FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

EUDAT Link: Integrating Dendro Platform With EUDAT, a Pan-European Data Management Network

Fábio Filipe Jesus da Silva

DISSERTATION REPORT

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Maria Cristina de Carvalho Alves Ribeiro

Co-Supervisor: Ricardo Filipe Carvalho Amorim

June 27, 2016

EUDAT Link: Integrating Dendro Platform With EUDAT, a Pan-European Data Management Network

Fábio Filipe Jesus da Silva

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Carla Alexandra Teixeira Lopes

External Examiner: Alberto Manuel Brandão Simões

Supervisor: Maria Cristina de Carvalho Alves Ribeiro

June 27, 2016

Abstract

There has been an increasing number of initiatives to make scientific research available to a broad range of users and contribute to better science. From data management plans that lead researchers to think about open access strategies beforehand, to repositories that make research results publicly available, all of them are part of the broad definition of *Open Science*. Recent guidelines are pushing researchers to actively manage their data and ensure it gets to an appropriate repository at the end of their work. Nevertheless, research data can't be understood without additional information about it, so context must be provided by researchers themselves, in order to facilitate its understanding by external parties. We call this information metadata. The domain-level description can only be done by researchers, who should do it ideally in early stages in their research workflow. They often do it using less conventional formats like free text on paper notes that easily get lost. Thus, a better description can be achieved if they are given appropriate tools to describe their data without requiring much effort or losing focus on their project.

The Dendro platform, created at the Faculty of Engineering in the University of Porto, is a collaborative and descriptive tool for researchers during the initial stages of data production, allowing them to make descriptions with domain vocabularies. They are more aware of their data's characteristics initially, so Dendro acts as a staging platform that can later connect to external repositories specifically designed for data preservation and dissemination on the long run.

EUDAT is proposed as a pan-European platform of reference for research data management, comprising a large set of tools that fulfill the main requirements within this field, plus having a set of experimental services that can be adopted at institutional level.

This work leverages Dendro's capabilities to handle data description across several domains and produce extensive metadata records, by providing means for data to be available, which means a deposit into EUDAT. Disclosure restrictions may be part of some research projects, so privacy levels were implemented within Dendro to protect sensitive data. This deposit is achieved by two different approaches. First, through an OAI-PMH server that exposes metadata from projects on a regular basis to EUDAT's B2Find module. Secondly, a project's data and metadata are packaged and sent to B2Share module upon request from the researcher, using Dendro's interface.

This project, which makes the bridge between institutional and international levels, assists researchers on the complex tasks associated with data deposit, which is accomplished by using all the standard-compliant metadata they already added on Dendro.

Our evaluation proved the effectiveness of this integration. After completing all the required tasks, researchers testing the platform made sure data and metadata were correct on each module. To test the dissemination aspect, they searched for specific terms and the datasets were retrieved, meaning that metadata sent by Dendro was essential for indexing purposes and that a proper description has an important role for the dissemination of data. Researchers felt our tools were fundamental, as they allowed them to properly describe data in a simple way that can be used as part of their routine, as well as then to export their projects to international repositories that provide identifiers ready to cite, with no additional effort.

Resumo

De modo a contribuir para uma ciência melhor, o número de iniciativas para tornar trabalhos científicos abertos à comunidade tem vindo a crescer. Desde planos de gestão de dados, que levam os investigadores a pensar previamente acerca de estratégias de acesso aberto, até repositórios que disponibilizam publicamente dados de investigação, todas elas fazem parte da vasta definição de *Open Science*. Recentes directrizes têm vindo a levar os investigadores a gerir activamente os seus dados e a assegurar que estes são depositados num repositório apropriado no final do projecto. No entanto, os dados de investigação não podem ser compreendidos por terceiros sem informações adicionais acerca destes. Por esse motivo, os investigadores devem dar-lhes contexto usando metadados. A descrição ao nível do domínio deve ser efectuada pelos investigadores, que o devem fazer idealmente em fases iniciais da investigação. Por vezes já tendem a fazê-lo usando formatos pouco convencionais, tais como notas em papel que facilmente se perdem ou deterioram. Como tal, melhores descrições podem ser alcançadas se os investigadores tiverem ao dispôr ferramentas apropriadas para tal, sem que lhes seja requerido muito esforço ou que percam o foco no seu trabalho.

A plataforma Dendro, criada na Faculdade de Engenharia da Universidade do Porto, é uma ferramenta colaborativa e descritiva focada nos investigadores durante a fase inicial de produção de dados, permitindo-lhes efectuar descrições com vocabulários de domínio, quando estão mais cientes das características dos dados. Desta forma, o Dendro funciona como uma área inicial que pode depois conectar-se a repositórios externos, especialmente concebidos para a preservação e disseminação a longo prazo.

O EUDAT pretende ser a plataforma pan-Europeia de eleição para a gestão de dados de investigação. Consiste numa série de ferramentas que cumpre com os principais requisitos desta área, possuindo alguns serviços experimentais que podem ser adoptados a nível institucional.

Este trabalho eleva as capacidades do Dendro para cuidar da descrição de dados em vários domínios e produzir extensos registos de metadados, fornecendo meios para que os dados estejam depois disponíveis, o que significa depositá-los no EUDAT. Restrições de divulgação podem fazer parte de alguns projectos de investigação, pelo que foram implementados níveis de privacidade no Dendro de forma a proteger dados sensíveis. O depósito é alcançado por duas abordagens diferentes. Primeiro, através de um servidor OAI-PMH que expõe regularmente os metadados de projectos ao módulo B2Find do EUDAT. Em segundo lugar, os dados e metadados são empacotados e enviados directamente através da interface do Dendro para o módulo B2Share assim que investigador assim desejar.

Este projecto, que faz a ligação entre os níveis institucional e internacional, assiste os investigadores durante a árdua tarefa de depósito dos mesmos. Tal é conseguido através do uso dos metadados no Dendro, que seguem padrões e que foram previamente fornecidos pelo próprio investigador.

A avaliação atestou a eficácia desta integração. Após efectuarem as tarefas necessárias, os investigadores certificaram-se que todos os dados e metadados estavam correctos nos módulos

do EUDAT. Para testar o aspecto da disseminação, pesquisaram por termos específicos e os dados foram retornados, o que significa que os metadados do Dendro foram importantes e que a descrição dos dados tem um papel importante para a disseminação destes. Os investigadores sentiram que tanto o Dendro como o módulo de integração foram fundamentais, dado que permitiram descrever eficazmente os seus dados, assim como exportar os projectos para repositórios internacionais que fornecem identificadores prontos a citar, sem esforço adicional por parte deles.

Acknowledgments

For the given support and guidance during the development of my thesis, I would like to thank my supervisors, Cristina Ribeiro and Ricardo Amorim.

Fábio Filipe Jesus da Silva

"A curiosidade, instinto de complexidade infinita,
leva por um lado a escutar às portas e por outro a descobrir a América;
- mas estes dois impulsos, tão distintos em dignidade e resultados,
brotam ambos de um fundo intrinsecamente precioso,
a actividade do espírito"

Eça de Queiroz

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

AIP	Archival Information Package
API	Application Programming Interface
CERN	*Conseil Européen pour la Recherche Nucléaire*
CKAN	Comprehensive Knowledge Archive Network
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
DIP	Dissemination Information Package
DMP	Data Management Plan
DOI	Digital Object Identifier
EGI	European Grid Infrastructure
EML	Ecological Medatata Language
EUDAT	European Data
HPC	High-Performance Computing
HTTP	Hypertext Transfer Protocol
ICPSR	Inter-university Consortium for Political and Social Research
IRI	International Resource Identifier
KML	Keyhole Markup Language
MARC	Machine Readable Cataloging
METS	Metadata Encoding and Transmission Standard
MODS	Metadata Object Description Schema
NSF	National Science Foundation
OAI-PMH	Open Archives Information Protocol for Metadata Harvesting
OAIS	Open Archives Information System
OWL	Web Ontology Language
PDI	Preservation Description Information
PRACE	Partnership for Advanced Computing in Europe
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RDM	Research Data Management
REST	Representational State Transfer
SaaS	Software as a Service
SIP	Submission Information Package

Chapter 1

Introduction

With the large volume of data produced nowadays and all the forms it can assume, science is becoming more and more *data intensive*, which means data is getting an important role on the researchers' routines and science uses this data, so new infrastructures and tools—*e-Science*—should be developed to support this new paradigm, often called the *Fourth Paradigm* [Bel09, TTH11]. As the amount of produced data increases and concepts like Big Data arise, long-term preservation of this data is becoming a systematic challenge, which can undermine its chances of being properly preserved and disseminated.

As result, the preservation workflow has been improved with many standards and methods. As this evolution still occurs, the collaboration among researchers from all the globe becomes easier, and this means better, reusable data and well-documented research.

The current research data management workflow is constantly improved to cover aspects such as metadata standards, collaboration and data fitness for reuse. Recent studies have shown that some institutions can struggle to keep up with the requirements from this nature.

1.1 Research Data

The increasing amount of produced data poses new challenges concerning data sharing and reuse. On one hand, data reuse is becoming a possibility due to initiatives like Open Access. On the other hand, for a researcher to feel comfortable reusing data, an extensive description is required. Datasets come from several different fields and have different formats. Even within the same research group, data can assume several formats, as it can come from sources like observations, experiments or simulations. Therefore, giving a proper context to data is key for future interpretations from third parties. For this correct contextualization, one needs a deep knowledge about data and its creation process.

In this work, the focus is on the *Long Tail of Science*, i.e. small groups of researchers that have few resources available but produce data from many different domains.

Curators are usually responsible for the description task, providing information that can contextualize the dataset's nature. *Curation* involves maintaining and adding value to digital research data, enhancing the long-term value of datasets [Dig13]. However, most of the time curators have a limited knowledge of the domain, so the risk of losing data increases—other researchers searching for datasets may not find them due to the absence of domain information that is essential for their retrieval. Commonly, their role is to gather basic information like authors and abstracts.

Researchers, who are the data producers, must be involved in the data description as soon as possible, using appropriate metadata that are normally provided by well established metadata schemas [CRR14], and even ontologies for a high-level solution that covers all requirements about data description on a specific domain [DHSW02].

1.2 Data Repositories

Besides data description, it is important for a research community to choose a data repository that fits its needs across the data management workflow [AR12]. By doing so, researchers will allow external communities to find, understand and reuse the produced data, contributing to a collaborative research environment.

Most of the research data management platforms like CKAN[1], Zenodo[2] or DSpace[3] are implemented to collect "final" data that can be already cited, so they focus on receiving data at the end of the research workflow. Besides that, these repositories often limit the metadata that can be used, by confining it to general schemas (e.g. Dublin Core[4]) or predefined descriptors. CKAN is more flexible, since it allows some arbitrary metadata to be added, although not ensuring compliance with standards [RRC14].

Describing data on a regular basis is important, starting as soon as it's collected, since it will create more domain metadata gathered at that moment. Leaving this task for later stages can increase the risk of forgetting or losing particular aspect of data, so an early description can be way more effective.

In order to bring the data description to earlier stages in the workflow, institutions can implement a staging area and involve researchers in the process. Dendro[5], developed at the Faculty of Engineering of the University of Porto, is a collaborative platform that acts as a staging area. It uses an underlying ontology-based data model and provides a simple Web interface, so researchers can participate in the process in a simplified way and provide rich descriptions of their data at the same time.

Figure 1.1 shows a workflow that starts by data creation and ends on a final repository [RRC14]. Initially, data is produced, described and shared (1). Here, the description of data is made using descriptors organized in ontologies. After that, data and corresponding metadata are packaged and

[1] http://ckan.org/
[2] https://zenodo.org/
[3] http://www.dspace.org/
[4] http://dublincore.org/
[5] http://dendro.fe.up.pt/

deposited into a final repository (2). As new domain metadata specifications evolve and are shared on the Web (3), new ontologies for research domains can emerge as standards. Finally, data can be reused by other researchers (4).

