FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Development of an Open Data Portal and a Data Space Platform for Maia's Municipality Data Lake

Fábio Cunha Morais

Mestrado em Engenharia Informática e Computação

Supervisor: Mariana Curado Malta

July 28, 2025

Development of an Open Data Portal and a Data Space Platform for Maia's Municipality Data Lake

Fábio Cunha Morais

Mestrado em Engenharia Informática e Computação

July 28, 2025

Resumo

A Câmara Municipal da Maia (CMM) possui um data lake com bastantes conjuntos de dados que abrangem uma grande variedade de informação, e está a pensar aproveitá-los através da criação de duas plataformas de partilha de dados.

Um portal de dados abertos com o objetivo de disponibilizar informação acessível a todos os cidadãos, seguindo standards europeus, e uma plataforma de espaço de dados para o projeto OMEGA-X, que inclui uma infraestrutura federada e uma marketplace de dados e serviços.

O desenvolvimento do portal de dados abertos estava em curso, e uma plataforma contendo alguns conjuntos de dados já tinha sido construido usando o OpenDataSoft. Porém, a metadata pertencente à camara municipal encontrava-se num nível básico, contendo apenas um modelo personalizado atendendo às necessidades do CMM, não cumprindo standards europeus. Adicionalmente, todo o trabalho desenvolvido pela câmara tinha sido feita diretamente pela interface da plataforma, o que, tendo em conta a dimensão do projeto, comprometia a sua sustentabilidade.

Relativamente ao sspaço de dados, a câmara municipal ainda estava numa fase preliminar, explorando o Sovity, o software indicado para o projeto OMEGA-X, e a compreender a sua complexidade.

Esta dissertação tem como objetivo ajudar o CMM em ambas as soluções de partilha de dados. No caso do portal aberto de dados, dar um passo importante na evolução dos metadados, evoluindo de um modelo específico à Maia para uma estrutura alinhada com os standards europeus e que maximiza as funcionalidades disponibilizadas pelo OpenDataSoft, assim como, desenvolver pipelines bem documentas para uma gestão automatizada da plataforma. Para o data space, pretende desenvolver uma prova de conceito usando o Sovity para estudar a sua complexidade.

Os resultados serão, posteriormente, testados. Para o portal de dados abertos, os princípios do FAIR vão ser usados para aferir a qualidade dos dados e metadados e as metricas definindas num estudo publicado pela universidade de Southampton para avaliar o estado do portal. Para avaliar a prova de conceito do data space, três casos de uso foram definidos.

Abstract

The Maia City Council (CMM) owns a data lake containing many datasets covering a wide range of information, and it is planning to leverage this data by creating two data-sharing platforms.

An open data portal with the goal of providing accessible information to all citizens, following European Standards and a Data Space platform for the OMEGA-X project, including a federated infrastructure and a data and service marketplace.

The development of the open data portal was underway, and a platform containing some datasets had already been built using OpenDataSoft. However, the metadata owned by the city council was still at an elementary level, containing only a custom model to attend to the CMM's needs, not complying with European standards. Additionally, all the work carried out by the city council had been done directly on the platform's interface, which, considering the dimensions of the project, compromised sustainability.

For the data space, the city council was still doing preliminary work, exploring Sovity, the software indicated for the OMEGA-X project, and understanding its complexity.

This dissertation aims to assist the CMM in both of these data-sharing solutions. For the open data portal, it seeks to take an important step in the evolution of the metadata, evolving from a single Maia-specific model to a structure aligned with European standards and that maximises the functionalities provided by OpenDataSoft, as well as to develop well-documented pipelines for an automated management of the platform. For the Data Space, it intends to develop a proof of concept using Sovity to study its complexity.

The results will then be tested. For the open data portal, the FAIR principles will be used to assess the data and metadata quality, and the metrics defined in a study published by the University of Southampton to evaluate the state of the portal. To evaluate the data space proof of concept, three use cases were defined.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. The work conducted during this dissertation tackles the following goals:

SDG 11 Make cities and human settlements inclusive, safe, resilient and sustainable

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

SDG 17 Strengthen the means of implementation and revitalise the Global Partnership for Sustainable Development

SGD	**Target**	**Contribution**	**Performance Indicators and Metrics**
11	11.3	The development of a data platform open to all citizens improves participation and collaboration in urban planning and governance.	Proportion of municipal planning processes that incorporate citizen input collected through the open data platform
16	16.6	Publishing institutional data through an open portal supports the development of effective, accountable, and transparent institutions at the local level	Increased citizen satisfaction with public services due to greater transparency.
	16.10	Facilitating access to public information through an open data platform supports the protection of fundamental freedoms and access to information	Number of public sector datasets made available through the platform that were previously only accessible via formal information requests
17	17.17	Data spaces that allow companies and institutions to trade data safely increase public-private partnerships, contributing to sustainable development.	Number of public-private partnerships made through the data space platform.

Acknowledgements

I want to start by thanking my supervisor, Prof Mariana Malta, not only for the continuous guidance but for being comprehensive and kind and reassuring me whenever I felt overwhelmed.

I also want to thank the Maia City Council for the opportunity to participate in this project. It was my first work experience in a real-world environment.

A special thanks to Eng. Pedro Pimenta, with whom I worked closely, for receiving me with open arms and including me in meetings and decision-making. Also, for helping me be more organised. I will not forget the logbooks!

To my family, thank you for the unconditional love. For all the sacrifices made to grant me a good future, and for always showing me how to be a better person.

Finally, I want to thank my friends for always being there for me. For the late-night talks, the nonsense rants and discussions and for pulling me back whenever I felt lost.

To all these people, I offer my sincerest appreciation. This dissertation would not be possible without you.

Fábio Morais

"The heart's not like a box that gets filled up; it expands in size the more you love."

Spike Jonze

Contents

List of Figures

List of Tables

List of Acronyms

AMA	Agency for Administrative Modernisation
API	Application Programming Interface
CMM	Maia City Council
COGD	Central Government Open Data
CRUD	Create Read Update Delete
DCAT-AP	DCAT Application profile for data portals in Europe
DCMI	Dublin Core Metadata Initiative
EC	European Commission
EIF	European Interoperability Framework
FAIR	Findability Accessibility Interoperability Reusability
HTTP	HyperText Transfer Protocol
ICT	Information and Communication Technology
LD	Linked Data
LOD	Linked Open Data
LGOD	Local Government Open Data
NLP	Natural Language Processing
OECD	Organisation for Economic Cooperation and Development
OGD	Open Government Data
REST	REpresentational State Transfer
RDF	Resource Description Framework
UI	User Interface
UN	United Nations
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Chapter 1

Introduction

The project that results in this thesis aims to assist the Maia City Council(CMM) in developing data-sharing solutions for its data lake. This introductory chapter describes CMM's current situation, the motivation for this project, and the problems the project aims to solve.

1.1 Context

Over the years, the Maia City Council has gathered many datasets in a data lake covering a wide range of information, including transport routes, administrative records, public service usage, and environmental data. These datasets present different characteristics. An example is the variation of data formats, from geospatial coordinates to demographic statistics, or the update frequency ranging from yearly to continuous real-time measurements.

CMM plans to develop two data-sharing solutions for this data lake, as shown in Figure 1.1.

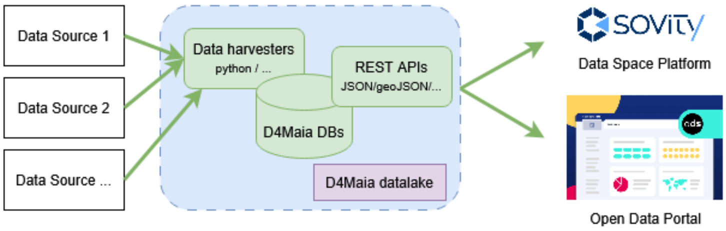


Figure 1.1: Project scheme overview

The first platform consists of a web portal with an open data nature, which aims to provide accessible information to all citizens following European standards. Its development was already underway, with the city council already managing to build a platform, using OpenDataSoft, containing some datasets.

CMM opted for OpenDataSoft due to its continuous user support. Given the limited technical expertise of its members, this factor was considered more important than the flexibility offered by open-source options or the cost of the subscription fee.

The second platform consists of a data space platform for the OMEGA-X project, including a federated infrastructure and a data and service marketplace. Even though the project focuses on the energy sector, it is intended for in the future to extend to other sectors [18].

1.2 Motivation

With the continuous growth of interest in data, the open data movement is gaining power.

This initiative has been proven to bring many benefits. These are particularly clear when considering governmental open data.

In fact, with this project, CMM aims to improve citizen participation and collaboration by allowing them to make better-informed decisions, which can provide the City Council with helpful feedback [10].

On the other hand, with the data space platform, the CMM plans to encourage collaboration between companies and organisations, leading to several benefits, namely, allowing for start-ups and small-sized companies to gain access to essential data they otherwise had no means to collect, allowing for larger companies to earn revenue from possibly unused data, and avoiding redundant processes across companies that could be wasting resources and causing environmental impact.

1.3 Problem

Despite the progress made by CMM in developing the open data portal, the metadata owned by Maia is still at an elementary level. It consists of a single schema created by the city council to meet their needs, considered by the AMA (Agency for Administrative Modernisation) as the first level in the metadata evolution model [2]. Thus, it is necessary to take the next step in the metadata maturation process by aligning it with European standards. It should also be extended to maximise the functionalities provided by OpenDataSoft, the software used to create the portal.

Additionally, all the work carried out by the city council had been done directly on the platform's interface. Given the size of the data lake and the possibility of growing, its sustainability was compromised. It was therefore necessary to create well-documented pipelines for the automated management of the platform.

Regarding the data space project, the city council was still carrying out preliminary work by exploring Sovity[1], an Eclipse Data Connector pointed by the OMEGA-X as the indicated software and so, it considered that developing a proof of concept in this software to understand its complexity should be next step.

1.4 Objectives

The objectives of this dissertation are:

[1] Website accessed on 5/06/2025.

1. Align the metadata maturation process with European standards and maximise the functionalities provided by OpenDataSoft, the software used to create the portal.
2. Create well-documented pipelines for the automated management of the OpenDataSoft platform.
3. Develop a proof of concept using Sovity[2], an Eclipse Data Connector pointed by the OMEGA-X, to understand the complexity of the software.

1.5 Research questions

- How to improve a metadata catalogue based on internal organisational requirements, such as the CMM's, to align with European standards?
- How to develop automated pipelines for managing an open data portal built in OpenDataSoft?

1.6 Methodological Approach

This section presents the methodology adopted for the project. Firstly, the methods used to conduct the literature review are presented, followed by the work methodology.

1.6.1 Introduction

As mentioned in 1.4, this project focuses on three tasks: the maturation of the Maia City Council's metadata to meet European standards and maximise the functionalities provided by OpenDataSoft, the creation of well-documented pipelines for the automated management of the OpenDataSoft platform and the development of a proof of concept for a data space platform using Sovity.

It was therefore important to establish a method to conduct the literature review.

At the time of this process, these tasks hadn't been defined yet. In fact, the project was still being studied with the CMM, so it was necessary to later adapt this methodology.

1.6.2 Literature Review Method

The literature review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, a widely recognised methodology for conducting systematic reviews. While PRISMA includes 27 items, only those from the methods section were considered relevant here, as the introduction is covered in Chapter 1, and the remaining sections do not apply to the literature review. Specifically, items 5 through 9 from the methods section were selected as essential for this review.

[2] Website accessed on 5/06/2025.

1.6.2.1 Item 5: Eligibility Criteria

The following eligibility criteria were established to ensure the relevance and accessibility of the selected documents for this review.

- Documents from 2001 to now
- Full-text access documents
- Published in English
- Articles from journals & conferences, or Books chapters or books.

1.6.2.2 Item 6: Information Sources

The documents were searched using the following databases:[3]

- IEEE Xplore
- Scopus
- Science Direct
- ACM Digital Library

1.6.2.3 Item 7: Search Strategy

Three queries were defined and executed in each database to cover the different dimensions of the dissertation. They are:

- Q1: **(Local OR Municipal) AND Government AND Open Data AND (Platform OR Portal)** - This query was defined as a general approach to address the initial project focused on creating an open data platform for municipal government. A quick analysis revealed that "local" and "municipal" are commonly used in this context.
- Q2: **Government AND Data Marketplace Platform** - This query was defined to support the second project, aimed at developing a platform that serves as a government data marketplace for the OMEGA-X project.
- Q3: **(Local Or Municipal) AND Government Data AND (Visualization OR Display)** - This query was defined to explore the principles of data visualisation and display. Initially, it was meant to search only in the context of government data. However, due to the excessive number of results, it was changed to concentrate specifically on local government data. As with the initial query, "local" and "municipal" are used interchangeably in this context.

The searches were conducted using the titles and abstracts of the papers.

[3] All websites were accessed on 31/10/2024.

1.6.2.4 Item 8: Selection Process

Searches were carried out on the sources referred to in section 1.6.2.2 using the queries mentioned in section 1.6.2.3 and the criteria defined in section 1.6.2.1. The eligibility step was done by manually combining filters for each database.

The results are shown in Table 1.1.

	Q1	Q2	Q3	Total
IEEE	26	5	49	80
Scopus	222	32	9	263
Science Direct	5	26	123	154
ACM	18	33	16	67
Total	**271**	**96**	**197**	**564**

Table 1.1: Preliminary Results of the Literature Search

The resulting documents then underwent a selection process where abstracts were reviewed to categorize each document as ’interesting,’ ’unrelated,’ or ’uncertain.’ Documents marked as ’uncertain’ were fully reviewed to make a final inclusion decision. Each document was selected based on specific query criteria. For the first query, documents are needed to discuss the process of creating, maintaining or evaluating open data platforms or comparisons of software used in building such platforms. The second query focused on documents addressing the creation of a data marketplace platform within a government context. The third query targeted documents that provide recommendations for displaying datasets, including requirements, regulations, and guidelines. Documents focused on highly specific data types outside the context of Maia’s data lake were excluded.

The final results are presented in Table 1.2.

	Q1	Q2	Q3	Total
IEEE	6	1	9	16
Scopus	30	1	4	35
Science Direct	1	1	1	3
ACM	8	3	0	11
Total	**45**	**6**	**14**	**65**

Table 1.2: Results of the Literature Search

1.6.2.5 Item 9: Data Collection Process

All files were downloaded and imported into EndNote, organised into groups according to the source. During import, duplicates across sources were detected and removed. In EndNote, each entry contained the abstract and DOI, providing easy access to the full version.

1.6.2.6 Method adaptation

After the final definition of the goals, the results for queries 2 and 3 were both discarded. To obtain the necessary information for the development of the project, different approaches were used.

To obtain the necessary theoretical context on metadata, the book [30] was used.

For the refinement of CMM's metadata, the AMA guide to open data was used [2], since the AMA is the public institute responsible for leading the digital transformation, including the implementation of open data portals in Portugal. In addition, some studies, published in the Publications Office of the European Union [11] and the metadata documentation of the OpenDataSoft software, were used to address this task.

The documentation of this software was also used for the development of the automated pipelines for the management of the portal.

Finally, in terms of the data space project, the focus was to study the Sovity software, so its documentation was used.

1.6.3 Work Methodology

A five-step process was undertaken to develop the solutions for the Maia City Council. These steps included integration within the CMM, an analysis of the current situation, the development and documentation of methods and solutions, and finally, an evaluation of the results. Each step is detailed in the following subsections.

1.6.3.1 Integration

The starting point involved integration with the Maia City Council. This step consisted of meetings to introduce the members, present the CMM's situation, define objectives for the project, demonstrate the software, share permissions and define the communication methods.

1.6.3.2 Research Stage

Once the integration was concluded and the objectives for the project were outlined, it was necessary to adapt the research approach to the new needs. This process is described in 1.6.2.6.

1.6.3.3 Development

With the research stage concluded, it was time to develop the solution. This process was conducted in proximity to the CMM. Weekly meetings were scheduled to present the developed work, validate it, and discuss specific next steps. All documentation, including investigation, development and meeting notes, was kept in a logbook.

The specific description of the development carried out for this project is presented in 3.

1.6.3.4 Evaluation

Finally, the developed work underwent an evaluation phase. This evaluation combined metrics and topics addressed in the research stage. Suggestions to address the exposed limitations were then provided for future work.

1.7 Summary

The Maia City Council (CMM) has an extensive data lake and is looking to develop two data-sharing solutions to utilise it. An open data portal aiming to provide accessible information for all citizens and a data space platform including a federated infrastructure and a data and service marketplace for the OMEGA-X project.

This dissertation presents the work developed to assist CMM in this development by improving Maia's metadata, evolving from a single Maia-specific schema to a structure aligned with European standards and that maximises the functionalities provided by OpenDataSoft, the software used in the creation of the portal, developing well-documented pipelines for the automated management of the platform, and develop a proof of concept using Sovity to understand the complexity of this software.

The next chapter presents the Theoretical Background and the State-of-the-art.

Chapter 2

Background and State of the Art

This chapter is divided into a background section and the state of the art.

The background provides a brief insight into the history of open government data and a more theoretical description of concepts such as interoperability, the FAIR principles, and metadata.

Based on these concepts, the state of the art explores the requirements of this type of platform and provides a list of metrics to evaluate its state. Finally, it compares OpenDataSoft to other open data portal software based on the required functionalities to determine if it suits the platform's development.

2.1 Background

2.1.1 Open Government Data Portal

Interest in open government data (OGD) has been growing since President Obama's open data initiative in 2009. With movements like the Open Government Partnership in 2011 and the G8 Open Data Charter in 2013 [5]. Government agencies, organisations, and citizens are realising the usefulness of OGDs. Many International organisations like the European Commission (EC), the Organisation for Economic Cooperation and Development (OECD), and the United Nations(UN) have developed guidelines and directives to promote open data policies [25].

