseer, v. 8, n. 4, p. 1082–9873, 2002.

INTERNATIONAL, O. K. *The open definition*. 2005. Https://www.opendefinition.org. Last accessed in January 28, 2019.

MURRAY-RUST, P. Open data in science. *Serials Review*, Elsevier, v. 34, n. 1, p. 52–64, 03 2008.

OBAMA, B. Transparency and open government, memorandum for the heads of executive departments and agencies. *http://www.whitehouse.gov/the_ press_ office/TransparencyandOpenGovernment*, 2009.

MOLLOY, J. C. The open knowledge foundation: open data means better science. *PLoS biology*, Public Library of Science, v. 9, n. 12, p. e1001195, 2011.

BAUER, F.; KALTENBÖCK, M. Linked open data: The essentials. *Edition mono/monochrom, Vienna*, 2011.

OECD. *Open Government in Latin America*. [s.n.], 2014. 260 p. Available from Internet: <https://www.oecd-ilibrary.org/content/publication/9789264223639-en>.

COX, P.; ALEMANNO, G. Directive 2003/98/ec of the european parliament and of the council of 17 november 2003 on the re-use of public sector information. *Official Journal of the European Union*, v. 46, n. L 345, 2003.

OECD. *Open Government Data*. 2017. Http://www.oecd.org/gov/digital-government/open-government-data.htm. Last accessed in March 22, 2019.

UBALDI, B. Open government data. OECD Publishing, 2013.

KASSEN, M. A promising phenomenon of open data: A case study of the chicago open data project. *Government Information Quarterly*, v. 30, n. 4, p. 508–513, 2013.

WESSELS, B.; FINN, R.; WADHWA, K.; SVEINSDOTTIR, T. *Open Data and the Knowledge Society*. [S.l.]: Amsterdam University Press, 2017.

ZIJLSTRA, T.; JANSSEN, K. *The new PSI directive-as good as it seems?. Open Knowledge Foundation Blog*. 2013. Http://blog.okfn.org/2013/04/19/the-new-psi-directive-as-good-as-it-seems/. Last accessed in March 22, 2019.

NETWORK, C. K. A. *CKAN*. 2017. Https://ckan.org/. Last accessed in January 10, 2019.

SOCRATA. *Data-driven innovation of government programs*. 2005. Https://socrata.com/. Last accessed in January 14, 2019.

OPENDATASOFT. *Standard metadata—OpenDataSoft Documentation 1.0 documentation*. 2018. Https://help.opendatasoft.com/platform/en/publishing_data/06_configuring_metadata/standard_metadata/. Last accessed in December 08, 2018.

GREENBERG, J. Metadata and the world wide web. *Encyclopedia of library and information science*, Marcel Dekker New York, NY, v. 3, p. 1876–1888, 2003.

GREENBERG, J.; GAROUFALLOU, E. Change and a future for metadata. In: SPRINGER. *Research Conference on Metadata and Semantic Research*. [S.l.], 2013. p. 1–5.

GREENBERG, J.; ROBERTSON, W. D. Semantic web construction: an inquiry of authors' views on collaborative metadata generation. In: *International Conference on Dublin Core and Metadata Applications*. [S.l.: s.n.], 2002. p. 45–52.

LISOWSKA, B. *Metadata for the open data portals*. Technical Report. Joined-up Data Standards Project, 2016.

STANDARDIZATION, I. O. for. Iso 23081-1 information and documentation - records management processes - metadata for records - part 1: Principies. 2006.

CHAN, L. M.; ZENG, M. L. Metadata interoperability and standardization–a study of methodology part i. *D-Lib magazine*, Corporation for National Research Initiatives, v. 12, n. 6, p. 3, 2006.

MAALI, F.; ERICKSON, J. *Data Catalog Vocabulary (DCAT)*. 2014. Http://www.w3.org/TR/vocab-dcat/. Last accessed in January 22, 2019.

COMMISSION, E. *Releases for DCAT application profile for data portals in Europe solution | Joinup*. 2018. Https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/releases/. Last accessed in January 22, 2019.

MOREIRA, B. L.; GONÇALVES, M. A.; LAENDER, A. H.; FOX, E. A. Automatic evaluation of digital libraries with 5squal. *Journal of Informetrics*, Elsevier, v. 3, n. 2, p. 102–123, 2009.

KUBLER, S.; ROBERT, J.; NEUMAIER, S.; UMBRICH, J.; TRAON, Y. L. Comparison of metadata quality in open data portals using the analytic hierarchy process. *Government Information Quarterly*, Elsevier, v. 35, n. 1, p. 13–29, 2018.

ZUIDERWIJK, A.; HELBIG, N.; GIL-GARCÍA, J. R.; JANSSEN, M. Special issue on innovation through open data: Guest editors' introduction. *Journal of theoretical and applied electronic commerce research*, SciELO Chile, v. 9, n. 2, p. i–xiii, 2014.

KIRÁLY, P.; BÜCHLER, M. Measuring completeness as metadata quality metric in europeana. In: IEEE. *2018 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2018. p. 2711–2720.