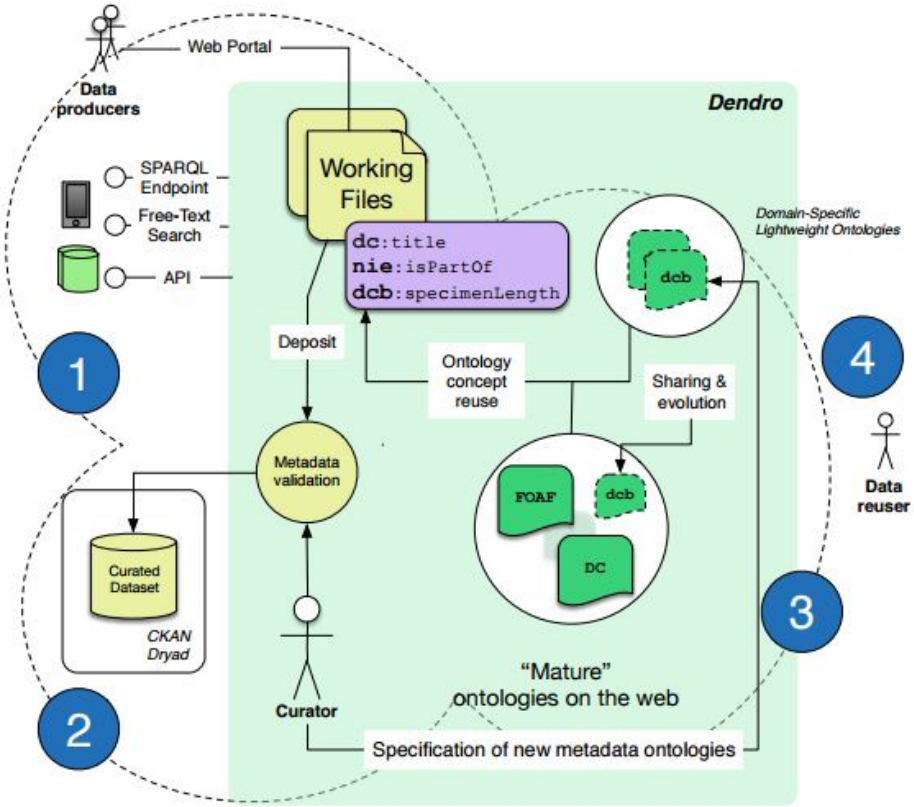

Figure 1.1: Dendro platform within RDM workflow. Available from [RRC14]

.

1.3 EUDAT Link: Integrating Dendro with EUDAT

Dendro is a staging area where researchers can work in a collaborative way to describe their data and transfer their expertise in the domain to standards-compliant metadata records, thus increasing the probability of data preservation in the long run.

Nevertheless, Dendro is not designed to address data preservation. It focuses instead in providing adequate conditions for research data to be deposited and later interpreted, so data must keep its flow towards a final repository where it can be preserved and disseminated. To achieve that, all the data and corresponding metadata must be sent to a final repository like CKAN, FigShare[6] or EUDAT. To address European requirements to deal with funded research projects, EUDAT provides a set of services that cover a part of the research workflow. Its modules integrate well known platforms such as CKAN and Invenio[7] to address specific needs. It's financed by Horizon2020

[6] https://figshare.com/
[7] http://invenio.readthedocs.org/en/latest/

program and has some well recognized partners like CERN, which can lead to a great visibility and potentially influence communities of European researchers to use it. Moreover, it is a remote solution that provides a functional product for whoever wants to use it readily, meaning no need to implement and maintain institutional solutions with several costs associated.

EUDAT Link aims at integrating the Dendro platform with EUDAT, creating a fully functional unidirectional bridge between the two. This module will make sure there is interoperability on each side, by validating data structures and making use of established standards. This module is built to fit within the research data management workflow, allowing research communities to export their data and descriptors to a long-term international repository—EUDAT—in an easy and flexible way. Besides this, such integration is part of a pilot program known as *Data Pilot*[8], which guarantees the project has the required support from EUDAT. In particular, this integration is part of the Data Pilot named *DataPublication@U.Porto*[9], that provided support during the development stages.

1.4 Document Structure

This document is segmented into five more chapters after this introduction. Chapter 2 refers to existing approaches for the entire data management workflow. From staging platforms to metadata standards and final repositories, all important concepts are defined. Section 2.1 focuses on metadata models and concepts like ontologies, while section 2.2 defines the OAIS standard for preservation. The section 2.3 gives a perspective about the OAI-PMH standard for metadata harvesting, while section 2.4 tells apart different kinds of repositories. Finally, section 2.5 explores Data Management Plans, a new concept that is emerging as a deliverable for research projects and section 2.6 points out some conclusions for the chapter.

Chapter 3 presents EUDAT Link as a solution for an integration problem involving Dendro and EUDAT, with section 3.1 explaining the workflow and its parts, namely Dendro, LatTablet, B2Share and B2Find. The chapter ends with some conclusions at section 3.2.

A more detailed explanation on the development of this solution can be found in chapter 4. As this integration was developed using a two-way approach, sections 4.1 and 4.2 provide in-depth details and concepts for each one.

In order to evaluate the proposed solution, we have invited a set of researchers to test Dendro from the very beginning until the final deposit in B2Share. Chapter 5 provides all the information about these tests and the tasks involved.

Finally, chapter 6 enumerates some conclusions and possible improvements to this project.

[8] http://eudat.eu/eudat-call-data-pilots
[9] https://www.eudat.eu/communities/datapublication-uporto

Chapter 2

Research Data Management

With the increasing number of data collections produced by small research groups with few resources, the term *Long Tail of Science Data*—vast number of smaller size collections generated by them—gets real and there is a need to standardize these highly heterogeneous datasets [PCHS07]. Metadata is then used for this purpose, as there are several standards already defined. While some metadata follow generic guidelines that fit almost any kind of data, some particular information, usually known by the person who created data, may be needed for specific domains like Ecology or Mechanics. Moreover, it is extremely important to know which metadata standards are available, so the future interpretation can be done correctly and repositories can exchange information without interoperability problems. Regarding this last point, some protocols must be followed between repositories, allowing the correct exchanging of data and metadata.

2.1 Metadata - First Step Towards Preservation

Typically, metadata is recurrently defined as *data about data*. Usually, metadata standards are followed, giving context to objects they portray [Gre03]. These standards emerged as a response to the dynamic nature of resources, as they are constantly changing and evolving, even more when they got to the Web [Gil08], so there was a need to implement patterns that should be followed in order to achieve an effective data preservation. Metadata can be divided into categories:

- **Descriptive** - describe and identify resources. Examples: identifiers (DOI[1], Handle[2]), physical attributes, specific domain information and bibliographic attributes (as title, author and language);

- **Structural** - responsible for providing details about a resource's internal structure. Examples: table of contents, chapters and index;

[1] https://www.doi.org/
[2] https://www.handle.net/

- **Preservation** - information for long-term preservation purposes, needed by repositories and that follow guidelines during its life cycle within workflows.

Metadata is frequently used to describe data in general, so it is used to describe research data as well. Researchers must then choose descriptors for their data, giving them context. For that, several models are already defined. Metadata models provide descriptors that allow the correct contextualization of a particular object. Many of them provide a general overview of data and can be used by almost everybody. However, others are very specific and most of the times can only be used by data creators, who are the experts about it.

2.1.1 Generic Metadata Models

These models provide descriptors that fit several kinds of data and resources. Therefore, almost no specific knowledge on the data domain is required, but only a general overview of it. Although they are fit for a wider set of domains, the context they provide is limited to high-level, administrative information such as the author and creation date.

2.1.1.1 MARC

Bibliographic records were created manually by librarians before the creation of computer systems. This way, each item in each library should have its own ID card, which was a clear repetition of work. After the invention of the computer, this task became easier, as the ID was created only once and could be shared later [Fur09]. As a result, MARC[3] (Machine Readable Cataloging) was created by the Library of Congress. From that point on, librarians would have to only buy the ID card from the Library of Congress and print it—that card was based on the item's corresponding MARC record. Several formats were implemented in several countries based on MARC, so the MARC21 standard was created as a result of the harmonization of the main ones. Nowadays, this is the most used standard in libraries' information systems [Fur09].

Each record has several fields with different purposes, such as identifying the author or the edition number. These fields are identified by a 3 digit code. Table 2.1 and Figure 2.1 show the most used fields and a practical example of one of those fields, respectively. More recently, MARCXML[4] standard was created, which is the corresponding XML schema. This was done so that MARC records could be exposed via Web, as XML is one of the standard formats for data interchanging.

[3] https://www.loc.gov/marc/
[4] https://www.loc.gov/marc/marcxml.html

Table 2.1: Most used MARC fields

Field	Description
010	Library of Congress Control Number (LCCN)
020	International Standard Book Number (ISBN)
100	Author
245	Title information
250	Edition
260	Publication information
300	Physical description
440	Series statement/added entry
520	Annotation or summary entry
650	Topical subject heading
700	Personal name added entry (joint author, editor or illustrator)

```
100 1# $a Pirsig, Robert M.
```

Figure 2.1: Example of MARC's field *100* (author). Adapted from [Fur09]

2.1.1.2 Dublin Core

Dublin core schema is one of the most used as a general description standard and it is maintained by the Dublin Core Metadata Initiative (DCMI), which was created in 1994. It was designed due to the need to describe Web resources. By 1999, 15 basic descriptors—Dublin Core Metadata Element Set[5] (DCMES) [ISO09]—were published, becoming an international standard in 2003 [Har10]. Descriptors like *title*, *creator*, *date*, *language* and *identifier* are part of this set. Figure 2.2 shows the description of an article using two triples that use Dublin Core for their predicates.

[5] http://dublincore.org/documents/dces/

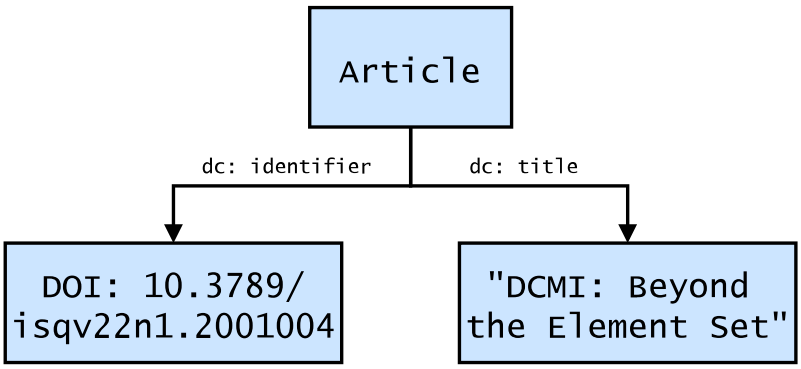

Figure 2.2: Describing an article using *dc:identifier* and *dc:title*. Adapted from [Har10]

2.1.1.3 METS

Metadata Encoding and Transmission Standard[6] (METS) is an initiative being developed by the Library of Congress Network Development and MARC Standards Office and it is focused on creating a pattern that promotes interoperability, scalability and digital preservation of digital objects [Can05] and to group several resources. It is expressed through an XML *schema*, and each METS document is divided into 7 sections—*METS Header*, *Descriptive Metadata*, *Administrative Metadata*, *File Section*, *Structural Map*, *Structural Links* and *Behavioral*—with different roles. Only *Structural Map* section is mandatory, as it is responsible for modeling the object's hierarchical structure and allows an easier navigation within the document, since it makes the connection between structure's elements to contents, as shown in Listing 2.1 [Can05].

```
<mets:structMap TYPE="PHYSICAL">
 <mets:div ID="ksldigbks" TYPE="digital book collection" LABEL="DIGITAL CASE E-BOOK
     COLLECTION">
  <mets:div ID="bk1" TYPE="BOOK">
   <mets:mptr LOCTYPE="URL" xlink:href="http://path/bk1-mets.xml"/>
  </mets:div>
  <mets:div ID="bk2" TYPE="book">
   <mets:mptr LOCTYPE="URL" xlink:href="http://path/bk2-mets.xml"/>
  </mets:div>
  <mets:div ID="bk3" TYPE="book">
   <mets:mptr LOCTYPE="URL" xlink:href="http://path/bk3-mets.xml"/>
  </mets:div>
 </mets:div>
</mets:structMap>
```

Listing 2.1: *Structural Map* section of a METS document

[6] http://www.loc.gov/standards/mets/

2.1.1.4 MODS

MODS[7] (Metadata Object Description Schema) format, also developed by the Library of Congress Network Development and MARC Standards Office, is an XML schema that has a subset of the MARC21's descriptors. Nevertheless, it has a simpler syntax than MARCXML, since the latter performs a direct conversion from MARC. The main reason to propose MODS was to have the possibility of providing a simplified version of MARC21 that could be easier and faster to create, using XML as format for interoperability with Web applications [Gue03]. Compared to Dublin Core, for example, MODS' creators claim that it provides a more complete description of assets, due to being more structured with sub-elements. Table 2.2 presents the correspondence between MODS elements and Dublin Core [Gue03].