The responsibility for developing these data portals has changed over time. Initially, they were almost reserved for central governments as central government open data (COGD) was the main focus. However, local governments have recently launched independently operated portals through cities, municipalities, counties, federal states, regions, and provinces. [1] argues that local government open data (LGOD) should be given the same attention as the COGD, as most public datasets are generated at the level of local authorities or agencies.

2.1.2 Interoperability

Interoperability is defined by the European Interoperability Framework(EIF) in [7] as "the ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, using the exchange of data between their Information and Communication Technology(ICT) systems".

The EIF is built upon twelve fundamental principles to establish interoperable European public services. These principles are grouped into four categories.

1. Principle setting the context for EU actions on interoperability.
2. Core interoperability principles.
3. Principles related to generic user needs and expectations.
4. Foundation principles for cooperation among public administrations.

For the development of European public Web Portals, the third group is essential as it describes the base requirements that this type of platform should address. This group includes the following principles:

- **User-Centricity** - User needs and requirements should drive the design and development of public services. The EIF recommends using multiple channels to deliver European public services, allowing users to choose the most appropriate option. It also refers to the importance of providing a single point of contact, hiding internal complexity, facilitating user access, and making available mechanisms that involve users in analysis, design, assessment, and further development. Additionally, users should only be asked to provide necessary information.
- **Inclusion and Accessibility** - All European public services should be accessible to every citizen, including persons with disabilities, the elderly, and other disadvantaged groups. These services should comply with e-accessibility and be widely recognised at European and international levels.
- **Security and Privacy** - A standardised security and privacy framework should be defined, along with processes that guarantee secure and trustworthy data exchanges between public administrations and interactions with citizens and businesses.
- **Multilingualism** - Information systems and technical architectures that support multilingualism should be used when establishing a European public service. Additionally, the level of multilingual support should be determined based on the needs of the expected user base.

2.1.3 FAIR principles

The FAIR principles, [29], were created by diverse stakeholders, including academia, industry, funding agencies, and scholarly publishers, as guidelines for improving the reusability of their data. It focuses on machine-actionability as humans rely on computational support to deal with data.

They have four foundational pillars: Findability, Accessibility, Interoperability, and Reusability [14]. Each of these principles is accompanied by a set of sub-principles that provide instructions on achieving the corresponding goal.

2.1.3.1 Findability

Findability refers to the ability to find data and metadata for both humans and computers. The following sub-principles support this:

- **F1 - (Meta)data are assigned a globally unique and persistent identifier**

  Each dataset should have a globally unique and persistent identifier, such as a DOI. This guarantees that data and metadata can be reliably found and cited. So datasets that use non-unique identifiers, like titles, do not meet this principle.

- **F2 - Data are described with rich metadata (defined by R1)**

  Each dataset should be described with rich metadata. This helps users understand and select the dataset that best fits their needs. So datasets with missing contextual metadata do not meet this principle.

- **F3 - Metadata clearly and explicitly include the identifier of the data they describe**

  Metadata should include the identifier of the data, enabling users to locate the dataset easily. Datasets that separate metadata and data without clear links fail to meet this principle.

- **F4 - (Meta)data are registered or indexed in a searchable resource**

  Existing metadata should be used to power search and filtering systems, helping users perform effective searches. So, even if the metadata is rich, if a platform doesn't integrate it into search systems, it does not meet this principle.

2.1.3.2 Accessibility

Accessibility refers to the ability of humans and machines to easily retrieve and use data and metadata through standard methods and protocols, guaranteeing long-term availability. The following sub-principles support this:

- **A1 - (Meta)data are retrievable by their identifier using a standardised communications protocol**

- **A1.1 - The protocol is open, free, and universally implementable**

  When a user knows a dataset's identifier, they should be able to access it using an open, free and universal protocol such as HTTP or FTP. So, platforms that use country-specific communication protocols do not meet this principle.

- **A1.2 - The protocol allows for an authentication and authorisation procedure, where necessary**

  Platforms may require users or machines to sign in if they can authenticate successfully. If authentication must be done directly through a user interface, preventing machines from authenticating automatically, the platform does not meet this principle.

- **A2 - Metadata are accessible, even when the data are no longer available**

  Datasets whose data becomes inaccessible should still keep their metadata. This guarantees that even if links become invalid, users do not waste time searching for no longer available data. The dataset does not meet this principle if the metadata is lost when the data disappears.

2.1.3.3 Interoperability

Interoperability refers to the ability of data to be universally accessible, reusable, and understandable by both humans and machines. The following sub-principles support this:

- **I1 - (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**

  Humans should be able to exchange and interpret data using commonly understood languages. Similarly, machines should read data without requiring specialised algorithms, translators, or mappings. So, if a company wants to publish datasets in a data space platform, but the data format must be adapted first, it does not meet this principle.

- **I2 - (Meta)data use vocabularies that follow FAIR principles**

  The controlled vocabularies used to describe datasets should be appropriately documented and made easily findable and accessible to anyone using the dataset. So, if a dataset uses internally undocumented vocabularies, it does not meet this principle.

- **I3 - (Meta)data include qualified references to other (meta)data**

  Datasets that depend on, extend, or complement other datasets should describe these relationships and cite the related datasets using persistent identifiers. So, if a dataset results from merging two other datasets but does not reference them, it does not meet this principle.

2.1.3.4 Reusability

Reusability refers to how well data and metadata are described to support their use in future research and integration with other compatible sources. It also requires clear citation information and usage conditions that are understandable by both humans and machines. The following sub-principles support this:

- **R1 - (Meta)data are richly described with a plurality of accurate and relevant attributes**

- **R1.1 - (Meta)data are released with a clear and accessible data usage license**

  The conditions under which the data can be used should be clearly stated in the metadata, in a way that is understandable to both humans and machines. So, a dataset where the license is missing or written only in free-text without machine-readable terms does not meet this principle.

- **R1.2 - (Meta)data are associated with detailed provenance**

  Detailed information about the provenance of data is essential for reuse, as it helps others understand how the data was generated, the context in which it can be reused, and how reliable it is. So, a dataset that does not document its origin, collection methods, or processing steps does not meet this principle.

- **R1.3 - (Meta)data meet domain-relevant community standards**

  Datasets should follow established community standards for data structure, formats, and documentation to confirm that they can be easily reused and integrated with similar datasets. This includes using common file types, shared metadata templates, and accepted vocabularies. So, if a dataset uses uncommon formats or lacks standard metadata, making reuse difficult, it does not meet this principle.

2.1.4 Metadata

The concept of metadata has changed over time, especially with the advent of the Internet and the Web of Data. Metadata went from simple definitions such as "data about data" [22], into more complex and specific forms such as "data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation." [8], as described in Zeng and Qin [30].

Specifically, in terms of open data, AMA refers to metadata as being responsible for adding all the descriptive information about the data by describing its content, format, limitations, frequencies, and updates to a dataset. In the open data guide [2], this agency adds that this information is essential for both users and computers, and, when well structured, it is decisive for the interoperability of the dataset.

AMA then presents a metadata evolution model consisting of four stages. Before analysing them, it's first important to define some concepts:

1. **Metadata Catalogue** - Curated collection of metadata about resources, for example, datasets [6].

2. **Metadata Schema** - Set of metadata elements defined for a specific purpose, within a given context. A metadata schema is also referred to, in some circles, as a “vocabulary” or an “RDF vocabulary.” [19]

3. **Application Profile** - Describes how a standard should be applied in a particular domain or application by containing constraints, like cardinality or possible values, that standards typically don't have. [26]

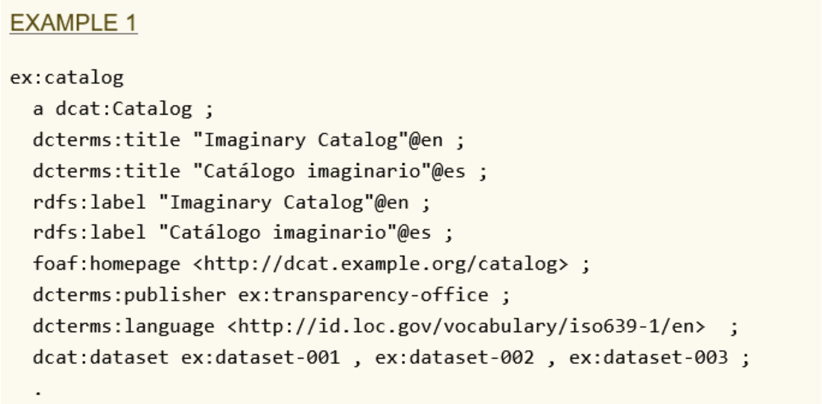
With these concepts well defined, the stages can now be examined:

1. **Internal Standard** - Even though AMA describes it as an internal standard, the expression internal schema would be more suitable, as standard normally implies a formal specification, such as an ISO standard. In this stage, the entity develops a metadata schema around logistics and needs, compromising interoperability and reusability.

   As mentioned in 1.4, this is the current state of the CMM's metadata.

2. **Dublin Core** - A more sophisticated metadata schema that describes digital objects based on XML and Resource Description Framework (RDF). It was published by the Dublin Core Metadata Initiative (DCMI) to facilitate search processes and information retrieval.

3. **DCAT** - Developed by the W3C, DCAT consists of an RDF vocabulary for representing data catalogues. Based around seven main classes: Catalog, Resource, Dataset, Distribution, DataService, DatasetSeries, and CatalogRecord. DCAT combines more generic Dublin Core terms like title, publisher, and language with new terms. Figure 2.1 shows an example in turtle format extracted from Browning and Albertoni [6].



EXAMPLE 1

```
ex:catalog
  a dcat:Catalog ;
  dcterms:title "Imaginary Catalog"@en ;
  dcterms:title "Catálogo imaginario"@es ;
  rdfs:label "Imaginary Catalog"@en ;
  rdfs:label "Catálogo imaginario"@es ;
  foaf:homepage <http://dcat.example.org/catalog> ;
  dcterms:publisher ex:transparency-office ;
  dcterms:language <http://id.loc.gov/vocabulary/iso639-1/en>  ;
  dcat:dataset ex:dataset-001 , ex:dataset-002 , ex:dataset-003 ;
  .
```

Figure 2.1: DCAT example in turtle format

4. **DCAT-AP** - Consists of a specification based on DCAT for describing European data portals. It specifies the mandatory properties and controlled vocabularies such as those for file types and licenses, as explained in European Commission [9]. Some countries are adhering to national application profiles. As Portugal doesn't yet own a national specification, AMA considers DCAT-AP to be the goal in evolving metadata for open data projects in Portugal.

2.2 State of the Art

2.2.1 OGD portals Requirements

Requirements are frequently discussed when evaluating the current state of e-government websites. This is particularly relevant to Open Government Data (OGD) portals ([20, 25]) but also applies to other types of platforms ([16, 17]). While both share several aspects, there are points of divergence. This section will focus specifically on OGD portals.

Most studies reference frameworks or guidelines from organisations such as the W3C and the World Bank [1]. In particular, the University of Southampton published some studies as part of the European Data Portal, an initiative of the European Commission, built upon these guidelines. The first study, [15], presents the list of requirements for OGD portals, while [13] complements by proposing metrics to evaluate them using a score system. Next up is the list of requirements and their metrics:

1. **Organise for Use** - The portal should be designed to focus on users and their experience. One point is awarded for each of the following items:

   (a) Each dataset is accompanied by a comprehensive descriptive record (going beyond a collection of structured metadata).

   (b) An extract of the data can be previewed (for easier sense making).

   (c) The portal provides recommendations for related datasets.

   (d) The portal enables users to review/rate datasets.

   (e) Keywords from datasets are linked to other published datasets.

2. **Promote Use** - The portal should adopt mechanisms to actively encourage users to engage with its datasets, with promotion efforts prioritising users rather than data publishers, as is often the case. One point is awarded for each of the following items:

   (a) The portal is connected with social media to create a social distribution channel for open data.

   (b) The portal provides users with online support for feedback, to request/suggest the publication of new datasets, and when problems arise during use (e.g. contact form, discussion forum, FAQs, helpdesk, search tips, tutorials, demos).

   (c) The portal provides a way for users to keep informed of updates to the data (e.g. news feed).

   (d) Datasets are accompanied by links or resources that provide user guidance and support.

   (e) Examples of reuse (fictitious or real) are provided (e.g. information contributed by other users, last reuse, best reuse, data stories).

3. **Be Discoverable** - The portal must follow good practices to ensure datasets can be easily found. One point is awarded for each of the following items:

   (a) The publisher/owner of the data has an open data portal (or similar search mechanism).

   (b) The publisher/owner of that portal publishes an updated, searchable list of datasets.

   (c) The publisher/owner of that portal publishes an updated, searchable list of datasets with synonyms.

   (d) The publisher/owner of that portal publishes a list of datasets which are known to exist but are not currently available (limiting the time wasted on abortive searches).

4. **Publish Metadata** - Publishing good quality metadata is essential to improve dataset reuse, support findability, enable proper cataloguing, and allow connections between related datasets. The following levels reflect the maturity of metadata publication, with one point awarded for each level achieved:

   1. Metadata Ignorance.
   2. Scattered or Closed Metadata.
   3. Open Metadata for Humans.
   4. Open Reusable Metadata.
   5. Linked Open Metadata.

5. **Promote Standards** - Adopting common standards for datasets and metadata is essential to ensure interoperability, enable cross-portal searches, and facilitate seamless data exchange between portals. One point is awarded for each of the following items:

   (a) A permanent, patterned and/or discoverable URI/URLs is used for each dataset (e.g. URI/URLSs can be used as universal, unique identifiers by appending a serial number or other internal naming system to a domain).

   (b) The portal uses versioning of datasets (to maintain the history of a dataset)

   (c) Dates are available in a standard format (facilitates the automated exploitation of date-type data and their conversion according to specific needs or constraints).

   (d) Metadata associated with each dataset is available in a standard format (e.g. using VOID or DCAT) to enable automated metadata retrieval and import of metadata from other data catalogues.

   (e) The metadata catalogue can be retrieved using a standard protocol (e.g. automatic retrieval of the metadata catalogue using RDF or HTTP GET).

6. **Co-Locate Documentation** - Documentation should be easily and immediately accessible directly from the dataset and context-sensitive. This approach allows users to access specific

information without needing to search for it, simplifying the process and improving access to relevant materials. The following levels reflect the availability of support information, with one point awarded for each level achieved:

   1. Supporting documentation does not exist.
   2. Supporting documentation exists, but as a document which has to be found separately from the data.
   3. Supporting documentation is found at the same time as the data (e.g. the link to the document is next to the link to the data in the search).
   4. Supporting documentation can be immediately accessed from within the dataset, but it is not context sensitive (e.g. a link to the documentation or text contained within the dataset).
   5. Supporting documentation can be immediately accessed from within the dataset and it is context sensitive so that users can immediately access information about a specific item of concern (e.g. a link to a specific point in the documentation or the text contained within the dataset).

7. **Link Data** - Linking datasets is essential for enabling cross-referencing, combining information from multiple sources, and improving data analysis by connecting related or external datasets, including previous versions and relevant recommendations. The following levels reflect the maturity of linked data implementation, based on Tim Berners-Lee's 5-star scheme for Linked Open Data, with one point awarded for each level achieved:

   1. **On the Web**: Make your stuff available on the Web (whatever format) under an open license.
   2. **Machine-readable data**: Make it available as structured data (e.g. Excel instead of image scan of a table).
   3. **Non-proprietary format**: Make it available in a non-proprietary open format (e.g. CSV instead of Excel).
   4. **RDF standards**: Use URIs to denote things, so that people can point at your stuff.
   5. **Linked RDF**: Link your data to other data to provide context.

8. **Be Measurable** - Open data portals should be measurable to assess how effectively they meet users' needs. The following levels reflect the maturity of measurement practices, with one point awarded for each level achieved:

   1. Portal has No analytics.
   2. Portal has Site analytics.
   3. Portal has Use analytics.

4. Portal has Impact analytics.

9. **Co-Locate Tools** - Co-locating tools are essential to engage a broader range of users with the datasets on an open data portal. Mapping and visualisation tools, in particular, can significantly improve an individual's ability to explore a dataset and assess its relevance. The following levels reflect the extent to which the portal integrates tools that support data exploration and user collaboration, with one point awarded for each level achieved:

    1. The portal does not provide visualisation or collaboration tools for users to engage with the datasets.
    2. The portal provides visualisation tools to enable users to engage with the datasets.
    3. The portal provides visualisation and collaboration tools to enable users to participate in the governance of the portal (e.g. dataset rating), but the engagement with other users is limited or mediated by the administrator.
    4. The portal provides visualisation and collaboration tools to enable users to collaborate innovatively with other users.

10. **Be Accessible** - Accessibility is one of the core principles of Open Government Data. Data should be available to the widest possible range of users and supported for various purposes. One point is awarded for each of the following items:

    (a) The portal uses human and machine-readable and non-proprietary formats (e.g. CSV, XML, RDF-based formats).

    (b) The portal provides different types of formats for the same dataset.

    (c) The mechanisms for accessing and interacting with datasets are documented.

    (d) Multilingual support is available on the portal.

    (e) The portal supports the visually and hearing impaired.

The study also included an assessment chapter where ten European government data portals were evaluated using these metrics. Dados.gov Portugal [21], the open data portal of the Portuguese public administration, and EU Open Data Portal [12], the official portal for European data, were among these portals, scoring 28 and 33 out of a possible 47 points, respectively.

2.2.2 Open Data Platforms software

According to [20, 28], the Comprehensive Knowledge Archive Network (CKAN) and DKAN, a Drupal-based platform built on top of CKAN, are the most popular platforms for creating open data portals. [4, 27] also present Socrata as a widely used platform.

It is, therefore, important to compare ODS to these three software. A feature comparison is presented in Table 2.1. This table is adapted from the one in [4], but it only includes the suitable platforms and required functionalities.