GAVRILIS, D.; MAKRI, D.-N.; PAPACHRISTOPOULOS, L.; ANGELIS, S.; KRAVVARITIS, K.; PAPATHEODOROU, C.; CONSTANTOPOULOS, P. Measuring quality in metadata repositories. In: SPRINGER. *International Conference on Theory and Practice of Digital Libraries*. [S.l.], 2015. p. 56–67.

ASSAF, A.; TRONCY, R.; SENART, A. Hdl-towards a harmonized dataset model for open data portals. In: *USEWOD-PROFILES@ ESWC*. [S.l.: s.n.], 2015. p. 62–74.

KUBLER, S.; ROBERT, J.; TRAON, Y. L.; UMBRICH, J.; NEUMAIER, S. Open data portal quality comparison using ahp. In: ACM. *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*. [S.l.], 2016. p. 397–407.

CKAN, C. K. A. N. *Metadata - CKAN*. 2017. Https://ckan.org/portfolio/metadata/. Last accessed in December 08, 2019.

SOCRATA. *Sample Metadata Schema*. 2017. Https://support.socrata.com/hc/en-us/articles/115008612447-Sample-Metadata-Schema. Last accessed in December 08, 2018.

PORTAL, E. O. D. *Bulk download of European Open Data Portal full content (metadata records)*. 01 2016. Http://data.europa.eu/euodp/en/data/dataset/bulk-download-of-odp. Last accessed in January 1, 2019.

PILBEAM, K. *Finance & financial markets*. [S.l.]: Macmillan International Higher Education, 2018.

YORK, S. of N. *NY Open Data*. 2005. Https://data.ny.gov/dataset/Open-ny-gov-Catalog/6quf-wz58. Last accessed January 28, 2019.

APPENDIX A – Main Frameworks Metadata Fields

In this appendix is presented a table with all the fields of the metadata schemas of the frameworks CKAN, Socrata and Opendatasoft.

Table A.1: Main Frameworks Metadata Fields Source: Author of This Dissertation

Metadata Schema	Label	Description
Ckan	Title	Allows intuitive labelling of the dataset for search, sharing and linking.
Ckan	Description	Additional information describing or analysing the data. This can either be static or an editable wiki which anyone can contribute to instantly or via admin moderation.
Ckan	Tags	See what labels the dataset in question belongs to. Tags also allow for browsing between similarly tagged datasets in addition to enabling better discoverability through tag search and faceting by tags.
Ckan	Unique identifier	Dataset has a unique URL which is customizable by the publisher.
Ckan	License	Instant view of whether the data is available under an open licence or not. This makes it clear to users whether they have the rights to use, change and re-distribute the data.
Ckan	API key	Allows access every metadata field of the dataset and ability to change the data if you have the relevant permissions via API.
Ckan	Multiple formats (if provided)	See the different formats the data has been made available in quickly in a table, with any further information relating to specific files provided inline.

Ckan	Groups	Display of which groups the dataset belongs to if applicable. Groups (such as science data) allow easier data linking, finding and sharing amongst interested publishers and users.
Ckan	Data preview	Preview .csv data quickly and easily in browser to see if this is the dataset you want.
Ckan	Revision history	CKAN allows you to display a revision history for datasets which are freely editable by users (as is thedatahub.org).
Ckan	Extra fields	These hold any additional information, such as location data (see geospatial feature) or types relevant to the publisher or dataset. How and where extra fields display is customizable.
Opendatasoft	Title	The title of the dataset.
Opendatasoft	Description	The full description of the dataset (HTML is accepted).
Opendatasoft	Keywords	One or more keywords for the dataset, mostly used to make it easier to find in the portal.
Opendatasoft	Last modification	The last modification date of the dataset (manually set).
Opendatasoft	Publisher	The publisher of the dataset (the name of a person or of an organization).
Opendatasoft	Identifier	Technical identifier of the dataset.
Opendatasoft	License	The license attached to the dataset; should always be filled for any public dataset.
Opendatasoft	Geographic area	The geographical coverage of the data.
Opendatasoft	Language	The language (as a two-letter language code) of the datasets data and metadata.
Opendatasoft	References	One or more links to indicate the references or sources of the dataset.
Opendatasoft	Theme	One or more themes associated to the dataset (Environment...).

Open-datasoft	Attributions	Link of a source of the dataset that should be mentioned for legal reasons (e.g. if the license demands the mention of a specific source or organization).
Open-datasoft	Timezone	Forces the dataset visualizations to use the defined timezone for the date and datetime fields. It avoids the dataset visualizations to depend on the timezone on which the users computer is set.
Socrata	Title	Title helps users discover, select, and differentiate between similar datasets.
Socrata	Description	Description helps users discover, select, and differentiate between similar datasets.
Socrata	Tags	Tags link technical language, secondary categories, and acronyms to your dataset, aiding in user-executed searches.
Socrata	Last Updated	Last updated indicates of the recency of the data. Helps users determine usage of data.
Socrata	Contact Email	Consider including publicly-visible Contact Email on each dataset, which can be used by users to ask questions.
Socrata	Unique Identifer	A Unique Identifier is required for dataset management.
Socrata	Public Access Level	While most data on the platform will be public, Public Access Level gives us a means to track protected or sensitive data and provide a means for internal users to discover and access non-public data.
Socrata	Agency / Department	Responsible Agency/Department is helpful for navigation and to ensure a single responsible party.
Socrata	License / Rights	A License reduces legal uncertainty for data consumers or users.
Socrata	Geographic Unit	Geographic Unit indicates the geographic level at which the dataset is collected; also helps track the need to aggregate or summarize data.
Socrata	Temporal Coverage	Temporal Coverage provides an easy way to determine the value of a dataset.