Table 2.2: Mapping between some elements of MODS and DCMES

MODS	**Dublin Core**
titleInfo[mandatory]	*Title*
name	*Creator / Contributor*
typeOfResource / genre	*Type*
publicationInfo	*Publisher / Date*
physicalDescription	*Format*
language	*Language*
abstract	*Description*
tableOfContents	*Description*
note	*Description*
subject	*Subject*
classification	*Subject*
relatedItem	*Relation*
identifier	*Identifier*

All the above models are used in a general way to describe data and resources. When it comes to research data that is often part of a very specific domain like Biology or Mechanics, these models reveal a lack of terms to describe data correctly, so there is a need for more specific ones.

2.1.2 Domain Specific Metadata Schemas

When it comes to specific domains, using generic metadata models will not be enough for the contextualization needs, which may lead to an incorrect interpretation of data. This gap led to the development of specialized models that could provide more precise details about domain datasets.

2.1.2.1 EML

EML[8] (Ecological Metadata Language) format is a model that addresses the Ecology subject and it is maintained by the Knowledge Network for Biocomplexity. It is a model built specifically for ecologists, something that did not exist until EML, in spite of the existence of other

[7] http://www.loc.gov/standards/mods/
[8] https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

patterns aimed at geospatial data that were recommended by the community of geographical sciences [MBH+97]. As other previous formats, this one is also designed using XML. Listing 2.2 exemplifies a simple EML document [KNB].

```
<?xml version="1.0"?>
<eml:eml
    packageId="eml.1.1" system="knb"
    xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 eml.xsd">

  <dataset id="ds.1">
    <title>Sample Dataset Description</title>
    <creator id="23445" scope="document">
      <individualName>
        <surName>Myer</surName>
      </individualName>
    </creator>
    ...
  </dataset>
</eml:eml>
```

Listing 2.2: Valid EML document

2.1.2.2 KML

Keyhole Markup Language, (KML[9]), maintained by the Open Geospatial Consortium (OCG), is a model used to encode representations of geographical data, which allows users to overlay their own content, like pinpoints, on top of base maps and satellite images on Earth browsers like Google Earth [SG08]. Like other XML formats, it uses a tree structure to represent its elements, which is represented in Figure 2.3. Both the KML document and corresponding files, like images and 3D objects, may be compressed into a zipped file with the *.kmz* extension. Besides Google Earth, KML files can be understood by applications like NASA WorldWind, ESRI ArcGIS Explorer, Adobe PhotoShop, AutoCAD and Yahoo! Pipes. Listing 2.3 shows an example of a pinpointed location using this schema.

[9] https://developers.google.com/kml/

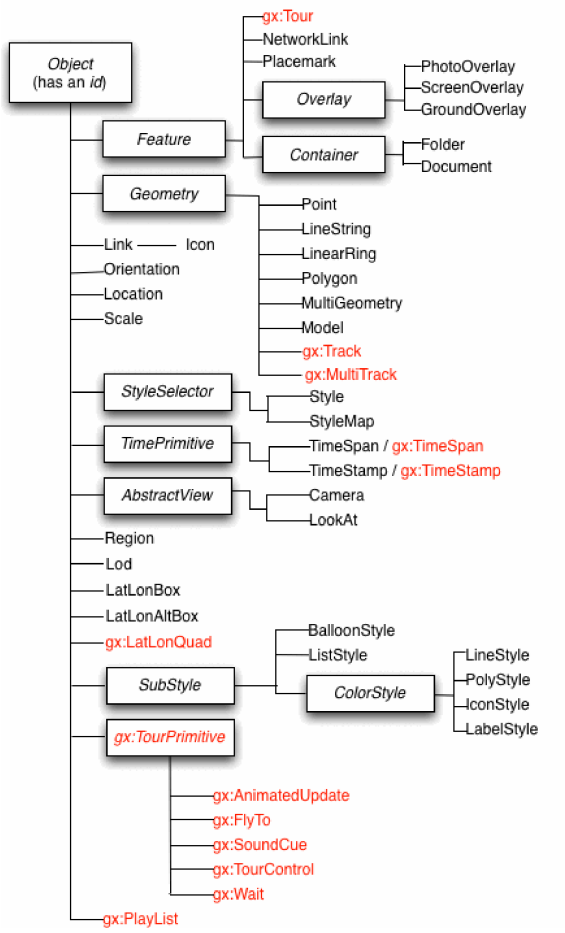

Figure 2.3: KML elements structure. Available from KML's website

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
<Placemark>
 <name>Simple placemark</name>
 <description>Attached to the ground. Intelligently places itself at the height of
     the underlying terrain.
 </description>
 <Point>
 <coordinates>-122.0822035425683,37.42228990140251,0</coordinates>
 </Point>
 </Placemark>
 </kml>
```

Listing 2.3: Location marker using KML

To sum up, it is fundamental to make a reasoned decision when choosing a model in order to allow data to endure in time and to be disseminated. Table 2.3 shows some other domain metadata models besides the ones mentioned above.

Table 2.3: Domain metadata models

Name	Subject	Maintained by
AgMES (Agricultural Metadata Element Set)	Food, Nutrition & Rural Development	ONU
CF (Climate and Forecast)	Climate Data & Forecasting	NetCDF Climate and Forecast Metadata Convention
Darwin Core	Biology	Biodiversity Information Standards (TWDG)
DDI (Data Documentation Initiative)	Social, Behavioral & Economic Sciences	DDI Alliance
INSPIRE (Infrastructure for Spatial Information in the European Community)	Spacial Data	European Commission
ISAD(G) (General International Standard Archival Description)	Archives Description	International Council on Archives (ICA/CIA)

Sometimes, there are relationships and meanings between data. As an example, several datasets are often related to the same project, building the concept of an underlying model of files and folders [CRR14]. Using metadata models as shown before does not allow researchers to represent this network of connections, so something at a higher level is needed, like ontologies.

2.1.3 RDF and Ontologies

An ontology allows to gather descriptors, i.e. terms used to describe data, from several metadata models, creating then a richer cross-domain environment. Instead of describing data using key-value pairs, researchers can use domain ontologies, which are built on top of RDF[10] (Resource Description Framework). By doing so, they are giving structure and sense to data, creating meaningful connections between resources. This approach allows a simpler data model and a superior flexibility, as it can grow as more ontologies for different domains are created [RRC14].

2.1.3.1 RDF

RDF is a model used to describe Web resources, from documents to physical objects, so they can be processed by machines and applications. By using IRIs (International Resource Identifier) from different contexts and schemas, it is possible to describe resources in an extensive way. In order to do so, data from IRIs is used, contributing to the popularity of *Linked Data* [Bl06]. Resources are described through triples using the syntax <*subject*> <*predicate*> <*object*> and an IRI can be used for each one of the three. Figure 2.4 illustrates a practical use of this framework,

[10] https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

where both Dublin Core and FOAF[11] (*Friend of a Friend*) are used. These triples can then be serialized using formats like TURTLE[12], JSON-LD[13] or XML.

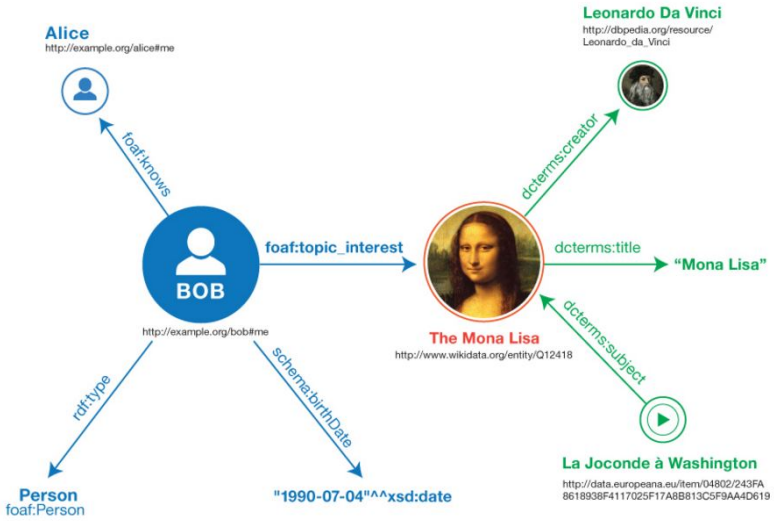

Figure 2.4: Informal RDF graph. Available from [SR14]

2.1.3.2 Ontologies

An ontology can be defined as the conceptualization of a domain, capturing its structure and restrictions, by specifying concepts and forming a vocabulary and axioms to express the knowledge base [Obi07]. Using a simpler definition, it can be assumed that an ontology is the representation of the domain.

Although triples can be represented using RDF, this framework only creates an underlying graph representing data, so there are no relationships between properties nor definition of classes and corresponding relationships. An ontology may include formal constraints of the terms in a vocabulary, as well as relationship among terms. There are some features implemented by ontologies that make this possible, namely [Bas12]:

- **Hierarchical relationships** - expressing hierarchies among terms, e.g. sub-classes;
- **Characteristics of properties** - defining properties using characteristics known to machines. Example: inverse properties, transitive properties, symmetric properties and functional properties;
- **Inference** - constraints may allow facts to be inferred without being explicit.

[11] http://xmlns.com/foaf/spec/
[12] https://www.w3.org/TR/turtle/
[13] http://json-ld.org/learn.html

In order to express ontologies, standards like RDFS[14] (RDF Schema) and OWL[15] (Web Ontology Language), which extends the former, may be used. Listing 2.4 shows a simple OWL class with an underlying RDFS and RDF structure [AH11].