Feature Name	Description	CKAN	DKAN	Socrata	ODS
Metadata	Allows metadata standards such as DCAT-AP.	+	+	+	+
Searchability	Allows searches by location, data title, description, or topic.	+	+	+	+
API	Offers APIs such as SPARQL and RESTful.	+	+	+	+
Machine-readable data	Compatible with Berners-Lee 5-star linked data.	+	+	+	+
Open Data Standard compliance	Provides support for Open Data Standards.	+	+	+	+
Data Visualisation	Provides tools for data preview and visualisation.	+	+	+	+
Licensing	Allows for most non-proprietary data licenses.	+	+	+	+
User Support	Has proper user support.	+	+	+	+
Open Data Categorisation	Different datasets are categorised into distinguished types.	+	+	+	+
Data linkage	Provides data linkage based on vocabulary and metadata standards such as the DCAT-AP metadata standard.	+	+	+	+
CMS Platforms	Contains an integrated CMS to manage the content inside the open data platform.	-	+	-	+
Data formats	Supports multiple formats including .pdf, .csv, RDF, and LOD.	+	+	+	+
User Interface Technologies	Uses: ReactJS and a Bootstrap-based GUI.	+	+	+	+

Table 2.1: Feature comparison of open data platform software

The table shows that regarding the selected features, the only difference between these four software platforms is that only DKAN and ODS include an integrated CMS. Therefore, although these two may support a more straightforward page customisation, all four software are valid options.

2.3 Final Considerations

In this chapter, several theoretical concepts were presented, alongside processes, evaluation metrics and requisites. These elements support the development of the project presented in the next chapter.

Chapter 3

Development

This chapter is divided into the development of the open data portal and the data space proof of concept.

3.1 Open Data Portal

As mentioned previously in 1.4, the open data portal for Maia was an ongoing project, and despite the work already carried out it was time to take an important step in the evolution of the metadata owned by the CMM, going from a single Maia-specific schema to a structure aligned with European standards and that maximised OpenDataSoft's functionalities. Additionally, it was necessary to create well-documented pipelines for automated management.

3.1.1 Metadata

3.1.1.1 Metadata Schemas Selection/Creation

In order to maintain compatibility with OpenDataSoft, the selection process was carried out using the features of this software. It was, therefore, important to understand how OpenDataSoft handled metadata.

This software allowed for four types of metadata: **Basic** and **Admin**, corresponding to public and private metadata models, respectively, that could be created by the platform's admins; **Interoperability**, consisting of a list of metadata schemas and application profiles defined by regulatory entities such as the European Commission and available for selection to be featured on the portal; and **Applicative Metadata** required for the internal functionalities of the portal.

Opendatasoft included two mandatory metadata models that could not be modified: **Standard**, a metadata model responsible for basic information such as dataset titles and descriptions, as well as fields required for features like filters and recommendation systems; and **Internal ODS Metadata**, an Applicative metadata model necessary for the internal mechanisms of the portal. This last model was only important for the ODS team, not being featured in the documentation and was, therefore, ignored.

For interoperability, the **DCAT-AP** was selected as it is the appropriate AP for European datasets, as explained in 2.1.4.

Finally, some fields of the old metadata catalogue provided by CMM were still not present on the new catalogue, fields essential for data management. These were compiled to create **D4Maia**, a custom metadata model for Maia. This model was created as Admin Metadata, as the information should only be visible to the CMM members.

The Table 3.1 presents side by side the specifications with a brief description for each field.

Table 3.1: Correspondence table between initial version of D4Maia, Standard ODS schema and DCAT-AP

D4Maia	Standard ODS	DCAT-AP	Description
nome	technical_identifier		ID of the dataset. Used for both the explore API and for the URL of the dataset.
descri	title	title(inherited[1])	Title of the Dataset.
descriplus	descri	descri(inherited[1])	Brief text describing the dataset.
dataultimaactuallocal	modified		Date of the last modification of the dataset's data or metadata.
comm			Comment about the dataset. Normally used to identify source datasets from which this dataset was derived.
editor			Information regarding the person within the CMM responsible for the dataset.
pcnome			Name of the person within the CMM responsible for the dataset. This value was normally included in the field "editor".
pcemail			Email of the person responsible for the dataset. This value was normally included in the field "editor".
		contact_name	Generic name of the department in the CMM responsible for the data.

		contact_email	Generic email of the department in the CMM responsible for the data.
fonte	publisher	publisher(inherited)	Entity that published the dataset, often, the organisation responsible for making the data available.
origem			Original source of the data. For example, if the dataset was published by a ministry based on census data, the ministry is the *fonte* and the census is the *origem*.
licenca	license		License of the dataset that indicates how it can be used.
licencaurl			Link to the license page.
tema	theme	theme	Broad categories or topics that group datasets, such as Health, Population or Finances.
formacalculo			Method or formula used to calculate or derive the dataset values.
medidaanalise			Unit of measurement used in the dataset.
divterrit	geographic_reference		Territorial division covered by the dataset. Not needed when the dataset contained georeferenced data.
periodactual	update_frequency		How often the dataset's data or metadata is updated. Useful for API listeners to know when to expect new information.
nivelacesso			Access level of the dataset, such as public or private.
formaactual			How the data is collected into the data lake, with values, manual or automatic.
		accrual_periodicity	How often the dataset is published.

tipodefonte		publisher_type	Type of source entity, such as governmental, non-profit, or academic.
primapref		temporal_startDate	Start date of the data.
ultimopref		temporal_endDate	End date of the data.
	keywords	keywords(inherited)	Specific terms describing the dataset content to improve search and discovery.
	reference		Link to the dataset stored in the data lake.
	attributions		References to source datasets from which this dataset was derived. It should always point to the data lake, as these are also kept there.
	metadata_language		Language used in the metadata.
	timezone		Timezone used in both and metadata.
	federated		Indicates if the dataset is part of a federated data system. Should always be false as all datasets are centralised in Maia's datalake.
		spatial_bbox	Automatically generated geographic bounding box of the dataset. Not editable. Included in metadata exports.
		spatial_centroid	Automatically generated central point of the dataset's geographic coverage. Not editable. Included in metadata exports.

1- The keyword "inherited" marks fields that, because they exist in both the Standard and DCAT-AP schemas, are defined only in the first schema, and the second presents a pointer to the value.

By analysing Table 3.1, it is visible that *nome* corresponds to *technical_identifier*, *descri* matches *title*, and *descriplus* aligns with *description*. Although the names of the fields initially suggested a different mapping, a closer look at their actual content justified this interpretation.

The field *nome*, despite its label, contained abbreviations with numbers, which are typical of identifiers. On the other hand, *descri* presented short, simple text characteristic of a title. This initial mismatch had to be corrected later in the process.

It is also clear that all text fields in the Standard ODS and DCAT-AP models have corresponding fields in D4Maia. At this stage of development, the intention was to keep the private metadata in Portuguese and the public metadata in English. This approach guaranteed that, during the metadata flow implementation, both languages would be considered, making it easier to adapt the platform to its final goal of supporting both Portuguese and English in the front and back office.

3.1.1.2 Metadata Database Adaptation and Improvement

After the selection/creation of the metadata catalogue, it was necessary to assess the metadata database and understand the necessary changes to account for the new requirements. A table was created to support this work by mapping the different columns of the database to the required fields of each specification, as well as a brief description for each field. For single-select fields, the possible values were also specified to ensure consistency. One iteration of the table can be found in Appendix A.

First, it was important to clarify the purpose and the content of each column. While some columns were more straightforward due to clear naming or easily interpretable content, others were more complex. This was the case, for example, with the columns *dataultimaactual*, *dataultimaactuallocal* and *dataultimaverifica*, whose names were very similar, even overlapping, and whose content, consisting of a single date, was not sufficient to determine their specific meaning. Additionally, certain columns like *SubProcesso*, *RegDate*, or *ObjEstrat* were mostly populated with null values, lacking sufficient information to decipher their meaning. This step was carried out in close collaboration with people from the CMM, who helped clarify the more complex columns and confirm the rest.

These columns were afterwards mapped to the corresponding fields in each metadata specification. From there, it was important to analyse which fields existed in the catalogue but had no direct correspondence in the database, in order to understand whether they could be derived from existing data or, in cases such as *keywords* and *accrual_periodicity*, if new columns needed to be created. At the same time, the fields in the database that were not being used were analysed to determine if their information was relevant for data management and should be included in the D4Maia schema, like the column *status*, or whether they were relevant for citizens and therefore could be incorporated into fields of the public schemas. An example of this was the column *formacalculo*, whose content, after analysis, was appended to the description field.

A special case was the column *Comm*. After analysing its content, it became clear that, even though the column was intended for CMM members to write brief and informal comments regarding the datasets, in many cases, it contained relevant information regarding the origin of the data. More specifically, in cases where the dataset was derived from another dataset, *Comm* was the only column that mentioned the parent. Since the field *Attributions* from the Standard ODS schema required this information, the content of the column was repurposed. In the case of derived datasets, and because the CMM always kept the original datasets in the Data Lake, even when it only intended to publish the derived ones on OpenDataSoft, a new notation was used to refer to the originals. This notation consisted of writing the *sigla* of the original dataset in between dollar signs, allowing it to be programmatically replaced by the original's *name* and appended to the description. Additionally, the *reference* of the original dataset would be added to the *Attributions*. With this in mind, the *Comm* column had to be cleaned, and any informal comments moved to the *ObserIntern* column.

Finally, new columns such as *ID_ODS* and *ODS2up* were created to support operations related to OpenDataSoft.

The final iteration of the mapping table can be consulted in Chapter 4, Table 4.5.

During that period, CMM was also looking to create a shared Google Sheet to serve as a workspace for its members. The goal was to simplify metadata updates for members less familiar with databases, including those who did not understand SQL, as well as to improve the traceability of changes. The members would make updates in the sheet, while a script developed by CMM would automatically synchronise these changes with the metadata database.

The structure of the Google Sheet was already adapted to accommodate the new database changes. A colour-coding system was implemented for the columns to indicate which schemas each column would populate. Additionally, single-select fields were limited to predefined dropdown values, as specified in the mapping table. Auxiliary columns created to support OpenDataSoft operations were restricted to a dedicated account, accessible only by the scripts described in 3.1.2.

After the structural alignment, it was time to assess the metadata itself. Both single-select and textual columns were considered. In the case of single-select fields, the process was simpler, as it only required reviewing all existing values and ensuring that they were all contained in the array of values for that column. Most issues found were related to inconsistent casing, accent marks, or confusion with values from other columns. In contrast, textual fields required a more thorough evaluation. From this process, several occurrences were noted, namely, the *title* presented inconsistent casing; some presented title case, where the main words were capitalised, and others had sentence case, where only the first word was capitalised. There were also inconsistencies in accent marks, in the use of units and their abbreviations, and confusion with other columns, similar to what occurred in single-select fields. Other issues included the use of synonyms, which, even if the platform supported semantic search, would reduce the portal's predictability, making it harder to browse, and inconsistencies in sentence structures, such as placing the unit of measurement at the beginning in some titles and at the end in others. There were also more specific cases, like the use of the masculine plural form, such as "alunos" when referring only to the male students, which can lead to ambiguity.

In most of these cases, there was no straightforward solution, and, therefore, it was necessary to discuss with members from each of CMM's departments responsible for the retrieval of data and metadata to discuss and establish consistent ways of handling each case. In this sense, there was a meeting to introduce and discuss these situations, but, more importantly, to raise awareness on the importance of collaboration in the definition of standards. The meeting resulted in the creation of a document to collect everyone's opinion and the schedule of future meetings.

Another script was then developed to assist in the detection of inconsistencies in textual columns. Its goal was to help detect three of the previously mentioned issues, specifically, irregular casing, inconsistent use of accent marks, and the use of synonyms. The script processed data from an XLSX file and printed text snippets to help locate the identified inconsistencies.

To identify irregular casing, spaCy [3], an NLP library, was used for sentence segmentation and tokenisation, using the pre-trained models *pt_core_news_sm* and *en_core_web_sm* for Portuguese and English columns, respectively. Punctuation and numbers were filtered out, as well as the first word of each sentence, since it was typically capitalised. A dictionary was then built where each key was a fully lowercased word, mapping to all the casing variations found in the text. If more than one format was present for a given word, it indicated a potential inconsistency and was therefore printed.

To detect inconsistent use of accent marks, a similar approach was followed. However, in this case, the first word of each sentence was retained, and all words were lowercased to ensure that the casing didn't influence the detection process.

Detecting synonyms was more complex. The same pre-trained SpaCy models used previously were applied to perform linguistic annotations on each text, and the expressions marked as noun chunks were selected. Stop words were then removed from these chunks in order to retain only the important elements. In addition to these noun chunks, individual terms marked as common or proper nouns were also added. The selected terms were then lemmatised and converted to lowercase. Finally, using the Sentence Tranformer and with the models *paraphrase-multilingual-MiniLM-L12-v2* and *all-MiniLM-L6-v2* for Portuguese and English columns, respectively, semantic embeddings were created for each term. Cosine similarity was then calculated between the embeddings and used to cluster the terms based on a threshold. Terms belonging to the same cluster were then classified as synonyms, and their original format was printed.

It is important to note that the script should only be considered as an assistance in the detection of inconsistencies, not replacing the manual review of each field. This limitation occurs not only because the script focuses solely on the three issues mentioned, not handling, for example, more complex structural inconsistencies, but also because it may fail to capture correctly all occurrences. This is especially true for the synonym detection, where even after tuning the similarity threshold, it still produces some false positives and fails to identify many synonyms.

3.1.2 Process Automation

Before the development of the automated pipelines, it was necessary to verify which APIs were provided by OpenDataSoft and understand if they supported both data and metadata management.

This evaluation revealed that ODS provided two REST APIs with distinct purposes. The first, the *Explore API*, focused primarily on platform users who access and consult datasets. It supported only GET requests, allowing retrieval of datasets, data, and associated metadata, but did not permit modifications. The second, the *Automation API*, was intended for platform administrators as it allowed them to manage all aspects of the platform and was therefore used in the development of the automated pipelines. Their documentation can be found in [24] and [23], respectively.

Since the CMM intended to maintain the same selection of datasets, and the process of creating and deleting them using the *Automation API* proved to be straightforward, the focus was placed on the management of metadata and data.

3.1.2.1 Metadata Management Automation

In terms of metadata management, the Automation API did not differentiate between uploading new metadata and updating existing one. It provided three types of requests that differed according to the scope of the update, whether it involved all schemas of a dataset, a specific schema, or a particular field within a schema. Initially, the schema-specific request was used because it made debugging easier, but it was later replaced by the catalogue request to reduce the number of API calls.

The flow of the script was structured as follows:

1. Retrieve from the metadata database all entries where the *ODS2up* field is set to "True", indicating that they had been recently modified and not yet updated on the platform.

2. Convert the metadata from the database format into the structure defined by the catalogue, applying the transformations described in 3.1.1.2. It was later agreed that this conversion process should not be limited to the OpenDataSoft pipelines but extended to all of the CMM's APIs.

3. Send the resulting metadata, now structured according to the catalogue, as the payload in the body of a PUT request to the endpoint `/datasets/{dataset_uid}/metadata/`.

4. If the field *nivelacesso* was set to "Público", send a request to endpoint `/datasets/{dataset_uid}/publish/` to publish the dataset. Otherwise, send a request to `/datasets/{dataset_uid}/unpublish/`, making the dataset available only in the back office.

5. For public datasets, create a scheduler using the value in the *periodactual* field via the endpoint `/datasets/{dataset_uid}/schedules/`. This scheduler ensured the dataset would be periodically republished.

6. Finally, upon successful completion of the process, set *ODS2up* in the database to "False", indicating that the dataset and the platform were once again synchronised. Update *status_ODS* to "publicado" if the dataset was published, or to "não publicado" otherwise.

It is important to clarify that the dataset_uid does not correspond to the technical_id. While the technical_id was the public identifier used in dataset URLs and in the Explore API, OpenDataSoft used a different ID in the Automation API for security reasons. Because of this distinction, it was necessary to first retrieve the list of datasets using the endpoint `/datasets/` and extract the corresponding UIDs to populate the column in the metadata table before running the update script. As for the technical_id, although it was often referred to as "metadata" in the documentation, this field could not be changed through the previously mentioned metadata endpoints. Instead, it could only be updated via the endpoint `/datasets/dataset_uid/`, by including the new value in the body of the request. This separation was justified, as changes to the technical_id reflected in the dataset's URL, potentially breaking any saved links. For this reason, the script responsible for updating the technical_id based on the database value was kept separate from the metadata update one.

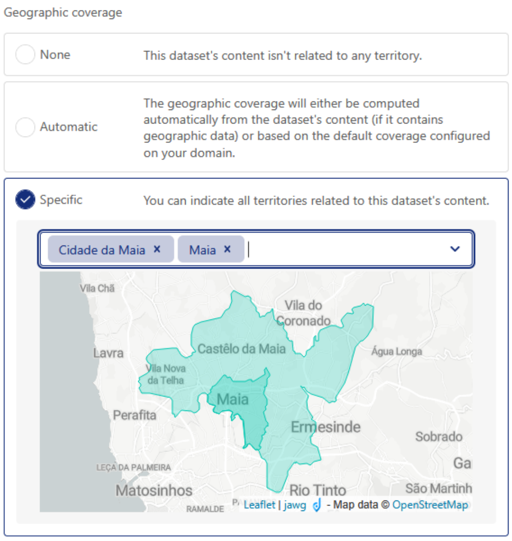
Regarding the *geographic_coverage* field in OpenDataSoft, it allowed three options, as shown in Figure 3.1:

- None, when the dataset is not related to any territory;
- Automatic when the data is georeferenced and, consequently, can be calculated automatically;
- Specific when the territory is specified.

In this software, the geographic coverage was specified using two fields. The first, *geographic_reference_auto*, a boolean that sets the georeference to automatic when true. The second, *geographic_reference*, a code representing a territory. When the first field was false and the second was left empty, the geographic reference was set to none.