Socrata	Data Dictionary	A Data Dictionary is essential to understanding how the data can be used. It can describe fields, differences between fields, and assess whether or not the data is appropriate for the intended use. Data Dictionaries could be published in both .csv and .pdf format.
Socrata	Permalink / Identifier	A Permalink helps provide continuity for accessing the dataset.
Socrata	Related Documents	Linking a Related Document provides the opportunity to include forms or other types of documents to help users understand the data. Not all datasets will have this information.
Socrata	Category	Category groups similar datasets together regardless of source and can be used to locate similar datasets.
Socrata	API Endpoint	An API Endpoint facilitates programmatic access to the data.
Socrata	Frequency of Data Change	Frequency - Data Change works together with the publishing frequency and helps set expectations for future updates as well as aids in planning.
Socrata	Frequency of Publishing	Frequency - Publishing works together with the Data Change frequency and helps set expectations for future updates as well as aids in planning.
Socrata	Public Access Level Comment	If the data is not public, consider providing an explanation and a means for people to access it if eligible.
Socrata	Data Steward	Consider including a Data Steward for each dataset to support the data coordinators and to answer dataset questions. This helps to track and triage data requests.
Socrata	Row Count	Row Count is a useful indicator of dataset size.
Socrata	Download URL	A Download URL provides access to the data for the purpose of open data.
Socrata	Link	A Link can provide more information on the origin of the dataset. Not all datasets will have this information.

ANNEX A – DCAT-AP Metadata Schema Fields

In this annex a table collected from (COMMISSION, 2018) is presented with all the fields of the metadata schema DCAT-AP Metadata Schema.

Table A.1: DCAT-AP Schema Metadata Fields Source: Source: Author of This Dissertation

Label	Description
type	This property refers to the type of the Dataset. A controlled vocabulary for the values has not been established.
title	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name.
description	This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description.
keyword/ tag	This property contains a keyword or tag describing the Dataset.
license	This property refers to the licence under which the Distribution is made available.
update/ modification date	This property contains the most recent date on which the Dataset was changed or modified.
publisher	This property refers to an entity (organisation) responsible for making the Dataset available.
contact point	This property contains contact information that can be used for sending comments about the Dataset.
identifier	This property contains the main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the portal.
access rights	This property refers to information that indicates whether the Dataset is open data, has access restrictions or is not public.
licence type	This property refers to a type of licence, e.g. indicating 'public domain' or 'royalties required'.
rights	This property refers to a statement that specifies rights associated with the Distribution.

spatial/ geographical coverage	This property refers to a geographic region that is covered by the Dataset.
temporal coverage	This property refers to a temporal period that the Dataset covers.
dataset distribution	This property links the Dataset to an available Distribution.
frequency	This property refers to the frequency at which the Dataset is updated.
release date	This property contains the date of formal issuance (e.g., publication) of the Dataset.
language	This property refers to a language of the Dataset. This property can be repeated if there are multiple languages in the Dataset.
landing page	This property refers to a web page that provides access to the Dataset, its Distributions and/or additional information. It is intended to point to a landing page at the original data provider, not to a page on a site of a third party, such as an aggregator.
documentation	This property refers to a page or document about this Dataset.
related resource	This property refers to a related resource.
theme/ category	This property refers to a category of the Dataset. A Dataset may be associated with multiple themes.
format	This property refers to the file format of the Distribution.
download URL	This property contains a URL that is a direct link to a downloadable file in a given format.
source	This property refers to a related Dataset from which the described Dataset is derived.
access URL	This property contains a URL that gives access to a Distribution of the Dataset. The resource at the access URL may contain information about how to get the Dataset.
byte size	This property contains the size of a Distribution in bytes.
checksum	This property provides a mechanism that can be used to verify that the contents of a distribution have not changed
conforms to	This property refers to an implementing rule or other specification.
end date/time	This property contains the end of the period
provenance	This property contains a statement about the lineage of a Dataset.

version	This property contains a version number or other version designation of the Dataset.
version notes	This property contains a description of the differences between this version and a previous version of the Dataset. This property can be repeated for parallel language versions of the version notes.
has version	This property refers to a related Dataset that is a version, edition, or adaptation of the described Dataset.
is version of	This property refers to a related Dataset of which the described Dataset is a version, edition, or adaptation.
linked schemas	This property refers to an established schema to which the described Distribution conforms.
start date/time	This property contains the start of the period
media type	This property refers to the media type of the Distribution as defined in the official register of media types managed by IANA.