```
<owl:Class rdf:ID="Agent">
 ...
 <rdfs:subClassOf rdf:resource="pers:Person"/>
 <owl:Restriction>
 <owl:onProperty rdf:resource="#affiliate"/>
 <owl:mincardinality rdf:datatype=
 "xsd:nonNegativeInteger">1</owl:cardinality>
 </owl:Restriction>
 </rdfs:subClassOf>
</owl:Class>
```

Listing 2.4: OWL class

Metadata models and ontologies themselves do not cover all long-term preservation requirements. They only fulfill description-related requirements, so factors like long-term preservation and dissemination are not covered. After data is described, it must go to a repository, along with metadata, where they can be preserved and disseminated. When it comes to how digital assets should be preserved into repositories, it is important to have standard guidelines that define how data is preserved after its ingestion and how it is then managed and redistributed.

2.2 OAIS Model for Digital Archives Management

In order to understand and apply necessary concepts to long-term digital preservation, the Open Archive Information System (OAIS) was created [SRGM04]. Developed by the Consultative Committee for Space Data Systems (CCSDS), it became an international standard on 2003 [ISO12] and is widely used as a standard among data repositories. It defines an architecture composed by several actors with different responsibilities on preservation and availability of an archive's information. In this particular context, we can assume the archive is the final repository, so this model defines how data and metadata must be sent to the repository, how they are managed within it and how they are retrieved for consumers later.

2.2.1 OAIS Entities

These are representations of the involved entities when dealing with data management and define specific responsibilities regarding their contribute to the overall workflow, as illustrated in Figure 2.5. Moreover, they can either represent people that directly interact with the objects, established protocols or standards that help maintaining a cohesive communication between the involved parts.

[14] https://www.w3.org/TR/rdf-schema/
[15] https://www.w3.org/2001/sw/wiki/OWL

- **Producer** - people or systems that provide information and metadata to be preserved, e.g., researchers or institutions;
- **Management** - control policies. Not involved in daily activity related to the archive;
- **Consumer** - people or systems that interact with OAIS services to get preserved information. There is a subset of consumers, known as *Designated Community*, composed by the people that are able to understand the preserved information [Ccs02]. In this case, we are talking about researchers.

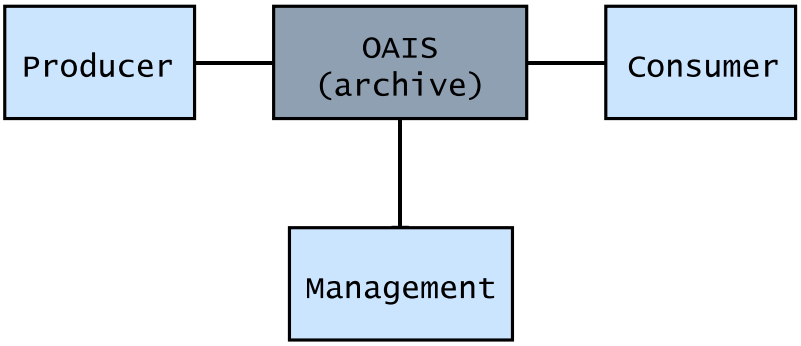

Figure 2.5: OAIS external environment. Adapted from [Ccs02]

2.2.2 Information Packages

Within this model, *Information Package* is defined as a package including *Content Information* and *Preservation Description Information* (PDI). The former is about the information itself (physical or digital object), whilst the latter represents the necessary information to preserve *Content Information* and may be one of five types:

- **Provenance** - source of *Content Information*;
- **Context** - description of relationships between *Content Information* and other information external to the *Information Package*;
- **References** - one or many identifiers that uniquely map *Content Information*;
- **Immutability Information** - strategies to protect information from non-documented changes, such as *checksums*;
- **Access rights** - Access terms, including preservation, distribution and use of *Content Information*.

2.2.2.1 Information Package Types

There are differences between *Information Packages* being preserved, submitted and distributed within OAIS. Three types can be distinguished, as shown in Figure 2.6. Figure 2.7 shows the overall workflow of this model.

- **Submission Information Package** (SIP) - provided by the *Producer*. Usually, it contains *Content Information* and the PDI;
- **Archival Information Package** (AIP) - one or many SIPs are converted into ore or more AIPs, which contains PDIs for corresponding *Content Information*;
- **Dissemination Information Package** (DIP) - upon Consumer request, all or part of an AIP is retrieved as a DIP, which can even contain several AIPs.

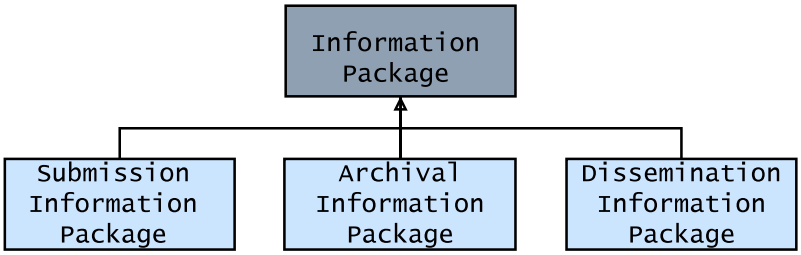

Figure 2.6: Information Package types. Adapted from [Ccs02]

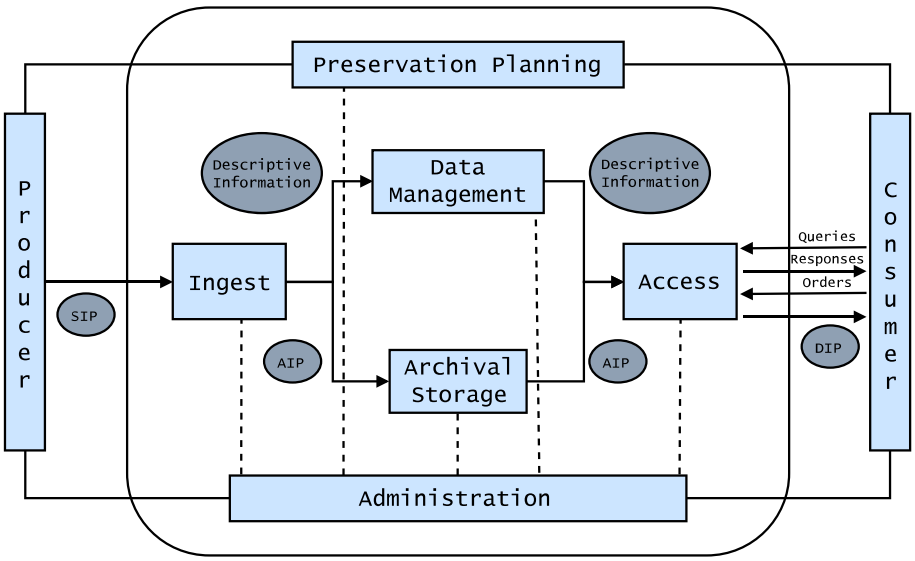

Figure 2.7: OAIS functional model. Adapted from [Ccs02]

Figure 2.7 shows that there are more entities involved throughout the OAIS processes, so it is convenient to point out the main tasks performed by each one of them:

- **Ingest** - it takes a SIP from *Producers*, preparing the contents for *Storage* and *Management*;
- **Archival Storage** - responsible for storing, managing and returning the AIP. After getting the AIP from *Ingest*, the package is stored permanently and some error inspection is done routinely. Besides that, it must send the AIP for *Access*;

- **Administration** - in general, it tries to ensure archive's system operability. In order to do so, negotiations are made with *Producers* to fulfill defined standards;

- **Preservation Planning** - it monitors the OAIS environment, trying to ensure that stored information keeps accessible and understandable, even if some used technologies become obsolete. Moreover, it builds templates for *Information Packages*, making it easier to implement them in SIPs and AIPs;

- **Access** - provides assistance to *Consumers*, allowing them to request and receive information. During this process, access constraints may be applied to protected information.

As the OAIS model is very abstract, in the sense that it defines concepts without elaborating much about technical questions, it can be adapted to any workflow. Within the context of this project, we can set researchers as *Producers* and *Consumers*, while EUDAT—the final repository—is our OAIS archive. Researchers send both data and metadata (SIP) that will be stored within EUDAT (AIP) and disseminated to other researchers (DIP).

With the fast growing activities related to data creation, the number of repositories has also grown, allowing several organizations to make their data available to external communities. Along with the increasing number of emerging platforms, new interoperability challenges arose.

2.3 OAI-PMH Protocol

As an attempt to fix the so called interoperability problems, OAI-PMH[16] (Open Archives Initiative Protocol for Metadata Harvesting) was designed. It is based on metadata harvesting from communities that might be interested in publishing their contents on the Web, providing a framework that does not impose high barriers to information discovery [Dek07].

When it comes to its implementation, two main entities can be identified:

- **Service Provider** - gathers metadata from *Data Provider*. Upon request to each repository, metadata is sent to databases. Data will then be available, together with search and navigation capabilities;

- **Data Provider** - repository exposing its metadata through OAI-PMH.

Regarding the communication between these two actors, OAI-PMH holds six requests—*verbs*—that might be used by *Service Provider*, as described in Table 2.4 [Dek07].

[16] https://www.openarchives.org/OAI/openarchivesprotocol.html

Table 2.4: OAI-PMH verbs

Verb	**Description**
Identify	Returns information about the repository
ListMetadataFormats	Returns metadata formats available used by the repository
ListSets	Returns metadata sets
GetRecord	Returns a specific metadata record
ListIdentifiers	Returns an item's unique identifiers
ListRecords	Lists all exposed records

All these *verbs* are called via HTTP. *Data Providers* answer using the XML format, and they are required to support DCMES. Please note that some concepts must be distinguished, namely:

- **Harvester** - the client that will issue the harvest request demanded by the *Service Provider*;
- **Resource** - digital object whose metadata is exposed by the repository;
- **Item** - metadata's entry point for a given resource, uniquely identified;
- **Record** - it contains metadata and possible secondary data about it. In order to identify a particular record, there is a matching between the item's identifier, metadata format prefix and corresponding *datestamp* (date and time of metadata's creation or modification).

Figure 2.8 shows an example of the above concepts.

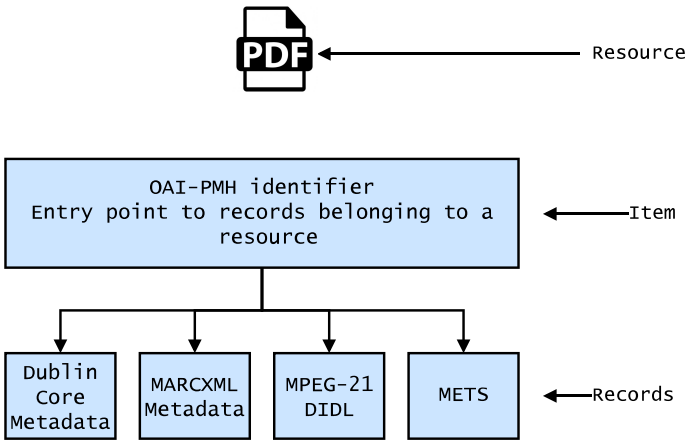

Figure 2.8: A PDF file within the OAI-PMH. Adapted from [VNLW04]

2.4 Repositories for Data Management

With the long tail of science, some solutions started to show up as an attempt to manage data from several communities, as well as from several domains, while providing efficient ways for data sharing and collaboration. Using these platforms can be helpful for the long-term preservation of digital information, as they provide tools for efficient management of data and usually have high visibility. However, there are different types of implementations that fit different needs, so it is fundamental that researchers and communities know the differences in order to make the choice that suits their datasets. These repositories vary greatly when considering aspects such as their architecture or the way they handle metadata records, and several studies show that their adopters often find it difficult to make a choice in the first place [Fay10, ACRdSR16].

2.4.1 Repositories as Installation Packages

These repositories are solutions that come as an installation package that must be deployed within local infrastructures. Afterwards, one can extend its features by installing plugins that fit a particular community or domain. With this approach, researchers have a highly customized solution and more control over their data. In addition, some repositories are open-source, which allows other people to develop more plugins that can adapt to particular needs. However, some of these are aimed at research publications in particular. They can be adapted to data management, but always with limitations.

- **DSpace**[17] (also served as a service by some companies) - a free, fully customizable solution to fit an organization's needs, with multi-domain support. Although Dublin Core is the default standard for metadata within DSpace, it is possible to add and change fields and use any non hierarchical schema. In order to use hierarchical ones like MODS or MARC, some technical skills are required. OAI-PMH is also supported, and it is possible to choose PostgreSQL[18] or Oracle[19] as the metadata database manager. It also benefits of high popularity and acceptance, as it has a large community[20]. DSpace's website provides a full list of repositories using it as a solution, which includes, among others:

    1. **Aichi University of Education, Japan**[21] - digital resources, digital works and papers done by this University's community;
    2. **European Commission, JRC Publications Repository**[22] - data produced by European Commission's Joint Research Centre on several domains, from agriculture to nuclear safety;

[17] http://www.dspace.org/
[18] http://www.postgresql.org/
[19] http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index-097090.html
[20] http://registry.duraspace.org/registry/dspace
[21] http://www.aichi-edu.ac.jp/english/
[22] https://ec.europa.eu/jrc/

- **ePrints**[23] - this package is focused on creating OAI-PMH compliant repositories. Although it is customizable, metadata fields are limited in a fixed way, which means a difficulty for domain data. Moreover, it is much confined to publications;

- **CKAN** - a full open-source suit. It has a modular architecture, allowing the development of extensions by users with appropriate knowledge. Besides default metadata provided by CKAN, it is possible to store arbitrary key-value pairs, which may not follow any predefined schema. There are some well-known case studies, showcased by CKAN itself, that prove this solution's acceptance:

  1. **IATI Registry**[24] - the International Aid Transparency Initiative (IATI) is a main repository for aid spending data published by funding organizations;
  2. **Helsinki Data Portal**[25] - regional information from Helsinki that may be used by everybody, from citizens to researchers;
  3. **Data.Gov.UK**[26] - UK's Government official open data portal. It provides easy data about governmental data for every user.

  It is possible to build an OAI-PMH compliant CKAN repository using already developed extensions.

Although these solutions offer high customization, not all communities or institutions have resources to implement and maintain them. Sometimes, the best solution is to search for an already configured remote platform that can be managed by third parties. By doing so, researchers only have to worry about more important aspects like their own data and leave all other background issues to third parties, even if a contract is required.

2.4.2 Repositories Served as a Service

This kind of repositories are served as SaaS (Software as a Service). Users can interact with them directly using a Web browser or a Desktop application. By doing so, they are free of some responsibilities like maintenance.

- **FigShare** - a generalist repository. It allows data submission along with a manuscript description, as well as in-browser preview features that avoid downloading files like images, spreadsheets, texts and zips. For each uploaded dataset, a DOI is assigned, making it easier to link and cite it. Although datasets are made public, the ownership is not dropped and the creator must be given credit in the form of citation. As for metadata, general descriptors like *title* and *authors* are available;

[23] http://www.eprints.org/uk/
[24] http://www.iatiregistry.org/
[25] http://www.hri.fi/en/
[26] https://data.gov.uk/