The code had to follow a format like "pt_70_1306", where "pt" stood for the country, in this case Portugal, "70" indicated the scope, here meaning municipality, and "1306" corresponded to the specific territory, which in this example was Maia. The specific codes for Portuguese districts, municipalities, and parishes could be found in the OpenDataSoft datasets *georef-portugal-distrito*, *georef-portugal-concelho*, and *georef-portugal-freguesia*, respectively. Although no official documentation was found for the scope IDs, these were determined by setting each scope manually through the platform's UI and retrieving their corresponding IDs via GET requests.

As these fields were specific to OpenDataSoft, and the conversion script needed to be flexible enough to handle other APIs used by the CMM, an extra step was added to translate the territory names into their respective codes before updating the metadata.



Geographic coverage

○ None — This dataset's content isn't related to any territory.

○ Automatic — The geographic coverage will either be computed automatically from the dataset's content (if it contains geographic data) or based on the default coverage configured on your domain.

☑ Specific — You can indicate all territories related to this dataset's content.

Cidade da Maia × Maia ×

Figure 3.1: Geographic coverage selection

The intention to make the platform multilingual, supporting both English and Portuguese, was not forgotten, but it encountered some issues. Although each metadata field included an extra field for translations, after contacting the ODS support, it was revealed that these fields were only accessible directly in the platform's interface. ODS planned to add translation support to the *Automation API* soon, but it was not available at the time. As a result, the approach of maintaining private metadata in Portuguese and public metadata in English was kept to facilitate future bilingual support in both the front and back office.

3.1.2.2 Data Management Automation

For the data management, the CMM requested the development of test scripts for each of the necessary operations instead of a fully fledged create/update script like metadata. It was, therefore, important to understand which operations *Automation API* supported for managing dataset resources.

The API supported all CRUD (Create, Read, Update, and Delete) operations for dataset resources.

The read and delete operations were relatively straightforward. The read operation only required the dataset ID to return the list of associated resources. For deletion, both the dataset ID and the resource ID to be removed were necessary.

The create and update operations were more complex but very similar, with the key difference being that the update operation required the ID of the resource to be overwritten. Both operations used endpoints that supported three types of input: direct file upload, HTTP connections (with or without Basic Authentication), and reusable connections to storage platforms such as Google Drive and SharePoint.

For direct file upload, it was first necessary to make a POST request to the endpoint `datasets/dataset_uid/resources/files/`, sending the file in the request body. The returned file ID would then be used in the datasource field of the create request to make the file available in the dataset.

For HTTP connections, it was sufficient to specify the URL in the datasource field and include any necessary API keys or authentication tokens in its headers subfield.

Finally, for storage platforms, a reusable connection to the intended service needed to be previously defined. This could be done directly through the data portal's back office or via the *Automation API*, which also supported management operations for these connections.

Additionally, OpenDataSoft supported the use of harvesters as a way of importing datasets. Unlike the other functionalities described so far, which operated on individual datasets, harvesters were designed to retrieve and create several at once from other platforms such as CKAN and Socrata, mentioned in 2.2.2, or even from other ODS portals. Because they handled datasets in groups, harvesters offered adapted versions of certain operations, such as publishing, unpublishing, and associating schedulers, that were executed at the group level rather than per dataset.

3.2 Data Spaces

As explained in Section 1.4, the chosen approach was to develop a proof of concept using Sovity, in order to evaluate the complexity of developing a data space with this technology.

Although Sovity was a paid software, it offered a free, open-source community edition. This was considered an ideal solution, as it allowed for the development of a simplified platform for the proof of concept.

This version of the software included a Docker setup to run a local demo environment, which was adopted for the proof of concept. The focus then shifted to understanding this demo and defining a set of use cases to serve as a basis for testing it.

3.2.1 Sovity - Community Edition Demo Analysis

The demo featured both a data provider and a data consumer. Each component ran multiple services connected through a Caddy reverse proxy, namely, a UI, a connector service, a management API, and a PostgreSQL database, used to store all the history of transactions and operations performed by the connectors.

The Docker setup allowed for customisation through environment variables defined in the docker-compose.yml file. These variables enabled configuring key aspects of each service, such as database connection details, network settings and API keys.

It was also possible to change the deployment type. There were two options: Control Plane with Integrated Data Plane, where both control and data plane functionalities ran together in the same connector instance, and Control Plane plus Data Plane (Standalone), where the data plane was deployed separately from the control plane, indicated for more complex platforms.

The management API was working and allowed for the automation of some processes, but the lack of documentation made it difficult to use.

3.2.2 Use cases

In terms of use cases, three were chosen to access the demo. For each use case, its purpose, flow and acceptance criteria are presented:

Use Case 1: On Request Data Offer

Purpose: This use case evaluates the ability to create data offers without specifying a data source. Instead, the offer is made visible to consumers, who can directly contact the provider to discuss and negotiate the terms outside of the platform.

Flow:

1. A data provider creates an on-request data offer.
2. A data consumer consults the list of available data offers and selects that offer.
3. The consumer subscribes to the data offer and contacts the provider directly to discuss terms.

Acceptance Criteria:

- The data provider can successfully create an on-request data offer in the platform without specifying a data source.
- The consumer can access and subscribe to the data offer.
- The platform enables or provides a mechanism for the consumer to contact the provider outside the platform.

Use Case 2: Data Offer with Source

Purpose: This use case evaluates the ability to create a data offer with a source and the associated metadata accessible to consumers.

Flow:

1. A data provider configures a data offer with a data source and associated metadata.
2. A data consumer browses and selects the shared data offer to access its page.
3. The data consumer then subscribes to that offer.
4. The consumer obtains access to the data source and its metadata.

Acceptance Criteria:

- The data provider can successfully create a data offer specifying the source and the metadata.
- The data consumer can discover and select the shared data offer.
- The user gets access to the data's source and metadata.

Use Case 3: Restrict Data Offer

Purpose: This use case evaluates the ability to create data policies that restrict access to certain users.

Flow:

1. A data provider publishes a data offer and sets an access policy so that only a specific group of users can access it, excluding the consumer.
2. The consumer tries to consult the available data offers.
3. The data offer is not presented, or he can't subscribe.

Acceptance Criteria:

- The data provider can create and publish a data offer with a specific access policy.
- The unauthorised consumers can't view the data offer or can't subscribe to the offer.

3.3 Summary

In this chapter, the work carried out for the two data sharing solutions, namely the improvement of the metadata catalogue to meet European standards, the development of management pipelines for the data portal and the data space proof of concept. It is now important to present the final results.

Chapter 4

Results

This chapter presents the table created to support metadata improvement, followed by the open data portal, focusing especially on functionalities powered by these metadata improvements.

It is important to mention that the work carried out regarding metadata focused on defining guidelines and mechanisms for its control, rather than directly editing the metadata itself. The adaptation and repopulation process remained an ongoing effort of CMM, so some inconsistencies may still appear in the current state of the portal, along with the use of mocked data.

This chapter describes the current state of the data space proof of concept, focusing on the use cases defined in 3.2.2, more specifically in their flow.

4.1 Open Data Portal

Table 4.1 presents the metadata table, which maps all columns from the CMM's metadata database to the schema fields that will use them. For each column, the table indicates its meaning, intended use, and, when applicable, the predefined set of possible values it should take.

Source Table	D4Maia	Standard ODS	DCAT-AP	Description
sigla	sigla	technical_identifier		Internal code of the dataset. Has to be unique. It will determine the URL of the dataset and so shouldn't be changed.
ID_ODS				ID generated by OpenDataSoft. **Don't change**, used to identify the dataset in the platform(foreign key).
ODS2up				Specify if the data in the source has already been updated in Opendatasoft. True – needs updating, False – does not need updating. **Don't change this field**.
nome	nome			Dataset title in Portuguese. Should be consistent to facilitate search tasks as well as understanding. Under analysis, possibly the unit of measurement should be kept in the end, and this process could be automated with the column medidaanalise. For now, it is private.
nome_en		title	title (inherited)	Dataset title in English. Equal to the Portuguese version. It will be shown to citizens.
descriplus	descri			Dataset description in Portuguese. Concatenated with formacalculo and comm. For now, it is private.
descriplus_en		descri	descri (inherited)	Dataset description in English. Equal to the Portuguese version. It will be shown to citizens.
metainfurl	metainfurl			URL for the dataset metadata at the source. It will only be visible internally for management purposes.
dataultimaactual	dataultimaactual			Date of the last update from the original source. It will only be visible internally for management purposes.
dataultimaactuallocal	dataultimaactuallocal	modified		Date of the update in the datalake. It will be shown to citizens.

Table 4.1: Metadata table (Part 1 of 5)

Source Table	D4Maia	Standard ODS	DCAT-AP	Description
dataultimaverifica	dataultimaverifica			Date of the last verification for updates at the source. It will only be visible internally for management purposes.
comm	comm			Comment in Portuguese. It will be added to the dataset's description. In cases where the dataset is composite (combines more than one dataset), reference them using the dataset sigla enclosed in $ symbols. For example: dataset originating from $nhabit-old$ and $ContraPess$. This will be converted into the full names, and the reference will be added to the 'attributions' field. It is important to remove HTML tags like <a> and internal comments, as these now belong in the 'ObserIntern' column. For now, it is private.
comm_en		descri	descri	Comment in English. Equal to the Portuguese version. It is concatenated with a public field.
curador	curador			Person responsible for the dataset. It is important to keep it consistent with the currently used structure: ACRONYM (email). For example: GBG (gabrielb@gmail.com). This field will only be visible internally; citizens will only see the general email of the City Council for this area and a generic name.
status	status			Availability of the dataset. Should have the values "disponível" e "indisponível". It will only be visible internally for management purposes.
fonte		publisher	publisher (inherited)	Source that published the data. It will only be visible internally, except in cases where the 'origin' field is empty, in those cases, it will replace it.

Table 4.2: Metadata table (Part 2 of 5)

Source Table	D4Maia	Standard ODS	DCAT-AP	Description
origem		publisher	publisher (inherited)	Origin and responsible for the data published by the source. It will be visible to citizens in most cases, being replaced by the source when it is null. It is important to fix the cases where this field contains the license URL.
licenca	licenca	license		In most cases, it will be Creative Commons CCZero / CC-BY-SA 4.0. It is important to be consistent and write the value out in full and with proper spacing to ensure it matches the values used in Opendatasoft. It will be shown to citizens.
tema	tema			Dataset area/topics in Portuguese. It will be used by citizens to filter datasets and simplify the search process. Therefore, it is important to be consistent. For now, it is private
tema_en		theme		Dataset area/topics in English. Equal to the Portuguese version. It will be shown to citizens
formacalculo	formacalculo			Formulas used to calculate the values. It will be added to the dataset description and, therefore, will be visible to citizens. It is important to remove HTML tags such as <p>, bullet points, and parentheses that do not contribute to its understanding, as well as internal comments, which now belong in the 'ObserIntern' column. Maintain consistency. For now, it is private.
formacalculo_en		descri	descri	Formulas used to calculate the values in English. Equal to the Portuguese version. It is concatenated with a public field.
medidaanalise	medidaanalise			Unit of measurement. Might be concatenated into the title field. For now, it is private.

Table 4.3: Metadata table (Part 3 of 5)

Source Table	D4Maia	Standard ODS	DCAT-AP	Description
divterrit	divterrit	geographic_coverage		This will be used to assign each dataset to a territory in order to allow users to filter datasets. It is not relevant for datasets with georeferenced data. It should be consistent, using only the values "Metropolitana", "Região Norte", "Concelho", "Freguesia", "Portugal"
periodactual	periodactual	update_frequency		How often the dataset's data or metadata is updated. Useful for API listeners to know when to expect new information. Should have standardised values so they can be mapped to the ODS values. It will be shown to citizens.
nivelacesso	nivelacesso			Access level of the dataset. The values should be either "Público" or "Privado".
formatosdisp				Format in which the data is available, such as JSON, CSV, etc. It will only be visible internally for management purposes.
formaactual	formaactual			Method of update. Should have the value "Automático" or "Manual". It will only be visible internally for management purposes.
obserintern	obserintern			Internal comments regarding the dataset. The secondary information presented in an informal way at Comm should be moved here. It will only be visible internally for management purposes.
tabela_sql	tabela_sql			Id to connect with the data. It will only be visible internally for management purposes.
accrualperiodicity	accrualperiodicity		accrual_periodicity	How often the dataset is published. In cases where data is added, the period should be derived from the time the dataset is republished with these new values. Admits the same range of values as periodactual. It will be shown to citizens.

Table 4.4: Metadata table (Part 4 of 5)

Source Table	D4Maia	Standard ODS	DCAT-AP	Description
keywords	keywords	keywords	keywords (inherited)	List of words in English that identify the dataset's area/topic. It will be visible to citizens and is important for them to quickly determine if the dataset is relevant to their purpose and to identify similar datasets. It will be shown to citizens.
publishertype	publishertype		publisher_type	Type of data source. It should have one of the following values: "Academia-ScientificOrganisation", "Company", "IndustryConsortium", "LocalAuthority", "NationalAuthority", "NonGovernmentalOrganisation", "NonProfitOrganisation", "PrivateIndividual", "RegionalAuthority", "StandardisationBody" or "SupraNationalAuthority"
references		references		URL of the dataset in the datalake. It will be shown to citizens.
status_ODS	status_ODS			Used to manage the status of datasets in Opendatasoft. It can have the values 'publicado' or 'não publicado.'. **Don't change this field**.
TC_ED				End date referring to the period the data covers. It will be shown to citizens.
TC_SD				Start date referring to the period the data covers. It will be shown to citizens.

Table 4.5: Metadata table (Part 5 of 5)

With the metadata table presented, it is now time to explore the current state of the open data platform:

As the user visits the portal, the front page, depicted in Figure 4.1, offers a standard text search as well as a list of themes to help users navigate the portal more easily.

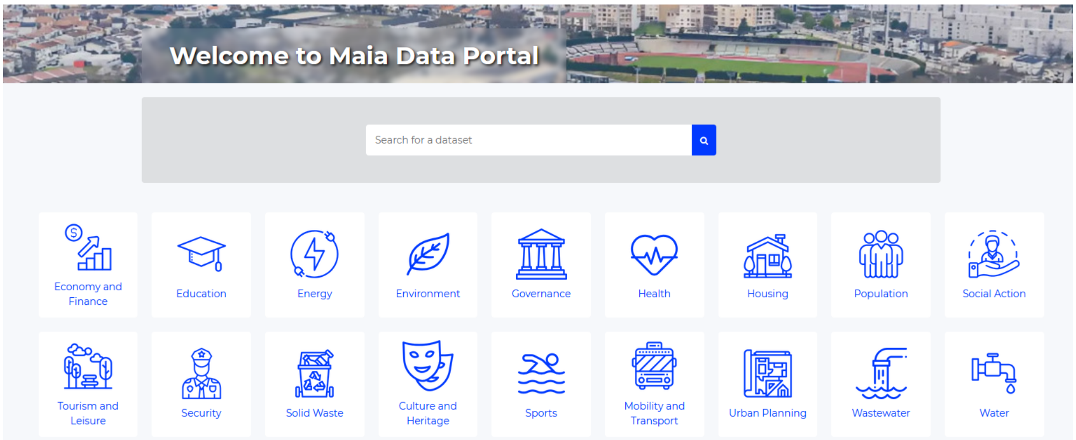
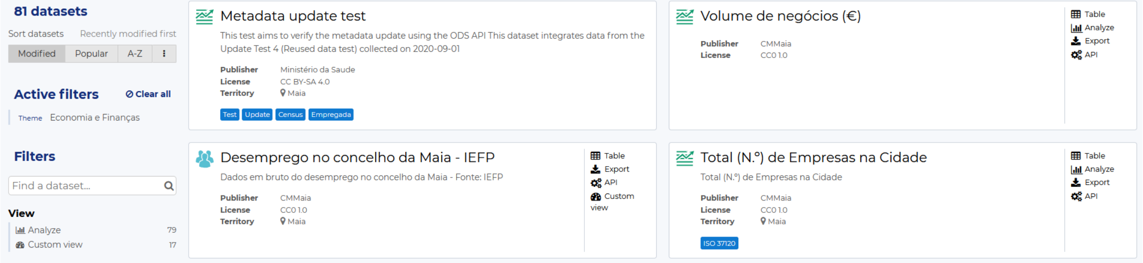


Figure 4.1: Open data portal frontpage

The user may then press on one of the themes, and it will be redirected to a page containing the datasets belonging to that theme. For example, when the "Economy and Finance" theme is selected, the user visits the page represented in Figure 4.2



81 datasets

Sort datasets Recently modified first

Active filters

Theme Economia e Finanças

Filters

View

Analyze 79
Custom view 17

Metadata update test

This test aims to verify the metadata update using the ODS API This dataset integrates data from the Update Test 4 (Reused data test) collected on 2020-09-01

Publisher Ministério da Saude
License CC BY-SA 4.0
Territory Maia

Test Update Census Empregada

Volume de negócios (€)

Publisher CMMaia
License CC0 1.0

Desemprego no concelho da Maia - IEFP

Dados em bruto do desemprego no concelho da Maia - Fonte: IEFP

Publisher CMMaia
License CC0 1.0
Territory Maia

Total (N.º) de Empresas na Cidade

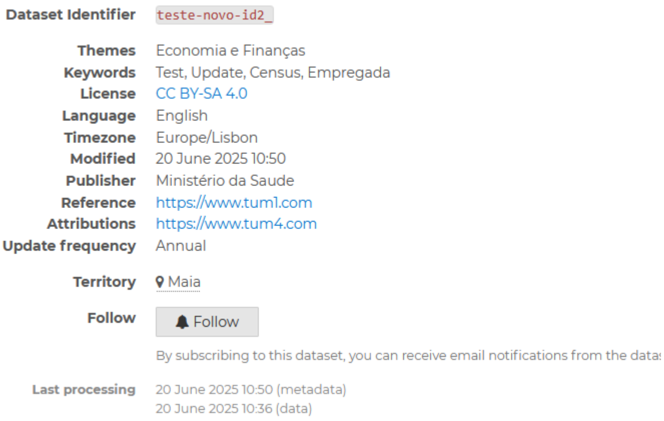
Total (N.º) de Empresas na Cidade

Publisher CMMaia
License CC0 1.0
Territory Maia

ISO 37120

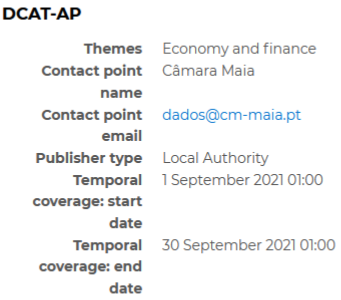
Figure 4.2: Economy and Finance page

When visiting the dataset, the user can consult the public schemas of metadata, Figures 4.3 and 4.4.



Dataset Identifier	teste-novo-id2_
Themes	Economia e Finanças
Keywords	Test, Update, Census, Empregada
License	CC BY-SA 4.0
Language	English
Timezone	Europe/Lisbon
Modified	20 June 2025 10:50
Publisher	Ministério da Saude
Reference	https://www.tum1.com
Attributions	https://www.tum4.com
Update frequency	Annual
Territory	Maia
Follow	Follow By subscribing to this dataset, you can receive email notifications from the datas
Last processing	20 June 2025 10:50 (metadata) 20 June 2025 10:36 (data)

Figure 4.3: ODS Standard metadata schema

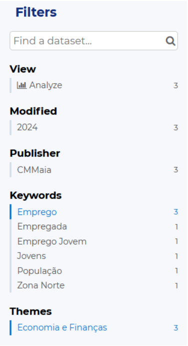


DCAT-AP

Themes	Economy and finance
Contact point name	Câmara Maia
Contact point email	dados@cm-maia.pt
Publisher type	Local Authority
Temporal coverage: start date	1 September 2021 01:00
Temporal coverage: end date	30 September 2021 01:00

Figure 4.4: DCAT-AP

Additionally, features like the dataset search filters, Figure 4.5, and the recommendation system for finding similar datasets, Figure 4.6, contribute to the improvement of the user experience.

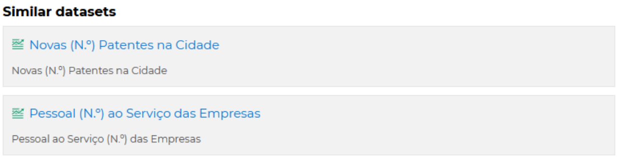


Filters

Find a dataset...

View	
Analyze	3
Modified	
2024	3
Publisher	
CMMaia	3
Keywords	
Emprego	3
Empregada	1
Emprego Jovem	1
Jovens	1
População	1
Zona Norte	1
Themes	
Economia e Finanças	3

Figure 4.5: Filters for dataset search



Similar datasets

Novas (N.º) Patentes na Cidade

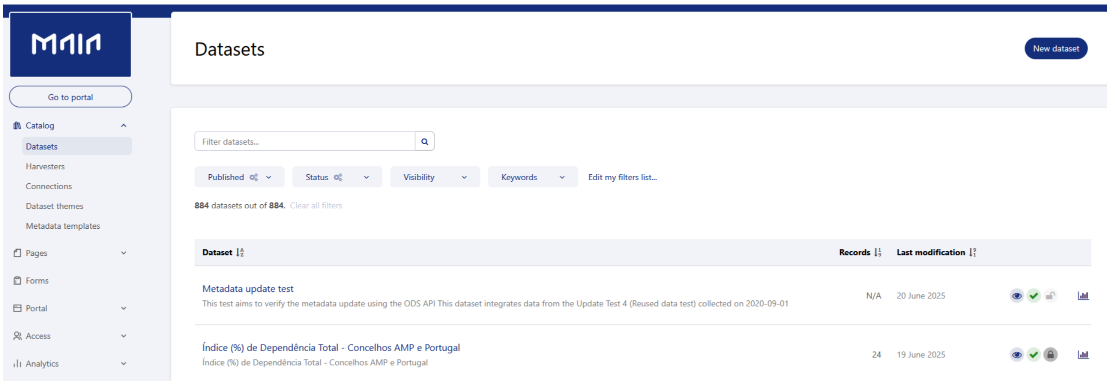
Novas (N.º) Patentes na Cidade

Pessoal (N.º) ao Serviço das Empresas

Pessoal ao Serviço (N.º) das Empresas

Figure 4.6: Recommendation system

There is also a page reserved for CMM members, the back office, shown in Figure 4.7. This page includes search and filter functionalities similar to those in the front office.



Datasets

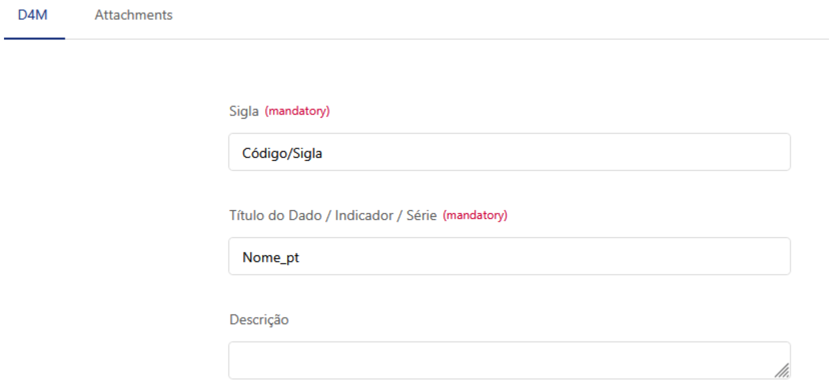
Filter datasets...

884 datasets out of 884. Clear all filters

Dataset	Records	Last modification
Metadata update test This test aims to verify the metadata update using the ODS API This dataset integrates data from the Update Test 4 (Reused data test) collected on 2020-09-01	N/A	20 June 2025
Índice (%) de Dependência Total - Concelhos AMP e Portugal Índice (%) de Dependência Total - Concelhos AMP e Portugal	24	19 June 2025

Figure 4.7: Back office page

In terms of metadata, the back office presents all available schemas, including the private *D4M* (Figure 4.8).



D4M Attachments

Sigla (mandatory)

Código/Sigla

Título do Dado / Indicador / Série (mandatory)

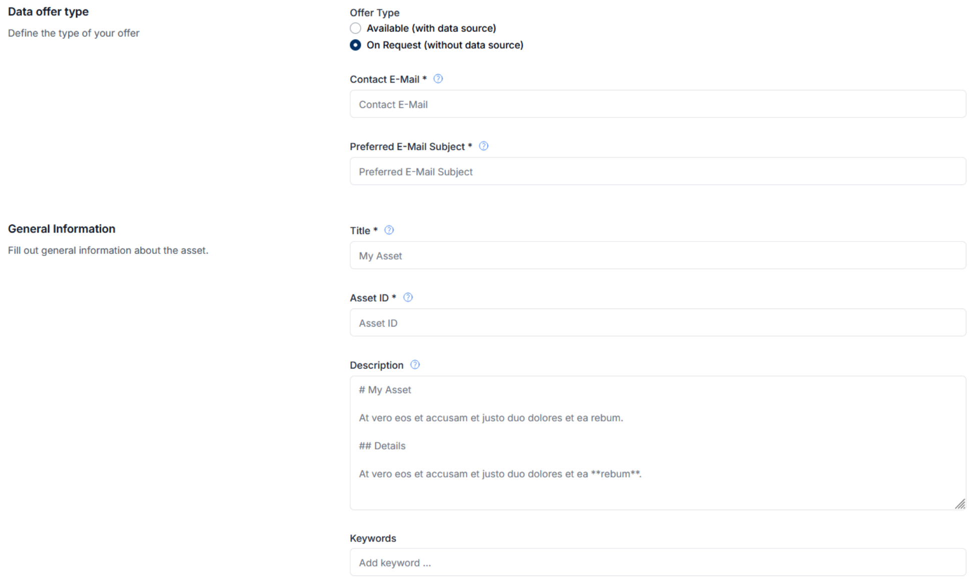
Nome_pt

Descrição

Figure 4.8: D4M schema

4.2 Data Space

For Use Case 1, "On Request Data Offer", the provider accesses the new data offer page, Figure 4.9, selects the offer type "On Request" and provides an email address, as well as the subject for the email he will receive when other user wants to reach him out. Additionally, some metadata fields have to be filled out along with the publishing mode, Figure 4.19. In this case, the unrestricted publish should be selected.



Data offer type

Define the type of your offer

Offer Type

○ Available (with data source)

● On Request (without data source)

Contact E-Mail *

Contact E-Mail

Preferred E-Mail Subject *

Preferred E-Mail Subject

General Information

Fill out general information about the asset.

Title *

My Asset

Asset ID *

Asset ID

Description

My Asset

At vero eos et accusam et justo duo dolores et ea rebum.

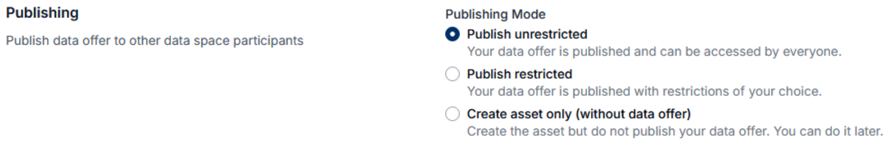
Details

At vero eos et accusam et justo duo dolores et ea **rebum**.

Keywords

Add keyword ...

Figure 4.9: New data offer page - On Request



Publishing

Publish data offer to other data space participants

Publishing Mode

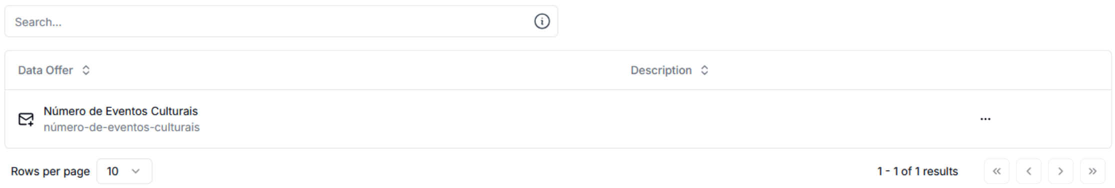
● Publish unrestricted
Your data offer is published and can be accessed by everyone.

○ Publish restricted
Your data offer is published with restrictions of your choice.

○ Create asset only (without data offer)
Create the asset but do not publish your data offer. You can do it later.

Figure 4.10: Publishing mode - Unrestricted

After this, the consumer can navigate the catalogue browser, Figure 4.11, and select the desired data offer. This will take the user to the offer page shown in Figure 4.12.



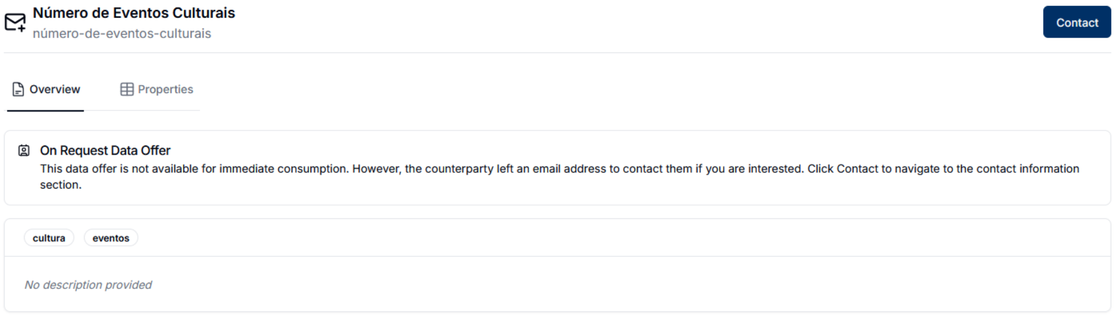
Search...

Data Offer	Description	
Número de Eventos Culturais número-de-eventos-culturais		...

Rows per page 10

1 - 1 of 1 results

Figure 4.11: Catalogue browser



Número de Eventos Culturais
número-de-eventos-culturais

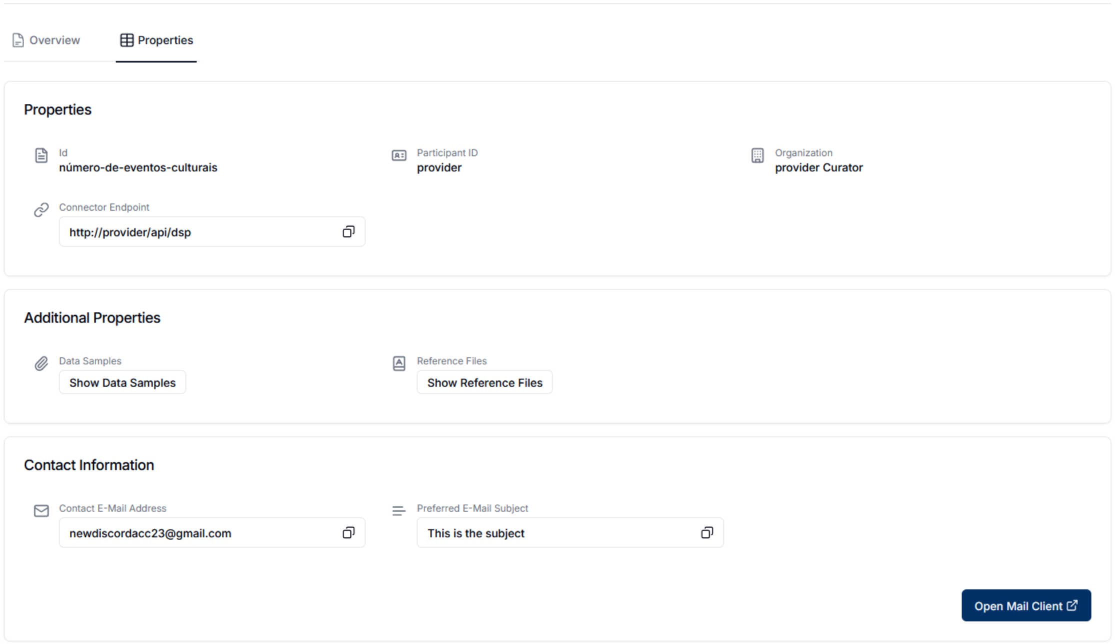
On Request Data Offer
This data offer is not available for immediate consumption. However, the counterparty left an email address to contact them if you are interested. Click Contact to navigate to the contact information section.

cultura eventos

No description provided

Figure 4.12: Data offer page - On Request

The user can then navigate to the properties tab, Figure 4.13 via the "Properties" or "Contact" button. On this page, the user can press the button "Open Mail Client" to be redirected to their default email application with a new email draft addressed to the provider's contact.



Properties

Id: número-de-eventos-culturais

Participant ID: provider

Organization: provider Curator

Connector Endpoint: http://provider/api/dsp

Additional Properties

Data Samples

Reference Files

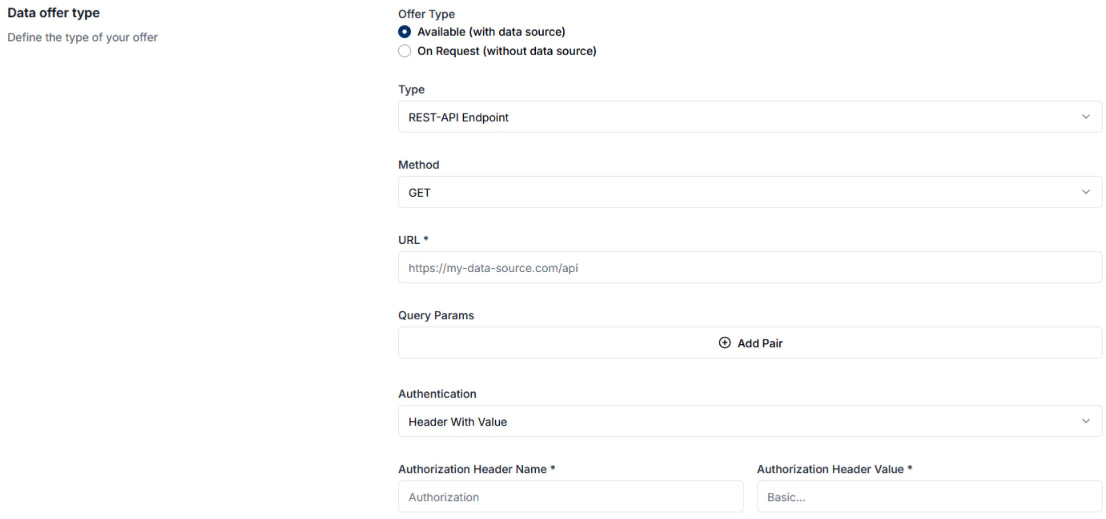
Contact Information

Contact E-Mail Address: newdiscordacc23@gmail.com

Preferred E-Mail Subject: This is the subject

Figure 4.13: Properties tab

Use Case 2, "Data Offer with Source", has a similar flow. The main differences reside in the selection of the offer type "Available", which replaces the email option with the configuration of a REST API endpoint, Figure 4.14, still being necessary to fill out some metadata fields and the publishing mode.



Data offer type

Define the type of your offer

Offer Type
● Available (with data source)
○ On Request (without data source)

Type: REST-API Endpoint

Method: GET

URL *: https://my-data-source.com/api

Query Params

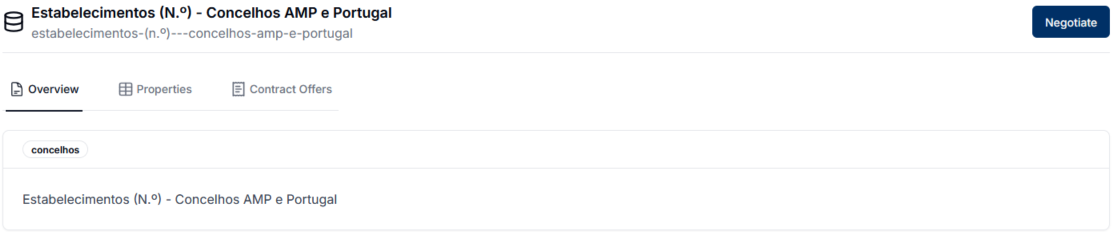
Authentication: Header With Value

Authorization Header Name *: Authorization

Authorization Header Value *: Basic...