- **Zenodo** - created by OpenAIRE[27] and CERN, it also provides a DOI to every publicly available dataset. Furthermore, it is OAI-PMH compliant. It accepts any file format from all domains and is currently developing an auto-metadata feature, which means gathering general information from uploaded files like the title, description and authors. As for now, all metadata is limited to general descriptors stored as MARC records and can be exported as MARCXML, Dublin Core and DataCite Metadata Schema[28]. It has been growing in popularity, as it is largely supported by CERN, storing large amounts of data from the Large Hadron Collider;

- **EUDAT** - this pan-European initiative has the ambition of connecting researchers from every domain and help them to manage their research data along all its lifecycle, covering steps like deposit, sharing and long-term preservation. It offers shared services that are spread across Europe and stores data in some of Europe's most powerful supercomputers. EUDAT is divided into 5 services/modules:

  1. **B2Drop** - synchronization and exchange of research data, so researchers can keep their data synchronized and up-to-date and exchange it with others, like team members. This is somehow similar to a Dropbox[29];
  2. **B2Share** - storing and sharing research data. Built on Invenio platform, it allows researchers to store, preserve and share data. It assigns an identifier to data and allows the insertion of some domain metadata, like biological descriptors, which is already a good improvement on general descriptions, and it is OAI-PMH compliant. Moreover, it provides a comprehensive REST API for integration with communities. This particular service is optimized for researchers who cannot long-term preserve both data and metadata locally;
  3. **B2Safe** - safe replication of research data. This service allows the implementation of data management policies, preventing issues like data loss. Some of its features are the definition of policy rules, data replication, managing identifiers and integrity checking;
  4. **B2Stage** - brings data to computation. This module is responsible for transferring research data between EUDAT's storage resources and high-performance computing (HPC) workspaces. In collaboration with other entities like the European Grid Infrastructure (EGI) and the Partnership for Advanced Computing in Europe (PRACE), it allows researchers to extract computational results out of their data;
  5. **B2Find** - a discovery service based on harvested metadata. It allow users to easily find datasets. It can be used to find meaningful data for reuse and get overviews of available data.

When comparing all these solutions, there are a lot of variables that may influence a community depending on its requirements and domain, which include [ACRdSR16]:

[27] https://www.openaire.eu/
[28] https://schema.datacite.org/
[29] https://www.dropbox.com/

- **Deployment** - installation package *vs* SaaS;
- **Storage location** - remote *vs* local;
- **Costs** - free *vs* paid;
- **Schema flexibility** - fixed descriptors *vs* flexible possibilities;
- **Compliance with standards** - like OAI-PMH and exporting metadata schemas.

As seen before, all those platforms shared a common point: they all focus on the final, refined data at the end of the workflow. There is clearly a gap before this stage, as researchers need to document their data as they produce it.

2.4.3 Early Stage Platform

By gathering metadata in the early stages, the chances of a dataset being preserved and made available to the community can be improved. Another benefit is a better collaboration, as documented data is easier to share and to be understood by others. Bringing the description closer to researchers means improvement on metadata quality, as they are domain experts and fully understand the data they produce. Overall, a flexible, intuitive and easy to use tool—researchers are not necessarily IT experts—can be a valuable contribution for the lifecycle of datasets.

Frequently, researchers already describe data without noticing. From paper notes to virtual annotations in their personal computers, their descriptions is extremely valuable. Letting these descriptions unnoticed means a reduced probability of long-term survival, which is not desirable.

Solutions like Dropbox, Microsoft OneDrive[30], Google Drive[31] or Copy[32] may seem like a solution for making collaboration possible. Nevertheless, simple annotations (if possible) about data can't be considered proper metadata, and managing datasets using these platforms is not the best method.

Dendro platform focuses precisely on this. Using an ontology-based data model instead of a relational one, as shown in Section 1.2, it assists researchers on the daily routine of data management, by letting them choosing terms from several ontologies, from general to specific ones. It also provides recommendations that they can optionally use for richer descriptions and a friendly Web interface, which fulfills one of the requirements mentioned before. It is also collaborative, and the underlying linked data graph is built as researchers use the platform and describe data. Along with data and metadata, ontologies can also be shared, so the *meaning* of data is present all the way. Comparatively to a relational database, the fully ontology-based data model used by Dendro provides *meaning* to data, allowing a richer description and contributing to the preservation of data. This solution brings data description to researchers' everyday data production, by providing a complete tool that is simple for users with no data management skills [RRC14]. By

[30] https://onedrive.live.com/about
[31] https://drive.google.com
[32] https://www.copy.com

allowing researches to use their own ontologies or standard ones, Dendro is proven to be a flexible platform. The collaborative feature is also a plus, since this way users can skip using third-party solutions like Dropbox and be focused on one single platform together with their partners.

2.4.4 Indexers

More and more research groups make their data and metadata available publicly through repositories or even their own Web pages, as shown by Figure 2.9. As of the 2nd of February 2016, the Registry of Open Access Repositories (ROAR)[33] had 4171 registered repositories, although most of them have institutional purposes. Therefore, it is useful to have tools that can gather metadata from all sources and make it available for people that may be interested. Here, we will refer to those tools as *indexers*. Either by using OAI-PMH to harvest metadata from repositories, like OAIster[34], or just crawling the Web, like Google Scholar[35], indexers may be aimed at particular domains or have general purposes.

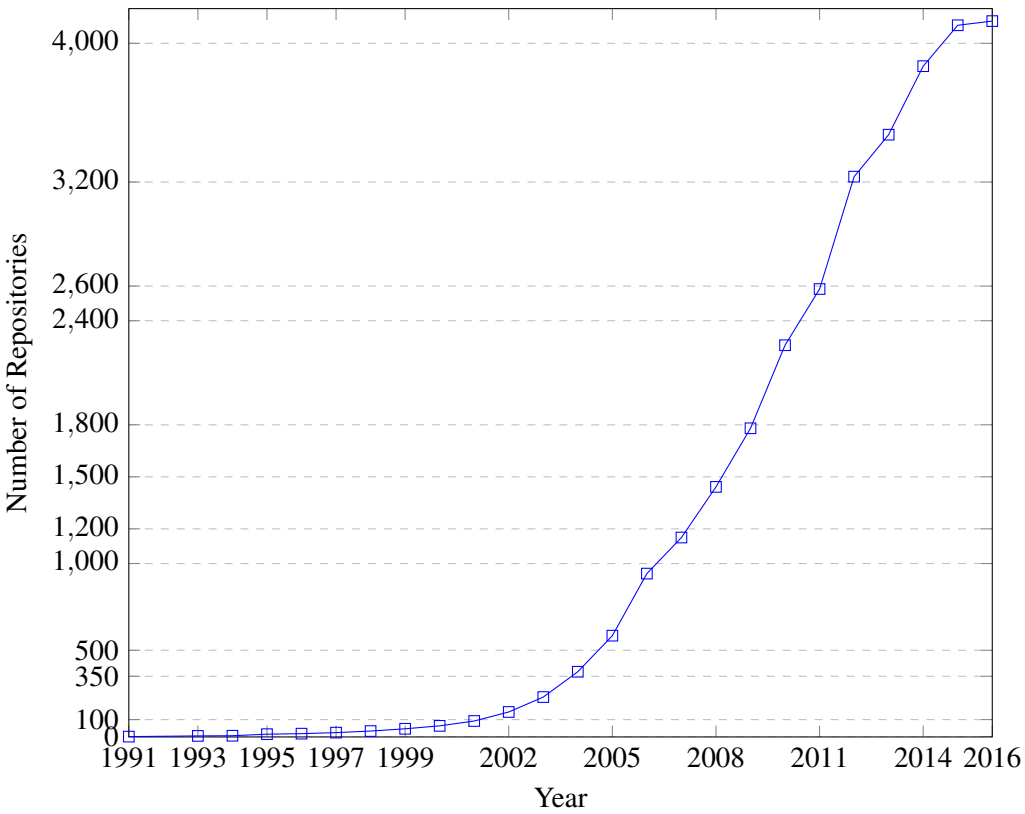

Figure 2.9: Number of Registered Repositories, according to ROAR

The success of these portals, as well as the importance of making data available for open access instead of imposing subscriptions, has already been shown by other studies [NOR08a, NOR08b, HB04]. They play an important role, as they provide a centralized access point from where users

[33] http://roar.eprints.org/
[34] http://www.oclc.org/oaister.en.html
[35] https://scholar.google.pt/

can search for all kinds of data for their own research. As so, some of these institutional portals can be portrayed:

- **OpenDOAR**[36] - a repository list for open access academic repositories. Each added repository is individually verified by the staff, which is a plus. It allows a refined search to match several criteria factors, from subject to content type;

- **Google Scholar** - a search tool maintained by Google. It is multidisciplinary, even indexing websites with scholarly materials. Researchers can upload their papers to their own websites and use a link referring it from their publications page. Google will then automatically index them—as long as websites host mostly scholarly articles and corresponding abstracts are easily shown and available for free. When it comes to an institutional repository, Google recommends using ePrints, DSpace or Digital Commons[37];

- **Coremine**[38] - a specific search tool for the biomedical area. It is very particular, as it uses text mining, allowing users to narrow down their search, as well as the possibility of navigating through the graph of related concepts, using then PubMed[39] to find articles regarding the search. Figure 2.10 shows the retrieved graph for a search about *influenza* virus, which returned 132 articles. By clicking on the edge that goes to *Inluenza in Birds* node, it is possible to filter articles regarding that particular subject, 16 is this case.

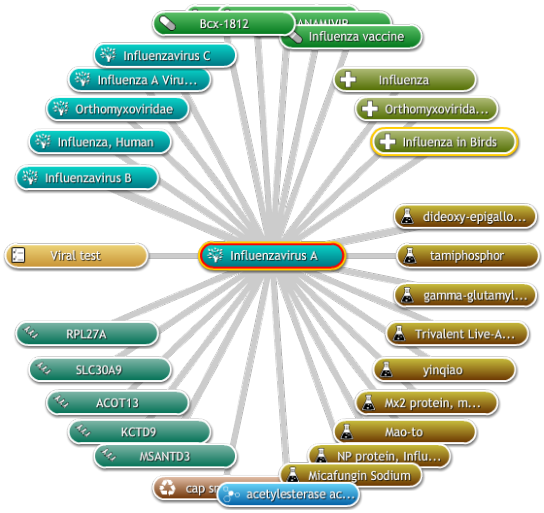

Figure 2.10: *Influenza*-related terms graph

When talking about indexers centered on data repositories instead, we can highlight re3data[40], which allows researchers to find repositories of different backgrounds. After searching for a particular subject or content, we are presented with a list with all results. By clicking one for more

[36] http://www.opendoar.org/
[37] http://digitalcommons.bepress.com/
[38] http://www.coremine.com/medical/#search
[39] http://www.ncbi.nlm.nih.gov/pubmed
[40] http://www.re3data.org/

details, we can find information like standards used by that particular repository, such as possible identifiers used, the underneath repository software and any possible programming interfaces as OAI-PMH.

Regardless of the portal to be used, the point here is that making data publicly available without restrictions is key for dissemination of data, as shown by the above mentioned studies.

2.5 Data Management Plans

In spite of all the mentioned practices like metadata formats and repositories, there is no absolute guarantee that researchers will follow them during their activities. Even if they are willing to do so, all their plans for data should be documented, so that all their intentions towards collaborative activities are clear from the beginning and it can be assured that their plan fits the research. With that in mind, many funding agencies and research institutes are asking for a Data Management Plan—DMP—as part of grant proposals; this is the case with Horizon 2020[41], the National Science Foundation (NSF)[42] and NASA[43] [NAS14], which can be seen as both an incentive and an obligation. By following this practice, researchers may have some benefits like anticipating problems, avoiding data loss and duplication, ensuring that created data is reliable and planing share policies in an early stage [RR14].

While different agencies may ask for particular points to be covered in the DMP, there are usually some general questions that must be answered, like the following ones, asked to the participants in Horizon2020's Open Research Data Pilot [Com16]:

- How will research data be handled during and after the project?
- What data will be collected, processed or generated?
- What methods and standards will be applied?
- Will data be shared or made open access? How?
- How will data be curated and preserved?

The first version of the DMP should be delivered within the first six months of the project for these participants. In order to cover the above general guidelines, Horizon2020 states that researchers must address particular points on a dataset by dataset basis, namely the dataset's name, description, standards and medatada, sharing policies and archiving and preservation procedures. If any change in the project occurs, this document must be updated and submitted again.

Another model for a data management plan is the one issued by NSF, which is largely adopted as a baseline. This agency started requesting a DMP for all grants since 2010, which even led to training and consultation from entities. One example was the University of Minnesota Libraries,

[41] https://ec.europa.eu/programmes/horizon2020/
[42] http://www.nsf.gov/
[43] https://www.nasa.gov/

which trained its researchers and created an institutional data repository [BJ15]. This means some entities have full interest in complying with open-data practices, giving their researchers the conditions they need to achieve that. The NSF model also specifies some points that may be included in the DMP[44]:

- Data produced during the project and its types;
- Standards used for data and metadata;
- Sharing and access policies;
- Policies for data reuse and re-distribution;
- Preservation actions to take place.

Depending on the field of research, particular guidelines must be followed, namely in Biological Sciences, Computer and Information Sciences, Education and Human Resources, Engineering, Geosciences, Mathematical and Physical Sciences and Social, Behavioral and Economic Sciences.

A more precise DMP is the one created by ICPSR (Inter-university Consortium for Political and Social Research), which is more objective, as it states well-defined points to be covered and an explanation for each one. These points go from metadata and security to audience, budget, data organization and legal requirements[45]. ICPSR even compares its topics with NSF's, making the mapping between the two. However, some of them have no equivalent.

Independently of the adopted model, a DMP can be seen as a tool to lead researchers to plan their project and think ahead of time about open-data practices and data sharing. Some studies have found that almost every DMP collected mentions data sharing and sharing policies to ensure long-term preservation and dissemination [BJ15]. This is a conclusive proof that a DMP can have an important contribution for the e-Science environment and the long-tail of data, and making it a required deliverable for grants can be fundamental.

2.6 Conclusions

Many factors contribute to create preservation conditions for datasets. From choosing metadata schemas and repositories, every factor may determine the success of research data management. Some key factors can then be highlighted when talking about long-term preservation of data:

- Using a staging area like Dendro can be fundamental for more complete descriptions of datasets, by complementing generic descriptors with domain ones. Since researchers have a deep knowledge about the data they produce, it is only normal that they must be given support and tools to do an easy and flexible description of datasets at an early stage;

[44] http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp
[45] https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/elements.html

- By using well-established standards or protocols, like Dublin Core for metadata representation and OAI-PMH for exposing metadata to existing harvesters, research communities are contributing for a better interoperability among platforms, which also means better and easier integrations;
- Indexers are a useful tool when searching for data already available and to make produced data available for others;
- Choosing a final repository is also vital, as several solutions are available. Deploying a local solution provides more control over data, but requires infrastructures. On the other hand, remote ones are easier to deploy, but can impose loss of control over deposited datasets. Some of the mains concerns regarding repositories are the location of data (local or remote), compliance with standards and open or close systems. Usually, open-source solutions like CKAN are better, as they provide tools for extending their features, which means better and more complete platforms configured for users and their specific requirements.

By following the right guidelines and choosing accordingly, researchers will be also helping others in their research, thus creating a rich collaborative environment where data makes sense for whoever uses it, no matter when or by which means.

Chapter 3

EUDAT Link, a Deposit Workflow for Dendro

Research data management can comprise several stages where data goes through and different stakeholders intervene. Assuming researchers are given a staging area for early description of data as part of the solution for the long-term preservation of datasets, they will provide domain descriptions that are essential for the data retrieval later on. Some of them already do their descriptions in less reliable ways, so they can be supported in order to facilitate their description job and afterwards.

After that initial preparation , all data and metadata should continue towards a final repository, where they will be deposited, preserved and shared. At this stage the dataset already includes an extensive metadata record, so new challenges emerge when trying to maintain the link between both these platforms, such as compliance with metadata standards and protocols like OAI-PMH.

3.1 The Workflow

Setting the tools to use along the course is as important as data itself. For the most complete and efficient workflow, some main components can be highlighted:

- A staging area for daily management of data and related activities, including deposit and both generic and domain descriptions;

- Tools to facilitate the integration of methods that researchers already use with the staging platform. These can either be electronic laboratory notebooks or plugins to streamline data capture;

- A final repository where data and metadata will persist and be shared with the community.

3.1.1 Dendro, the Staging Area

The description task is already made by researchers and they have an in-depth knowledge of data and its domain. Besides that, it is not efficient to let curators to do that job—again, they have no sufficient knowledge on the domain—, so Dendro emerges as a tool for improving research data at the beginning. By also offering storage and collaboration, Dendro is a good approach for this step. The underlying technology is also a point in favor. By using ontologies as a way to define the domain, datasets become integrated in a unique way that gives sense to them, using standards-compliant metadata most of the times.

Although other solutions could be considered, Dendro covers additional aspects, namely:

- Dendro is a local solution, with ongoing development, getting direct feedback from researchers;
- Platforms like Google Drive, OneDrive or EUDAT's B2Drop provide data exchange and synchronization. However, raw data is not properly described most of the times, so this would only work as a mere storage area. These solutions fall short when dealing with extensive, variable metadata records and are often limited to a basic set of descriptors, or none at all.

For these reasons, it was decided to use Dendro during the initial day-to-day phase, then sending all datasets for an international repository that will be responsible for long-term preservation and dissemination.

Along with Dendro, researchers may use other tools that can ease the initial description of data, such as electronic notebooks, overcoming less practical and standard methods like paper notes.

3.1.2 LabTablet, the Laboratory Notebook

This electronic laboratory notebook is designed to describe research data using standard metadata records [ACRR14]. Together with Dendro, LabTablet[1] helps researchers to manage their daily data by also gathering some metadata from the device's sensors, like geographic coordinates. Data can then be sent to a long-term repository like CKAN or synchronized with Dendro for further management and description. This kind of approach is suitable for field trips, as researchers can take their notes using the application, thus replacing obsolete techniques used before. The application also addresses scenarios where data can also be gathered, helping researchers to capture data at any time, later exporting them to the staging platform, along with the metadata record. From there, they will both be deposited into a final repository.

3.1.3 EUDAT, Modular Solution at International Level

When choosing the final repository, several advantages and disadvantages must be taken into account. EUDAT is a modular end-to-end solution and served as a service, which means reduced

[1] http://dendro.fe.up.pt/

dependency on other infrastructures and maintenance tasks. It is compliant with metadata standards like Dublin Core, which is essential, and part of its modules is free to use. As a plus, some of its modules—like B2Share—are built on top of other well known and vastly used platforms like CKAN and Invenio.

Recently, EUDAT has challenged European research communities to test their services, a program known as *Data Pilots*. These pilots aim at improving the existing platform and address specific researchers' needs that can only be identified with close partnerships with them. This call means proper support from EUDAT, as well as a big opportunity for participants being pioneers and getting further visibility. As part of Horizon2020 program, it brings a lot of attention and support from European communities, which may contribute to its development as a service for research data management.

Given these factors, EUDAT will be used as the final platform for all data and metadata coming from Dendro, reaching the final step of the RDM workflow. Specifically, B2Share is the module to be used as the final repository. Although EUDAT could be a full solution for the entire workflow, the initial phase would have some drawbacks. B2Drop, which can be seen as analogous to Dendro, does not allow for complete descriptions of data—like mentioned before, it is more closely related to Dropbox, in the sense that it provides data storage and general descriptions. It becomes natural that Dendro, as a local solution, can be used initially instead.

When analyzing the remaining EUDAT's modules, another one caught our attention. B2Find, as an OAI-PMH harvester, can be used to gather metadata from Dendro on a regular basis, as researchers may want to share their metadata from the beginning and get early visibility. This way, the goal of the pilot program can be reached by two different ways. Moreover, metadata can be harvested by other entities than just B2Find, so it is a good opportunity to develop this feature.

3.1.3.1 B2Share for Long-Term Preservation and Dissemination

This specific module is responsible for storing research data and guaranteeing its persistence over time. More importantly, it allows data to be later retrieved by other researchers, which is the main point when looking for a solution for data preservation. It lets any user to deposit data without any costs, then assigning a unique identifier to each dataset. Moreover, all provided metadata used to describe datasets are publicly exposed through OAI-PMH and ready to be harvested. Finally, it has a REST API that allows any community to integrate its services with a local solution.

The key point is then to make use of this API to integrate B2Share as part of Dendro's environment, making this repository another solution for the final stage, along with others—CKAN, DSpace, EPrints, Figshare and Zenodo. The export process starts upon user request, and all project's data should be sent, as well as all possible metadata that can be reused to describe the dataset within B2Share.

3.1.3.2 B2Find, the Metadata Harvester

On some occasions, some researchers may look forward to disclose their data from the begining of their project. Although rare, this scenario can enhance the project's visibility and promote a collaborative environment that may be interesting to institutions.

Dissemination protocols such as OAI-PMH are designed for this purpose and can be used to publicize datasets to the scientific community, provided there is a link between the instituion's repository and the search engine. In this context, B2Find can be used to harvest metadata and expose data stored in the Dendro platform.

The workflow must address different scenarios depending on the researcher's intentions on disclosing data. Upon the project creation, the researcher can define whether the project's metadata can be harvested by the B2Find module. Projects can then be systematically indexed by external harvesters, if specified by the researcher.

Our workflow is therefore defined as represented in Figure 3.1, covering all the research data management steps.

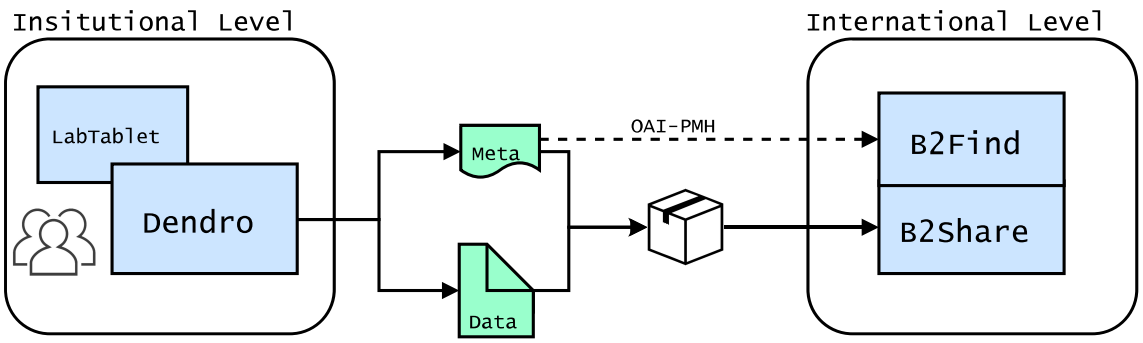

Figure 3.1: Complete Worfklow

3.2 Conclusions

Research data management requires a complete and streamlined data management workflow. By giving researchers a platform like Dendro for their everyday data management, together with LabTablet for a flexible set of tools for data description, the remaining of the workflow requires a data repository, and EUDAT's B2Share has been chosen here. This choice was made after considerations like visibility, support, innovation, compliance with standards and openness.

In the eventuality of researchers wanting to disclose their metadata sooner, they can define their projects with a privacy level that allows metadata to be exposed through OAI-PMH. This way, projects can get early visibility even before depositing data.

All the architecture takes into account the best possible description for data, by allowing domain terms since the beginning and respecting predefined standards. By taking all stages of data

into account in order to contribute to a better research environment, this solution can be considered complete and built according to well-established standards and protocols.

Chapter 4

Implementing EUDAT Link

The integration between Dendro and EUDAT was made using two different approaches. First, setting up an OAI-PMH server that gathers metadata from Dendro and makes it available for B2Find (and for any other harvester), always taking privacy concerns into account. Then, exporting a package with both data and metadata from a project following a decision by the researcher, using Dendro's interface. This package is sent to B2Share using its API.

4.1 B2Find approach

Before exposing any kind of metadata from projects, privacy concerns must be taken into account. Some projects have privacy constraints that don't allow researchers to disclose their data, so they must define a privacy level for their project by the time it is created. To accommodate this, and knowing in advance that Dendro uses an ontology-based data model—again, ontologies are representations of a particular domain, saved as RDF triples—, the Dendro ontology, *ddr*, was expanded with the *hasPrivacyStatus* predicate to address this requirement, as shown in Figure 4.1:

- *Public* - project's metadata is exposed and the project can be seen by any person;
- *Private* - project's metadata is not exposed and no one but its creators and collaborators can see it;
- *Metadata-Only* - project's metadata is exposed, but users must ask for access in order to see it.

Using this extension to the ontology, researchers are fully aware of the their projects' privacy, and they completely control it during the entire workflow. If they want to change the privacy level after the creation of the project, that is possible through the project's administration page.

OAI-PMH requires metadata to be provided as XML files, which means Dendro has to provide them in the first place. Dendro uses a triple store to save all its metadata, so there were no readily-available XML files to provide. In order to achieve this, a NodeJS[1] script was developed. This

[1] https://nodejs.org/en/

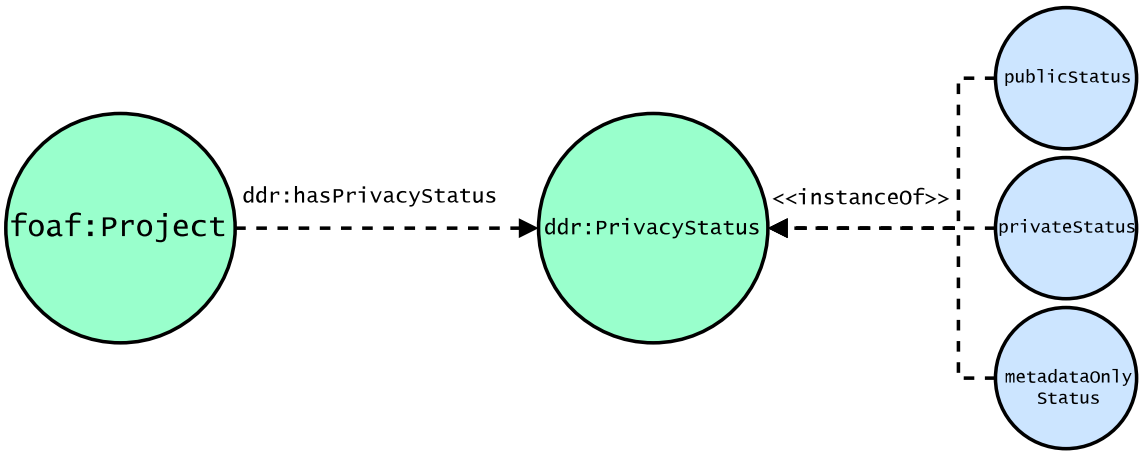

Figure 4.1: Privacy definitions within Dendro's Ontology

script connects to the triple store's SPARQL endpoint and selects all projects that are defined with public or metadata-only privacy status. For each one of those, all Dublin Core metadata is collected and an OAI-PMH compliant XML file, as shown in Listing 4.1, is built and saved in a proper folder. Of course this folder must be constantly updated with fresh metadata from Dendro, so this script is scheduled to run every six hours.