Figure 4.14: New data offer page - Available

For the data consumer, the navigation through the catalogue browser and the selection of the desired data offer is the same as in Use Case 1. The difference begins on the properties page, Figure 4.15, which includes an additional tab named "Contract Offers". Instead of the "Contact" button, a "Negotiate" button is available to redirect the user to this tab.



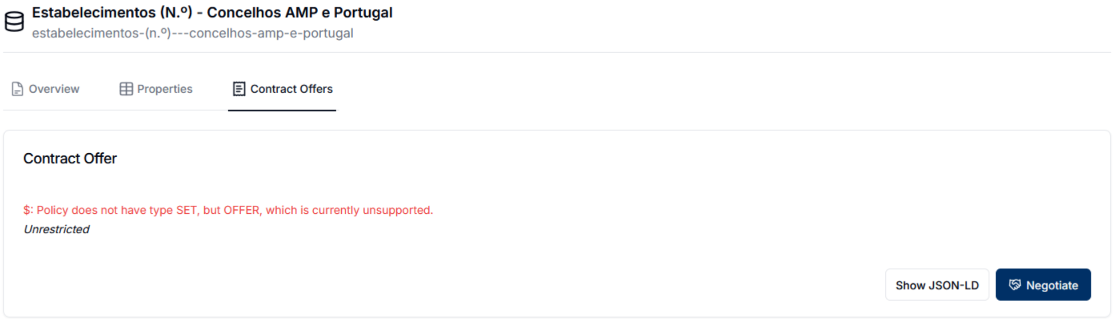
Estabelecimentos (N.º) - Concelhos AMP e Portugal

estabelecimentos-(n.º)---concelhos-amp-e-portugal

Estabelecimentos (N.º) - Concelhos AMP e Portugal

Figure 4.15: Data offer page - Available

As shown in Figure 4.16, in this tab, the user can consult the contract policy or press the "Negotiate" button again to subscribe to the data offer, after accepting the terms and conditions.



Estabelecimentos (N.º) - Concelhos AMP e Portugal

estabelecimentos-(n.º)---concelhos-amp-e-portugal

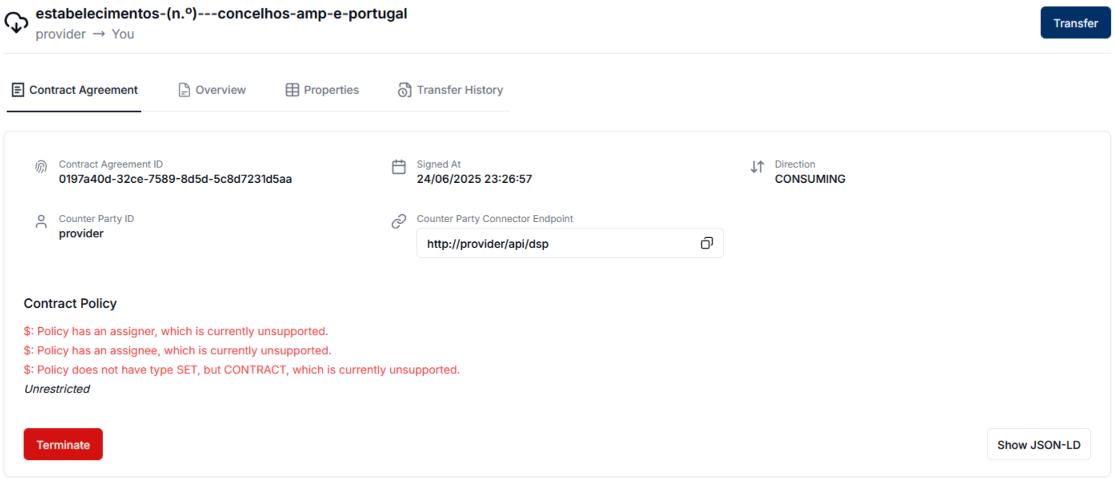
Contract Offer

$: Policy does not have type SET, but OFFER, which is currently unsupported.

Unrestricted

Figure 4.16: Contract Policy tab

Once subscribed, the data offer's page, Figure 4.17, contains two more tabs, the "Contract Agreement" and the "Transfer History". This page also includes a "Transfer" button that redirects to the transfer page, Figure 4.18, where the user can specify the endpoint to which the data should be transferred. Upon assigning a test endpoint, both data and metadata were proven to be received through a POST request.



estabelecimentos-(n.º)---concelhos-amp-e-portugal

provider → You

Contract Agreement | Overview | Properties | Transfer History

Contract Agreement ID: 0197a40d-32ce-7589-8d5d-5c8d7231d5aa

Signed At: 24/06/2025 23:26:57

Direction: CONSUMING

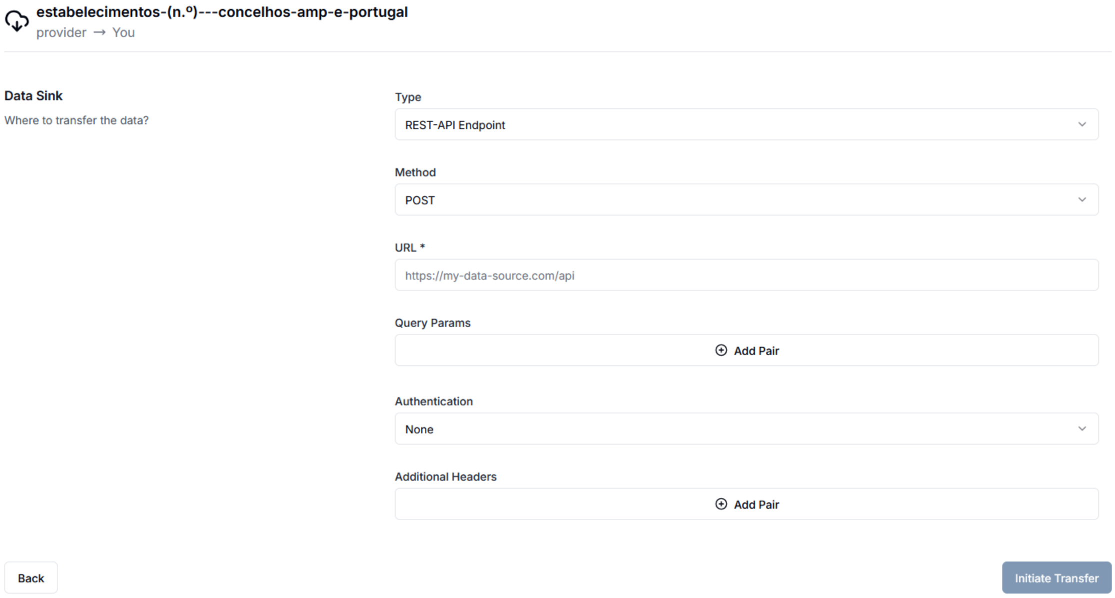
Counter Party ID: provider

Counter Party Connector Endpoint: http://provider/api/dsp

Contract Policy

$: Policy has an assigner, which is currently unsupported.
$: Policy has an assignee, which is currently unsupported.
$: Policy does not have type SET, but CONTRACT, which is currently unsupported.
Unrestricted

Figure 4.17: Data offer page - Subscribed



estabelecimentos-(n.º)---concelhos-amp-e-portugal

provider → You

Data Sink

Where to transfer the data?

Type: REST-API Endpoint

Method: POST

URL *: https://my-data-source.com/api

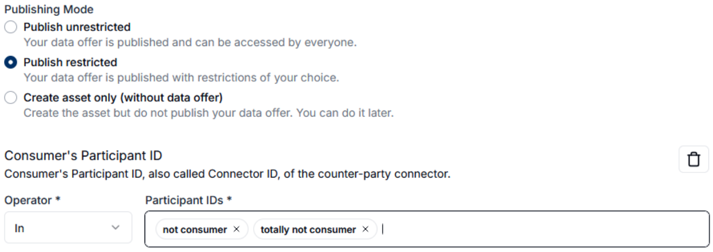
Query Params

Authentication: None

Additional Headers

Figure 4.18: Transfer page

Finally, for Use Case 3, "Restrict Data Offer", the only difference stems from the publishing mode, where the option restricted is selected. This opens an additional field where the operator can specify which users are allowed access by writing their IDs.



Publishing Mode

○ Publish unrestricted
Your data offer is published and can be accessed by everyone.

● Publish restricted
Your data offer is published with restrictions of your choice.

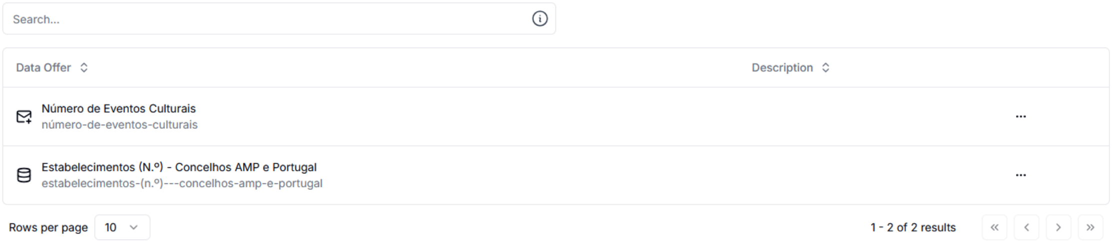
○ Create asset only (without data offer)
Create the asset but do not publish your data offer. You can do it later.

Consumer's Participant ID
Consumer's Participant ID, also called Connector ID, of the counter-party connector.

Operator *	Participant IDs *
In	not consumer × totally not consumer ×

Figure 4.19: Publishing mode - Unrestricted

As we can see in Figure 4.20, when the consumer browses the catalogue, only the previous offers are visible; the restricted dataset is not displayed.



Search...

Data Offer	Description
Número de Eventos Culturais número-de-eventos-culturais	
Estabelecimentos (N.º) - Concelhos AMP e Portugal estabelecimentos-(n.º)---concelhos-amp-e-portugal	

Rows per page 10

1 - 2 of 2 results

Figure 4.20: Catalogue browser with private offer

4.3 Summary

The final results of the CMM's data catalogue, the data portal and the data space proof of concept were presented in this chapter. In the next chapter, these results will be evaluated.

Chapter 5

Evaluation

This chapter presents the evaluation of both the open data portal and the data space proof of concept.

5.1 Open Data Portal

The evaluation of the open data portal was conducted using two different approaches. The first focuses on the quality of the data/metadata by evaluating a hypothetical dataset whose metadata follows the principles in Table 4.1, and whose data share the same characteristics as the majority of the datasets available, assumed to be published in the portal, based on the FAIR principles described in 2.1.3. The second approach evaluates the portal against the requirements presented in 2.2.1, considering the upgrades enabled by the metadata improvements and the new update pipelines.

5.1.1 Dataset FAIRness

The analysis of the dataset for each Fair principle is presented below:

Findability

- **F1** - The dataset includes a technical identifier corresponding to its unique ID within the portal and determining its URL. While this identifier is still subject to change during development, it has been excluded from the metadata update pipelines, as it is not intended to change after the portal's release.
- **F2** - The dataset is described with structured metadata, as shown in the previous chapter. A more detailed analysis of its richness and completeness is provided in **R1**.
- **F3** - The technical identifier described in F1 is kept in the dataset's metadata, ensuring a link between the data and metadata.

Accessibility

- **A1.1** - HTTP is an open source, free to use and globally implementable protocol, meeting the criteria of **A1.1**.
- **A1.2** - The *Explore API*, uses OAuth2 to enable authentication and authorization operations.
- **A2** - Currently, the platform does not define how metadata would be managed if the associated data became unavailable. A possible solution would be to preserve the metadata or, at minimum, maintain a reference to the metadata stored in the data lake to guarantee continued accessibility.

Interoperability

- **I1** - Both metadata and data are expressed using formal, structured, and machine-readable formats. The metadata catalogue includes DCAT-AP, based on RDF, which enables semantic representation.
- **I2** - The metadata uses controlled vocabularies based on DCAT-AP. Additionally, the metadata table allowed for the values of each field to be made more consistent.
- **I3** - Derived datasets include references to the sources in the attribution field or within the description, where these relationships are explained.

Reusability

- **R1.1** - Each dataset's metadata includes a license field. Additionally, the platform displays a URL where the license can be consulted.
- **R1.2** - Each dataset contains a publisher field indicating the entity responsible for the original publication and the data source. Additionally, the publisher type field specifies the nature of the publishing entity.
- **R1.3** - The metadata catalogue includes DCAT-AP, the indicated standard for European portals. In terms of data, it is necessary to analyse and adopt appropriate standard formats for each data type. For example, inconsistencies were identified in the representation of time series data, where some datasets used two separate JSON arrays, one for time and one for values, while others used an array of time-value pairs.

The dataset met almost all the principles, falling short in A2 and R1.3.

5.1.2 Portal Requirements

The analysis for the open data portal, considering each of the OGD portal requirements and their metrics, is presented below:

- **Organise for Use** - For this requirement, the portal receives **three points out of five (3/5)**.

  In fact, it allows data preview, includes a recommendation section within each dataset that links to similar datasets, and provides a keywords field in the metadata to support dataset linkage, metrics (b), (c) and (e), respectively.

  Regarding comprehensive descriptive records complementary to the metadata, metric (a), only a few datasets include a custom view page containing this type of information. It is, therefore, necessary to extend this functionality by including more information and applying it to other datasets.

  The review systems, from metric (d), could also be an interesting feature to implement in the future, as it would allow for the CMM to easily identify which datasets require improvements.

- **Promote Use** - In terms of use promotion, the platform requires more work, being awarded a **single point out of five (1/5)**.

  Currently, the only metric met is (c), as the platform offers a subscription system that allows users to receive notifications about dataset updates.

  In regard to social media connections, metric (a), the platform includes a section with links to other CMM websites and portals in various areas such as environment, education, and sports. Creating dedicated social media accounts to promote the platform and including their links in this section could be a smart step to increase visibility and engagement.

  In terms of user feedback, OpenDataSoft supports the creation of forms to collect user feedback. The use of this feature is already being explored by the CMM and is expected to be implemented soon.

  As for metric(d), ODS provides both user guidance and support resources, but these are presented externally to the platform, so the inclusion of their links is necessary. These would need to be selected taking into consideration certain datasets' characteristics, such as the data type.

  Finally, the creation of reuse pages, metric (e), would be an interesting way to promote the use of datasets, opting for fictitious scenarios at the release and eventually including real ones. These scenarios could also be included in the datasets for easier access.

- **Be Discoverable** - As for discoverability, the portal is awarded **two points out of four (2/4)**.

  By consisting of an open data portal featuring a search system, metrics (a) and (b) are met.

  The semantic search system, capable of handling synonyms, described in metric (c), is a functionality provided by OpenDataSoft, but it is not currently implemented in the portal.

  As for the list of unavailable datasets (d), these datasets are currently unpublished and stored only in the back office, making it easier to provide users with access to a list of them and help prevent wasted time searching for unavailable data.

- **Publish Metadata** - Regarding metadata, the portal is currently at **level five out of five (5/5)** of maturity, "Linked Open Metadata".

  In fact, the metadata catalogue contains the DCAT-AP standard, which is based on RDF and uses linked data principles by assigning unique identifiers and structuring metadata in a machine-readable format.

- **Promote Standards** - When it comes to standards, the portal is awarded **four points out of five (4/5)**.

  Each dataset includes a technical identifier that determines its URL. While this identifier is still being worked on, it is not intended to change after the portal's release, so metric (a) is fulfilled.

  In terms of metadata, the catalogue includes DCAT-AP, a standard format, and can be retrieved using the *Explore API*, a REST API that operates over HTTP, a standard protocol, therefore satisfying both metric (d) and (e).

  As for date types, (c), OpenDataSoft detects fields that contain dates. It then interprets that information using several standardised date formats and finally displays the dates in those columns according to the defined region's default date format, ensuring consistency and readability for users.

  The dataset version history, metric (b), is not currently supported by the portal. Before implementing this feature, it is important to consider the resources required, especially given the large number of existing datasets.

- **Co-Locate Documentation** - Concerning the documentation, and assuming that metadata is not included, the platform currently sits at the bottom, **Level one out of five (1/5)**.

  The only supporting documentation is included in the custom view page of each dataset. However, as mentioned in previous requirements, only a few datasets contain this feature, and so it is not accounted for.

- **Link Data** - The platform is currently at the **third level of Tim Berners-Lee's five-star scheme (3/5)** for Linked Open Data.

  Although Opendatasoft supports linked data formats such as JSON-LD, RDF/XML, and Turtle, the majority of datasets are provided in JSON, a machine-readable and non-proprietary format.

- **Be Measurable** - Regarding measurement practices, the portal is currently at the **third level out of four (3/4)** of maturity.

  It possesses tracking systems that are gathering information about dataset activity, such as the number of visits each dataset receives, the number of API calls made, and the number of downloads. It also collects information on the overall portal usage, such as the number of visits to each page and a record of textual searches performed by users.

To reach the next level of maturity, it is important to develop methods that measure the impact of the published data. But more importantly, and something that the metrics fail to evaluate, the priority should be to utilise the captured information to detect tendencies that might help make well-informed decisions regarding the future of the platform.

- **Co-Locate Tools** - Considering the co-locate tools, the platform currently occupies the **second level of four (2/4)**, as it only provides visualisation tools for the data, not supporting any form of user collaboration.

  Potentially, the review systems and forms already mentioned in previous requirements would help advance the platform to the third level.

  There was also a suggestion that turned up during discussions with the CMM. The inclusion of a functionality that allows users to create new datasets by combining and/or applying other transformations to existing ones. These user-generated datasets could then be submitted for review and, upon approval by an administrator, published on the portal.