```
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"xmlns:dc="
    http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http:
    //www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:modified>2016-05-23T09:24:31.554Z</dc:modified>
<dc:description>Data for the vehicle-reservation assignment problem in a car rental
     company</dc:description>
<dc:publisher>http://dendro.fe.up.pt</dc:publisher>
<dc:source>http://dendro-prd.fe.up.pt/project/carrental</dc:source>
<dc:title>Vehicle reservation assignment in car rental</dc:title>
<dc:creator>http://dendro-prd.fe.up.pt/user/beatriz</dc:creator>
<dc:language>EN</dc:language>
</oai_dc:dc>
```

Listing 4.1: OAI-PMH XML file

Setting up a server that follows the OAI-PMH protocol is a common and important task among harvesters and providers that want to gather or expose metadata. For that reason, the Digital Library for Earth System Education (DLESE) provides jOAI[2], a tool that allows to configure an OAI-PHM servlet within Apache Tomcat[3] in an easy way, exposing metadata through the URL *«machine-url»:«tomcat-port»/oai/provider* . After the installation, some configuration must be

[2] http://www.dlese.org/dds/services/joai_software.jsp
[3] http://tomcat.apache.org/

done in order to make the server fully functional. All this setup is handled by jOAI through its configuration webpage depicted in Figure 4.2, and the first step is to define the repository information, namely:

- Repository name;
- Repository administrator's email;
- Repository namespace identifier.

The first two are trivial and easy to understand, but the last one requires some more attention. The repository identifier takes the form of a Web address or domain name and it is used to uniquely identify the repository and to be part of every record's identifier later on. This record identifier has the format *oai:«namespace-identifier»:«file-identifier»*, where *«file-identifier»* is the unique name of the XML file itself, which comes after the project's unique handle used in Dendro. These identifiers follow the *Specification and XML Schema for the OAI Identifier Format*[4], and Dendro's identifier was set to *dendro-prd.fe.up.pt*. Altogether, every record's identifier takes the form *oai:dendro-prd.fe.up.pt:project/«project-handle»*. In order for any person that retrieves this metadata to reach the project's page, the *dc:source* element is set with that project's URL, which is almost identical to the record's identifier itself, as seen in the example.

Figure 4.2: jOAI—provider configuration page

The second step is to fill out information about metadata folders that provide the XML files in a per folder basis, specifically:

- Nickname;

[4] http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm

- Path to the directory;
- Format of files.

Given that this configuration is made for each folder, one can conclude that every XML file within a folder must follow the same format, as shown by the workflow in Figure 4.3. Since our script saves Dublin Core files in a folder, only that one was added here, with the format field being set as *oai_dc*, which is the *metadataPrefix*[5] used for Dublin Core. After this step, jOAI will index all files in that moment and for every eight hours, exposing them via OAI-PMH. Using the *ListRecords* verb, by typing *http://dendro-prd.fe.up.pt/oai/provider?verb=ListRecords&metadataPrefix=oai_dc* in the browser, the response shown in Listing 4.2 is returned. The identifier of the only record indexed is *oai:dendro-prd.fe.up.pt:project/carrental*, which means that metadata comes from a Dendro project with the handle *carrental*.

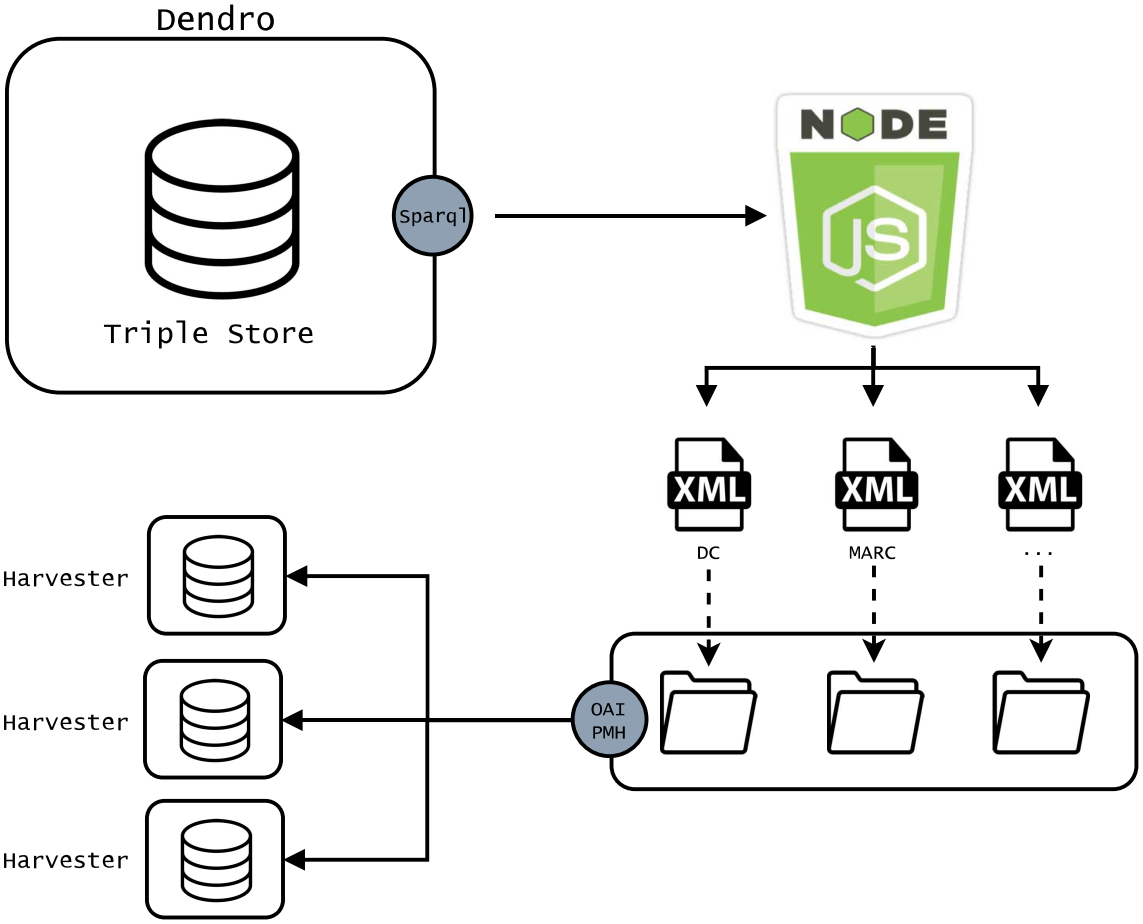

Figure 4.3: Complete workflow for the B2Find approach

[5] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm#MetadataNamespaces

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org
    /2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI
    /2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2016-05-30T18:39:43Z</responseDate>
<request verb="ListRecords" metadataPrefix="oai_dc">http://dendro-prd.fe.up.pt/oai/
    provider</request>
<ListRecords>
<record>
<header>
<identifier>oai:dendro-prd.fe.up.pt:project/carrental</identifier>
<datestamp>2016-05-30T16:51:13Z</datestamp>
</header>
<metadata>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"xmlns:dc="
    http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http:
    //www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:modified>2016-05-23T09:24:31.554Z</dc:modified>
<dc:description>Data for the vehicle-reservation assignment problem in a car rental
     company</dc:description>
<dc:publisher>http://dendro.fe.up.pt</dc:publisher>
<dc:source>http://dendro-prd.fe.up.pt/project/carrental</dc:source>
<dc:title>Vehicle reservation assignment in car rental</dc:title>
<dc:creator>http://dendro-prd.fe.up.pt/user/beatriz</dc:creator>
<dc:language>EN</dc:language>
</oai_dc:dc>
</metadata>
</record>
</ListRecords>
</OAI-PMH>
```