- **Be Accessible** – When evaluating accessibility, the portal scores **three points out of five (3/5)**.

  As mentioned before, the majority of datasets are available in JSON, a format that meets the criteria of metric (a). Additionally, metric (b) is also met, as the platform allows users to export datasets in other compatible formats.

  In terms of support for the visually and hearing impaired, metric (e), the platform implements some features provided by OpenDataSoft. And, even though they are mainly focused on assisting visually impaired users, the platform does not contain multimedia elements that might present challenges for hearing-impaired users. These features include the use of WAI-ARIA attributes in decorative elements, which prevents screen readers from reading out non-informative content, the tagging of images with alt attributes and meaningful titles, the attribution of accurate labels to links, and the full keyboard navigation support for all user interface elements.

  As for multilingual support, metric (d), even though it is not yet met, once translation support is added to the *Management API* and the CMM finishes populating the translation fields in the metadata database, the portal will offer both Portuguese and English versions. At that point, this metric can be considered fulfilled.

  Finally, while OpenDataSoft provides documentation on the mechanisms for accessing and interacting with datasets, required by metric (c), this documentation is not currently included in the portal. Consequently, this metric is not considered satisfied.

The portal finished with a score of twenty-seven out of forty-seven points (27/47). When compared with the study results described in 2.2.1, it falls just one point short of the dados.gov portal. The gap increases when compared to the European Open Data Portal. However, it is important to keep in mind that the study was published in 2020, and so, that evaluation may be outdated.

5.2 Data Space

For the evaluation of the data space demo, the acceptance criteria defined in 3.2.2 was used to validate the results presented in 4.2.

All use cases respected the defined criteria. The access to both types of data offers worked as expected, with the "on request" type leading to a new email draft addressed to the owner of the offer, and the "available" type to a transfer page where the user could specify the endpoint to which the data should be transferred. As for the access policies, these could be assigned during the creation of a data offer to restrict which users were allowed to view it.

5.3 Summary

In this chapter, the results produced in the last chapter were evaluated. Even though this evaluation proved satisfactory, there were still some limitations that should be addressed.

The next chapter presents a final overview of this project, including a discussion of these limitations and suggestions for future work aimed at overcoming them.

Chapter 6

Conclusion

6.1 Final Overview

The project that results in this thesis aimed to assist the Maia City Council in the development of data-sharing solutions for their data lake, specifically, an open data portal for the Maia citizens, and a data space platform for the OMEGA-X project.

In terms of the open data portal, the goal was to improve Maia's metadata catalogue by evolving from a single Maia-specific schema to a structure aligned with European standards and that maximised the functionalities provided by OpenDataSoft, the software used in the creation of the portal. As well as developing well-documented pipelines for the automated management of the platform.

Regarding the metadata, the first step was the selection/creation of the necessary schemas. This process resulted in the adoption of three schemas: DCAT-AP, a standardised interoperability schema recommended for European platforms; the standard ODS schema, a mandatory schema defined by OpenDataSoft; and D4Maia, a private schema created to address the requirements of the CMM.

With the catalogue defined, it was necessary to evaluate the database and determine it it was prepared for the new requirements.

This evaluation was made using a correspondence table. The database columns were studied and associated with their corresponding metadata fields. The necessary adaptations were then carried out. Some changes involved modifying the database structure, like creating new columns, while others focused on the content of existing columns, including repurposing them. In some cases, combinations of columns were mapped to a specific metadata field. Finally, new fields were introduced in the D4Maia schema.

The final iteration of the table served as documentation for the CMM members responsible for adapting and repopulating the database, as it included for each column the description, purpose, special notations, and possible values.

After the structural alignment, it was time to assess the metadata itself by carefully analysing the database's content in order to detect inconsistencies.

In most cases, the detected inconsistencies didn't have a straightforward solution, and so, it was necessary to discuss and establish consistent ways of handling them with members of the CMM.

This led to a multi-department meeting, an important step in raising awareness on the importance of collaboration in the definition of standards. This meeting resulted in more initiatives of this nature.

Additionally, a script based on NLP was created to assist in the detection of these inconsistencies. Due to its limitations, it could be used to detect some, but it could not replace a complete analysis.

This marked the end of the maturation process that aimed to align the metadata catalogue based on internal organisational requirements, owned by the CMM, with European standards.

It was then time to develop automated pipelines responsible for the platform's management using the *Automation API* provided by OpenDataSoft.

Since the CMM intended to maintain the same selection of datasets, and the process of creating and deleting them using the API proved simple, the focus was placed on the automated management of metadata and data.

For metadata, a script was developed. This script was responsible for retrieving from the database entries with the flag recently modified set to "True". These results would then be converted into the catalogue structure, applying all the necessary transformations. Using ODS's Management API, the script would update the platform. Depending on the defined privacy level, it would publish the dataset or make it only available in the back office. Additionally, a scheduler would be created for public datasets, responsible for republishing them in the provided update period. In the end, the recently modified flag would again be set to "False".

As for the data, the CMM opted to develop test scripts that covered the necessary operations. In this regard, all resource-related operations were tested. Specifically for the create and update operations, examples were developed for all the input types allowed.

With that, the development of pipelines for managing an open data portal built in OpenDataSoft according to the CMM's needs was finished. Additionally, the remaining management pipelines could be created by incorporating the functionalities demonstrated in the test scripts.

After the development stage, the open data portal was evaluated following two approaches.

The first aimed at assessing the quality of the data and metadata. It consisted of evaluating a hypothetical dataset whose metadata followed the principles defined in the correspondence table, and whose data shared the same characteristics as the majority of the datasets available, assumed to be published in the portal, based on the FAIR principles.

The second was to evaluate the state of the portal itself in terms of requirements using metrics defined in a study published by the University of Southampton.

The results of both evaluations were positive, but revealed some issues that still needed to be handled.

Finally, for the data space platform, the city council was still developing preliminary work, and it considered that the first step should be creating a proof of concept using Sovity to understand its complexity.

Upon investigating this technology, it was discovered that Sovity provided a free, open-source community edition, which was considered ideal.

This version of the software included a Docker setup to run a local demo, which was adopted for the proof of concept.

Three use cases were then created to test whether this software supported the basic functionalities.

The first use case consisted of creating "on request" data offers, allowing users to contact the providers and negotiate the terms outside the platform.

The second use case referred to the ability to create a data offer with a source, allowing users to extract the data and associated metadata after subscribing to the deal.

The last use case evaluated the ability to create data policies and restrict user access.

All use cases passed their acceptance criteria.

6.2 Limitations and Future Perspectives

6.2.1 Open Data Portal

During the evaluations, several suggestions were made for future work to improve the current state of the portal. These are presented below:

- Establish a plan on how the portal should handle the removal of a dataset. Opting for only removing the resources would be a good approach, allied with tagging the dataset as unavailable. It would also improve the discoverability, as users would have access to the list of unavailable datasets, preventing them from wasting time searching.

- Adopt community standards for data, resolving inconsistencies like the ones identified in the representation of time series data.

- Make sure that once the platform is finally released, no changes occur to the technical identifier, as it would change the dataset's URL.

- Once OpenDataSoft extends support for translation in the *Management API* and the translations fields in the metadata database are correctly populated, change the automated pipelines for metadata update in order to account for the Portuguese and English languages.

- Include in the datasets the links to the user guidance and support provided by ODS, selecting the appropriate ones based on the dataset characteristics.

- Include in the portal links to the documentation on how to access and interact with datasets that ODS provide. This could possibly be done in a new Help page.

- Add reuse examples to the portal. These could possibly be included in a new page or directly in the dataset.
- Further utilise the custom view feature to provide each dataset with comprehensive descriptive records that complement the metadata and relevant documentation, and extend this functionality to the remaining datasets. To support this, it is important to first determine how the necessary information can be collected and to establish automated pipelines to add and update this content across the datasets.
- Include forms in the portal in order to collect user feedback.
- Add a review system that allows users to rate datasets.
- Study the possibility of adding a functionality that allows users to create new datasets by combining and/or applying transformations to existing ones. These user-generated datasets could then be submitted for review and, upon approval by an administrator, published on the portal.
- Implement the semantic search system, a feature provided by OpenDataSoft.
- Identify ways to make better use of the data already being captured by the dataset and user activity tracking systems, and, secondly, develop methods to measure the impact of the published data.
- Create and connect the portal with social media to further promote the platform.
- Study the inclusion of a dataset version history, considering the storage requirements.
- Encourage the adoption of linked data formats like JSON-LD and RDF/XML.

In addition to the future work it is also important to note that initiatives such as the multi-department meetings to discuss and agree on standards, or the sharing of forms to collect feedback from each member, mentioned in 3.1.1.2, continue to take place, as the improvement of metadata occurs not only at the technical domain but also at human and organizational levels.

6.2.2 Data Spaces

Since the demo successfully covered all defined use cases, future work should focus on defining the remaining requirements for the platform and understanding if it would be better to keep using Sovity's community version, building upon the demo, or if it justifies to acquire the paid version, which contains a bigger range of functionalities, more documentation and continuous user support.

References

[1] Eric Afful-Dadzie and Anthony Afful-Dadzie. “Local Government Open Data (LGOD) Initiatives: Analysis of Trends and Similarities Among Early Adopters”. In: July 2018, pp. 299–308. ISBN: 978-3-319-94540-8. DOI: 10.1007/978-3-319-94541-5_30.

[2] Agência para a Modernização Administrativa. *Guia de Dados Abertos*. Accessed: 2025-06-07. 2016. URL: https://www.ama.gov.pt/documents/24077/24804/guia_dados_abertos_ama.pdf/aa97d8e8-c5fe-47ab-9500-734948c02b19.

[3] Explosion AI. *spaCy: Industrial-strength Natural Language Processing in Python*. https://spacy.io/. Accessed: 2025-06-03. 2015.

[4] Mohsan Ali, Charalampos Alexopoulos, and Yannis Charalabidis. *A comprehensive review of open data platforms, prevalent technologies, and functionalities*. Conference Paper. 2022. DOI: 10.1145/3560107.3560142. URL: https://doi.org/10.1145/3560107.3560142.

[5] Judie Attard et al. “A systematic review of open government data initiatives”. In: *Government Information Quarterly* 32.4 (2015), pp. 399–418. ISSN: 0740-624X. DOI: https://doi.org/10.1016/j.giq.2015.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0740624X1500091X.

[6] David Browning and Riccardo Albertoni. *Data Catalog Vocabulary (DCAT) – Version 3*. Accessed: 2025-06-27. Aug. 2024. URL: %5Curl%7Bhttps://www.w3.org/TR/vocab-dcat-3/%7D.

[7] European Commission. *European Interoperability Framework - Implementation Strategy*. Accessed: 2024-12-10. 2017. URL: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf.

[8] Dublin Core Metadata Initiative (DCMI). *Dublin Core™ Metadata Initiative Glossary*. 2005. URL: https://www.dublincore.org/specifications/dublin-core/usageguide/2005-05-26/glossary/.

[9] European Commission. *Linking data: Data Catalogue Vocabulary Application Profile (DCAT-AP)*. Accessed: 2025-06-27. 2022. URL: https://data.europa.eu/en/publications/datastories/linking-data-data-catalogue-vocabulary-application-profile.

[10] European Data Portal. *The Benefits and Value of Open Data*. https://data.europa.eu/en/publications/datastories/benefits-and-value-open-data. Accessed: 2024-10-01. 2024.

[11] European Union. *European Union Open Data Portal*. [Accessed 2025-04-19]. 2025. URL: https://op.europa.eu/en/.

[12] Publications Office of the European Union. *data.europa.eu – The official portal for European data*. Accessed: 2025-06-10. 2025. URL: https://data.europa.eu/en/.

[13] Publications Office of the European Union. *Sustainability of (open) data portal infrastructures – Open data portal assessment using user-oriented metrics*. Publications Office, 2020. DOI: doi/10.2830/51660.

[14] GO FAIR Initiative. *FAIR Principles*. Accessed: 2024-12-10. 2016. URL: https://www.go-fair.org/fair-principles/.

[15] Capgemini Invent et al. *The future of open data portals*. Publications Office, 2020. DOI: doi/10.2830/879461.

[16] M. Jibladze et al. “E-governance under the framework of open governance in Georgia: current situation, problems and opportunities”. In: *Public Administration and Policy* 27.2 (2024). Export Date: 25 November 2024; Cited By: 0, pp. 193–205. DOI: 10.1108/PAP-05-2023-0074. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200474267&doi=10.1108%2fPAP-05-2023-0074&partnerID=40&md5=61fca5f6efcd3307ef2782b5b93503d2.

[17] S. B. Lim and K. A. Kamaruddin. “Violated factors in building citizen-centric e-government websites: insights from the performance of the federal, state and local governments websites in Malaysia”. In: *Journal of Systems and Information Technology* 25.1 (2023). Export Date: 25 November 2024; Cited By: 2, pp. 109–132. DOI: 10.1108/JSIT-12-2021-0262. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150844288&doi=10.1108%2fJSIT-12-2021-0262&partnerID=40&md5=6bf472daf44dab14505ebe4feaf4e68f.

[18] Maia City Council. *Omega-X Project*. https://www.umaia.pt/pt/investigacao/projetos/2022/OMEGA-X. Accessed: 2024-10-01. 2022.

[19] Mariana Curado Malta. “Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web Semântica”. PhD thesis. 2014. URL: https://hdl.handle.net/1822/30262.

[20] A. Miletić, A. Kuveždić Divjak, and F. Welle Donker. “Assessment of the Croatian Open Data Portal Using User-Oriented Metrics”. In: *ISPRS International Journal of Geo-Information* 12.5 (2023). Export Date: 31 October 2024; Cited By: 3. DOI: 10.3390/ijgi12050185. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160288442&doi=10.3390%2fijgi12050185&partnerID=40&md5=7002e08ae625ff344b7

[21] Agência para a Modernização Administrativa. *dados.gov.pt – Portal de dados abertos da Administração Pública*. Accessed: 2025-06-10. 2025. URL: https://dados.gov.pt/en/.

[22] National Information Standards Organization (NISO). *Understanding Metadata*. Bethesda, MD: NISO Press, 2004. URL: https://www.niso.org/publications/understanding-metadata.

[23] Opendatasoft. *Opendatasoft Automation API v1 Documentation*. Accessed: 2025-06-05. 2025. URL: https://help.opendatasoft.com/apis/ods-automation-v1/.

[24] Opendatasoft. *Opendatasoft Explore API v2 Documentation*. Accessed: 2025-06-05. 2025. URL: https://help.opendatasoft.com/apis/ods-explore-v2.

[25] E. Papachristou and E. Gounopoulos. “Evaluation of data format quality of Open Government Data portals in southern EU countries”. In: *AIP Conference Proceedings*. Vol. 2909. Export Date: 25 November 2024; Cited By: 2. DOI: 10.1063/5.0184421. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179871582&doi=10.1063%2f5.0184421&partnerID=40&md5=5bafd1ee00b1b33cd1e638d471d03a09.

[26] Publications Office of the European Union. *Application Profiles*. https://op.europa.eu/en/web/eu-vocabularies/application-profiles. Accessed: 2025-06-28.

[27] A. Quarati. “Open Government Data: Usage trends and metadata quality”. In: *Journal of Information Science* 49.4 (2023), pp. 887–910. DOI: 10.1177/01655515211027775. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115263669&doi=10.1177%2f01655515211027775&partnerID=40&md5=a2d78a7a591c203eebd97d0bc93d9794.

[28] K. Rajamäe-Soosaar and N. Anastasija. “Exploring Estonia’s Open Government Data Development as a Journey towards Excellence: Unveiling the Progress of Local Governments in Open Data Provision”. In: *ACM International Conference Proceeding Series*. Export Date: 25 November 2024; Cited By: 0, pp. 920–931. DOI: 10.1145/3657054.3657161. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195262349&doi=10.1145%2f3657054.3657161&partnerID=40&md5=f7d00bc09381b09d3cdc2e73aa2d0214.

[29] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1 (2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[30] Marcia Lei Zeng and Jian Qin. *Metadata*. 2nd. ALA Neal-Schuman, 2016. ISBN: 9781555709655, 1555709656. URL: https://archive.org/details/metadata0000zeng_c8c4_2ed.

Appendix A

Intermediate Iteration of the Metadata Table

The following images show an intermediate iteration of the metadata table. Instead of converting it into a LaTeX table like the final table, images were used to better reflect the real state of the work, preserving the signs of collaboration with the CMM and of the ongoing process.

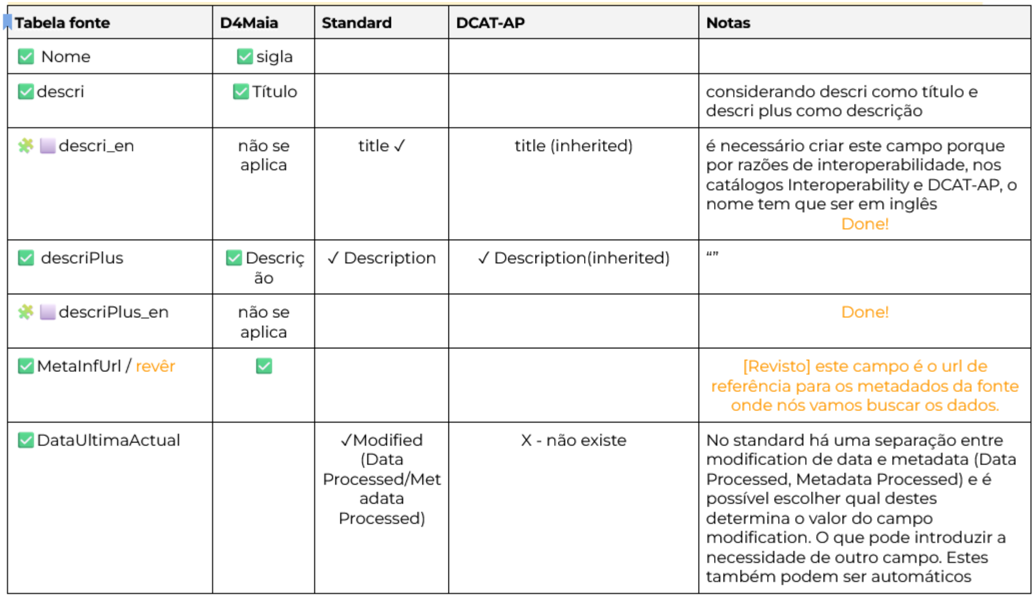
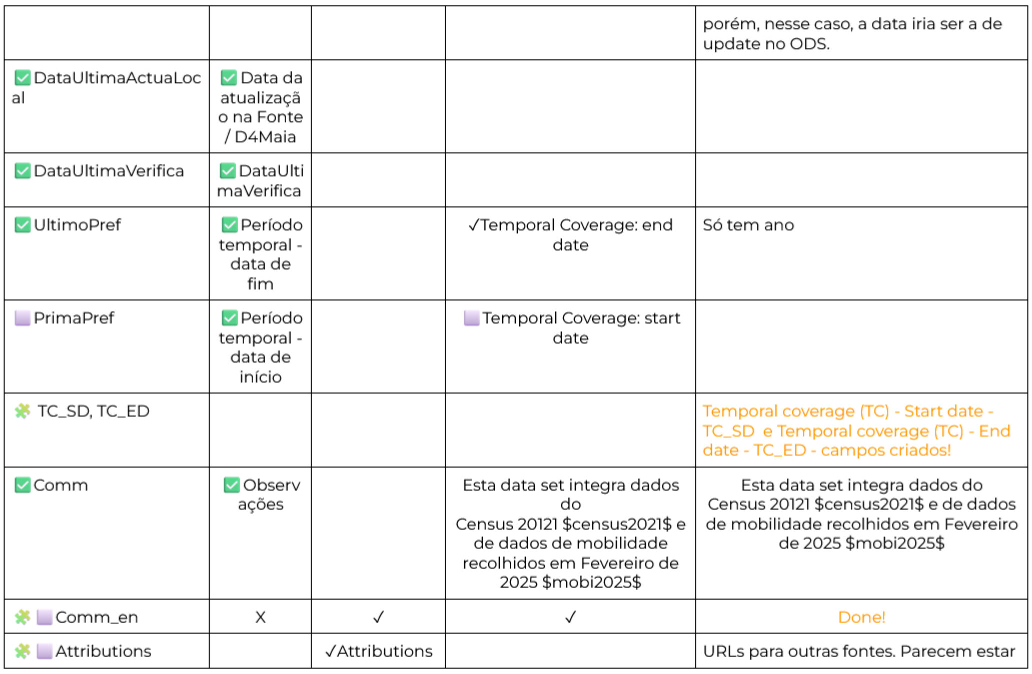


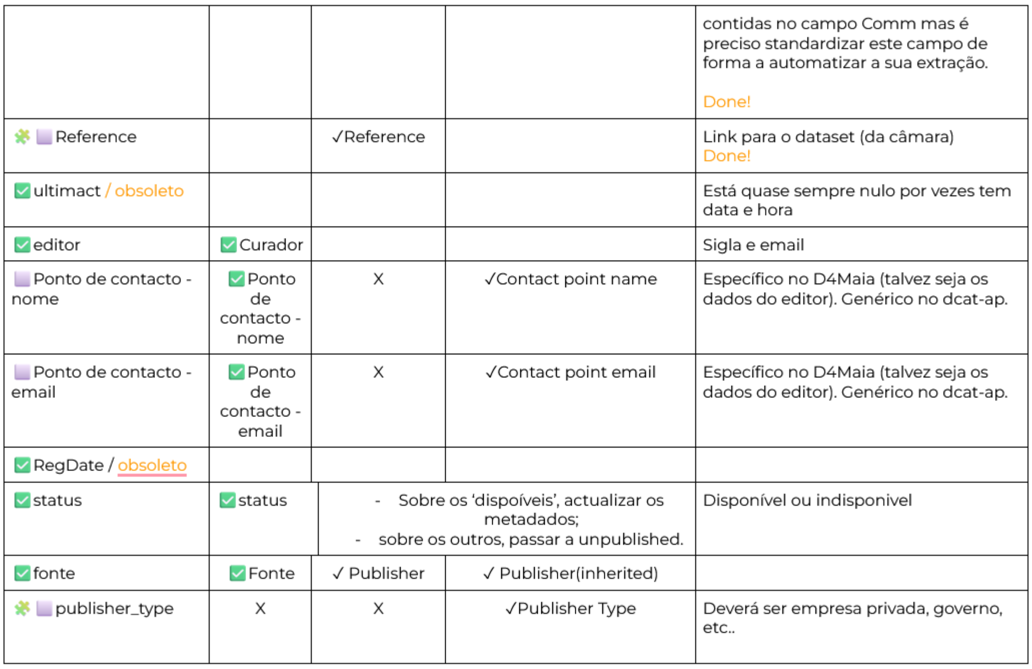
Tabela fonte	D4Maia	Standard	DCAT-AP	Notas
✅ Nome	✅sigla			
✅descri	✅Título			considerando descri como título e descri plus como descrição
🧩 🟪descri_en	não se aplica	title ✓	title (inherited)	é necessário criar este campo porque por razões de interoperabilidade, nos catálogos Interoperability e DCAT-AP, o nome tem que ser em inglês Done!
✅ descriPlus	✅Descrição	✓ Description	✓ Description(inherited)	“”
🧩 🟪descriPlus_en	não se aplica			Done!
✅MetaInfUrl / revêr	✅			[Revisto] este campo é o url de referência para os metadados da fonte onde nós vamos buscar os dados.
✅DataUltimaActual		✓Modified (Data Processed/Metadata Processed)	X - não existe	No standard há uma separação entre modification de data e metadata (Data Processed, Metadata Processed) e é possível escolher qual destes determina o valor do campo modification. O que pode introduzir a necessidade de outro campo. Estes também podem ser automáticos

Figure A.1: Intermediate iteration of the metadata table (Page 1)



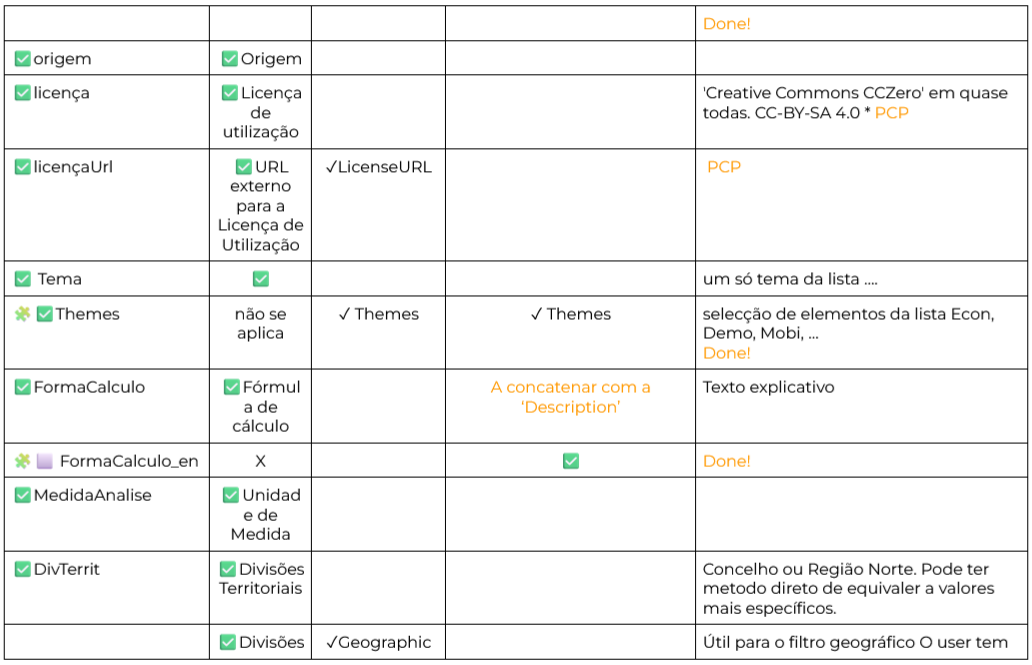
				porém, nesse caso, a data iria ser a de update no ODS.
☑ DataUltimaActuaLocal	☑ Data da atualização na Fonte / D4Maia			
☑ DataUltimaVerifica	☑ DataUltimaVerifica			
☑ UltimoPref	☑ Período temporal - data de fim		✓Temporal Coverage: end date	Só tem ano
🟪 PrimaPref	☑ Período temporal - data de início		🟪 Temporal Coverage: start date	
🧩 TC_SD, TC_ED				Temporal coverage (TC) - Start date - TC_SD e Temporal coverage (TC) - End date - TC_ED - campos criados!
☑ Comm	☑ Observações		Esta data set integra dados do Census 20121 $census2021$ e de dados de mobilidade recolhidos em Fevereiro de 2025 $mobi2025$	Esta data set integra dados do Census 20121 $census2021$ e de dados de mobilidade recolhidos em Fevereiro de 2025 $mobi2025$
🧩 🟪 Comm_en	X	✓	✓	Done!
🧩 🟪 Attributions		✓Attributions		URLs para outras fontes. Parecem estar

Figure A.2: Intermediate iteration of the metadata table (Page 2)



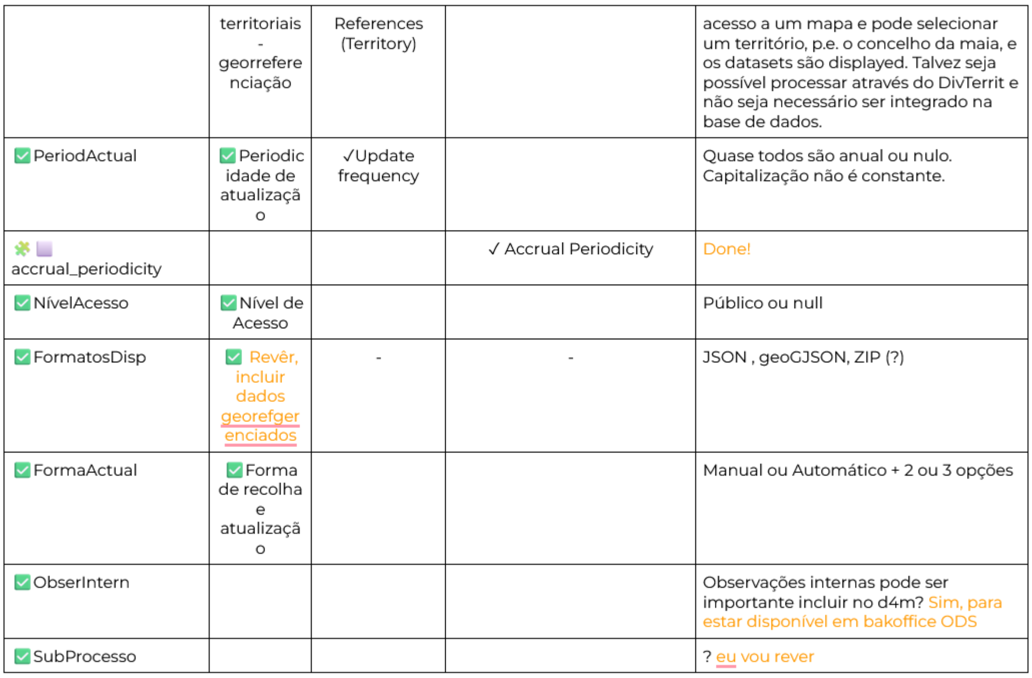
				contidas no campo Comm mas é preciso standardizar este campo de forma a automatizar a sua extração. Done!
🧩 🟪 Reference		✓Reference		Link para o dataset (da câmara) Done!
☑ ultimact / obsoleto				Está quase sempre nulo por vezes tem data e hora
☑ editor	☑ Curador			Sigla e email
🟪 Ponto de contacto - nome	☑ Ponto de contacto - nome	X	✓Contact point name	Específico no D4Maia (talvez seja os dados do editor). Genérico no dcat-ap.
🟪 Ponto de contacto - email	☑ Ponto de contacto - email	X	✓Contact point email	Específico no D4Maia (talvez seja os dados do editor). Genérico no dcat-ap.
☑ RegDate / obsoleto				
☑ status	☑ status	- Sobre os 'dispoíveis', actualizar os metadados; - sobre os outros, passar a unpublished.		Disponível ou indisponivel
☑ fonte	☑ Fonte	✓ Publisher	✓ Publisher(inherited)	
🧩 🟪 publisher_type	X	X	✓Publisher Type	Deverá ser empresa privada, governo, etc..

Figure A.3: Intermediate iteration of the metadata table (Page 3)



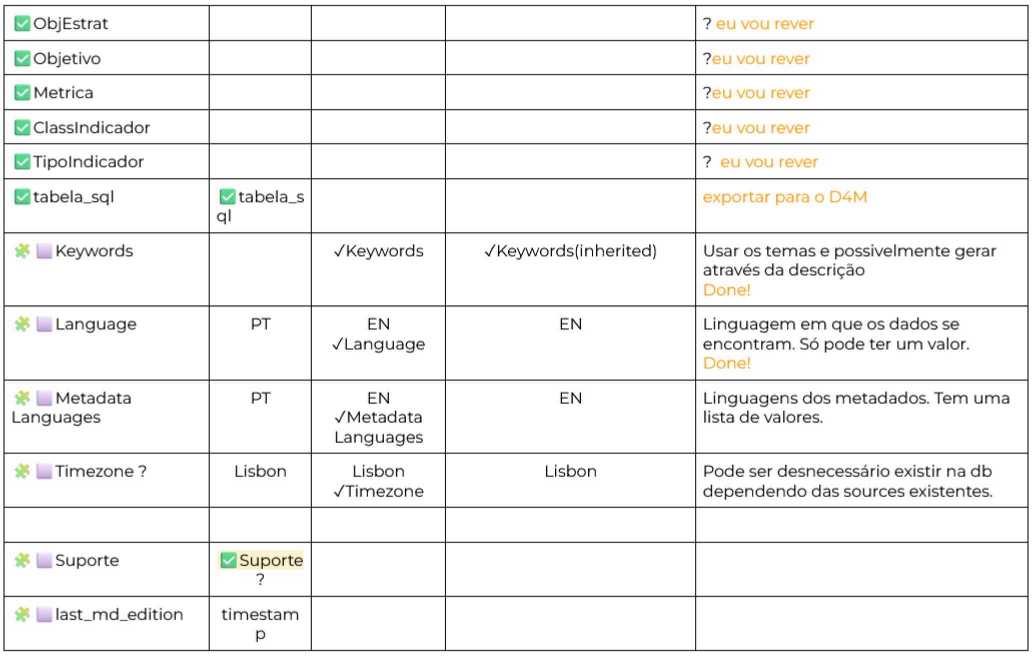
				Done!
☑origem	☑Origem			
☑licença	☑Licença de utilização			'Creative Commons CCZero' em quase todas. CC-BY-SA 4.0 * PCP
☑licençaUrl	☑URL externo para a Licença de Utilização	✓LicenseURL		PCP
☑ Tema	☑			um só tema da lista ....
🧩 ☑Themes	não se aplica	✓ Themes	✓ Themes	selecção de elementos da lista Econ, Demo, Mobi, ... Done!
☑FormaCalculo	☑Fórmula de cálculo		A concatenar com a 'Description'	Texto explicativo
🧩 🟪 FormaCalculo_en	X		☑	Done!
☑MedidaAnalise	☑Unidade de Medida			
☑DivTerrit	☑Divisões Territoriais			Concelho ou Região Norte. Pode ter metodo direto de equivaler a valores mais específicos.
	☑Divisões	✓Geographic		Útil para o filtro geográfico O user tem

Figure A.4: Intermediate iteration of the metadata table (Page 4)



	territoriais - georreferenciação	References (Territory)		acesso a um mapa e pode selecionar um território, p.e. o concelho da maia, e os datasets são displayed. Talvez seja possível processar através do DivTerrit e não seja necessário ser integrado na base de dados.
☑PeriodActual	☑Periodicidade de atualização	✓Update frequency		Quase todos são anual ou nulo. Capitalização não é constante.
🧩 🟪 accrual_periodicity			✓ Accrual Periodicity	Done!
☑NívelAcesso	☑Nível de Acesso			Público ou null
☑FormatosDisp	☑ Revêr, incluir dados georefgerenciados	-	-	JSON , geoGJSON, ZIP (?)
☑FormaActual	☑Forma de recolha e atualização			Manual ou Automático + 2 ou 3 opções
☑ObserIntern				Observações internas pode ser importante incluir no d4m? Sim, para estar disponível em bakoffice ODS
☑SubProcesso				? eu vou rever

Figure A.5: Intermediate iteration of the metadata table (Page 5)



☑ ObjEstrat				? eu vou rever
☑ Objetivo				?eu vou rever
☑ Metrica				?eu vou rever
☑ ClassIndicador				?eu vou rever
☑ TipoIndicador				? eu vou rever
☑ tabela_sql	☑ tabela_sql			exportar para o D4M
🧩 ☐ Keywords		✓Keywords	✓Keywords(inherited)	Usar os temas e possivelmente gerar através da descrição Done!
🧩 ☐ Language	PT	EN ✓Language	EN	Linguagem em que os dados se encontram. Só pode ter um valor. Done!
🧩 ☐ Metadata Languages	PT	EN ✓Metadata Languages	EN	Linguagens dos metadados. Tem uma lista de valores.
🧩 ☐ Timezone ?	Lisbon	Lisbon ✓Timezone	Lisbon	Pode ser desnecessário existir na db dependendo das sources existentes.
🧩 ☐ Suporte	☑ Suporte ?			
🧩 ☐ last_md_edition	timestamp			

Figure A.6: Intermediate iteration of the metadata table (Page 6)