Listing 4.2: Result of the *ListRecords* verb

At this point, everything is set and B2Find—or any other harverster—can now collect Dendro's metadata, making it then available for the community through its search interface.

As mentioned before, different privacy levels lead to different views of the project. In this case, metadata harvested by B2Find leads to either public or metadata-only projects, which means that researchers may have to request access to a project, depending on the privacy status the project's creator defined at the beginning—see Figure 4.4. Figure 4.5 shows the view of a public project, where the project structure is totally visible and there's the possibility of downloading it. Now if the project is a metadata-only one, user has to request access for it, as seen in Figure 4.6. After clicking the request button, the project creator will receive an email stating that this user is requesting for access to the project. Using the project administration page, the creator can then add the user as a collaborator. After that point, the user will also get an email confirming they have been added as collaborators, and can now see and edit the project.

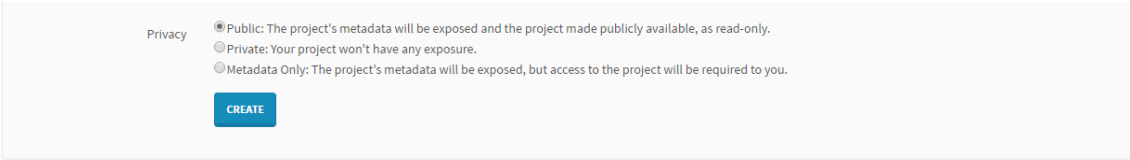

Figure 4.4: Project creation page—privacy section

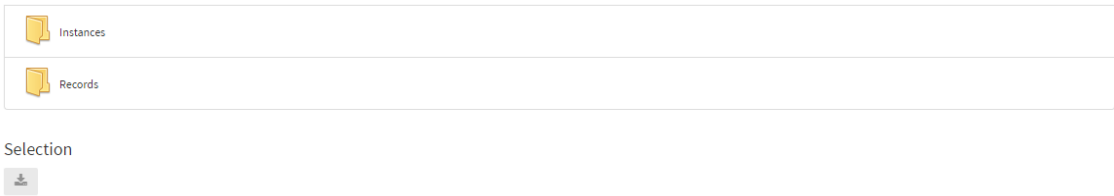

Figure 4.5: View of a public project

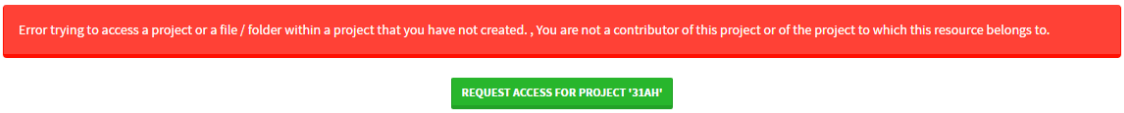

Figure 4.6: View of a metadata-only project

4.2 B2Share, the External Repository for Data and Metadata

When researchers feel their data is described accurately with domain metadata, they can finally export their project to B2Share. If researchers wanted to deposit their data directly through B2Share's webpage, they would have to create an account, download all their project from Dendro, upload it to B2Share and describe it all over again. However, that requires a lot of effort from them, which is not what we want and it is against the homogeneous process we are looking for. Therefore, our integration module connects the two ends and allows researchers to export their data directly via Dendro.

B2Share provides a REST API[6] that may be used for the export. Using this API is the right way to take advantage of the information prepared by the researchers and turn all the process efficient and intuitive. Figure 4.7 illustrates how this API works, which can be divided into three steps:

1. A deposit object is created within B2Share and the ID is returned, so it can be used for later requests;

2. Files can be sent to the deposit object using its ID;

[6] https://b2share.eudat.eu/docs/b2share-rest-api

3. Metadata is sent and the deposit object is committed, becoming an immutable record, which means it can no longer be modified. The record's URL is returned.

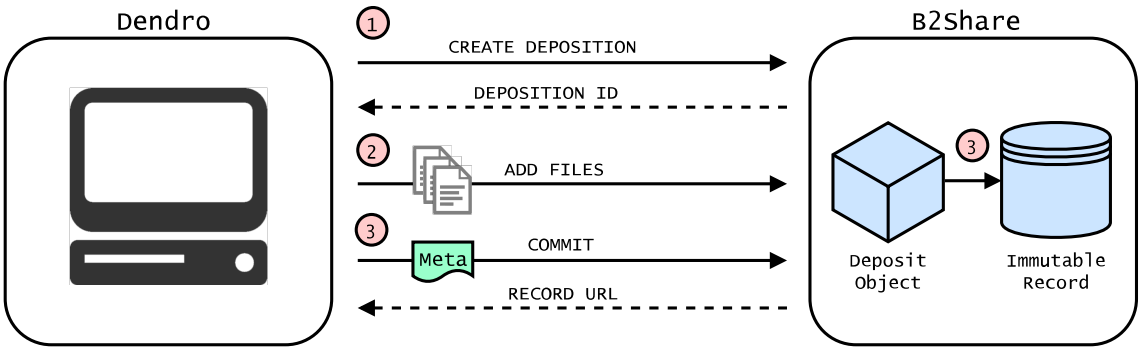

Figure 4.7: Necessary API calls to make a deposit into B2Share

All requests to the API require a token generated by B2Share. A B2Share account for Dendro was created and the corresponding token is used for the communication. All records are deposited into Dendro's account, unless researchers have a B2Share account of their own. In that case, they can use their token by the time of the deposit, and all data will go into their account. When exporting, researchers must first add a bookmark with information on the target repository. By doing so, they only need to insert information about a given repository once. They can even have two bookmarks for the same repository, such as one with the default configurations and another with their own token, as in Figures 4.8 and 4.9, respectively.

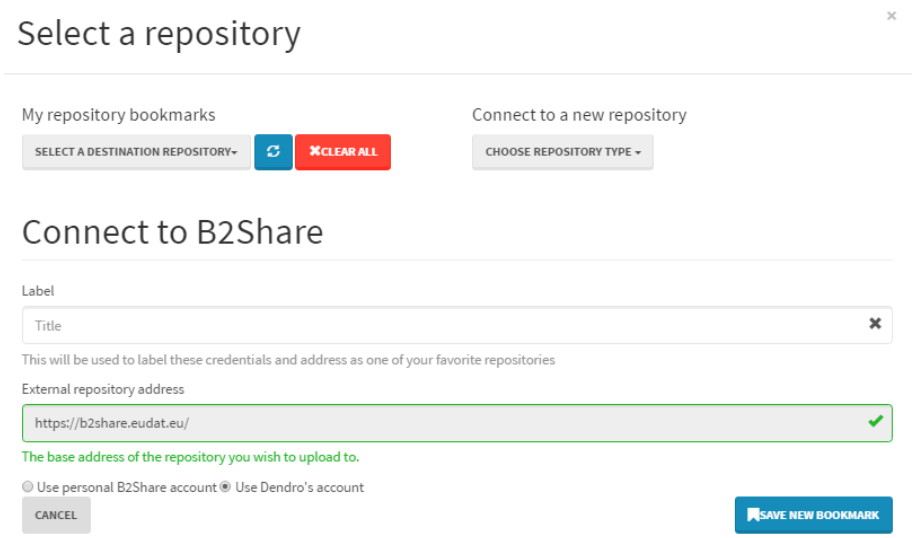

Figure 4.8: Saving a repository bookmark

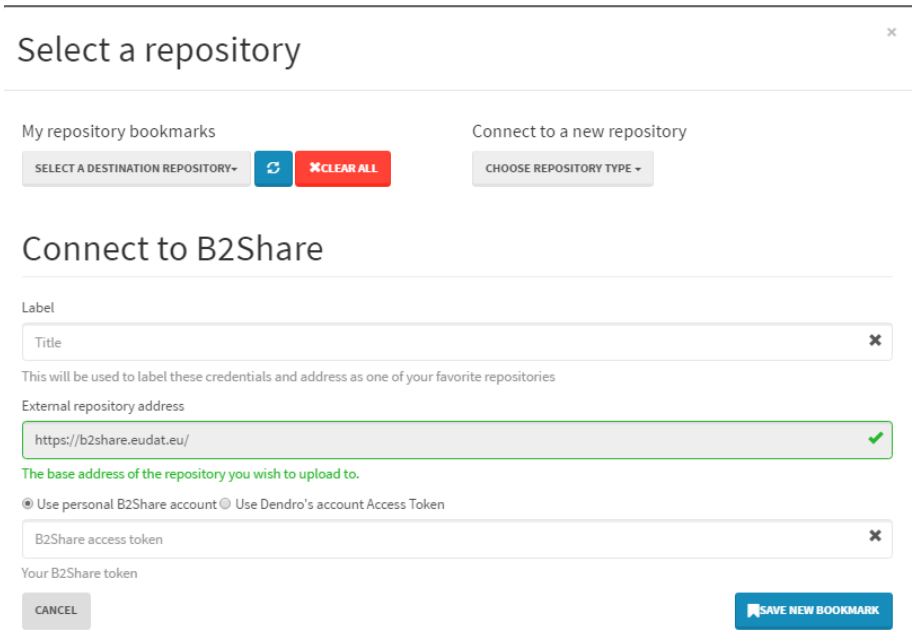

Figure 4.9: Saving a repository bookmark

After creating the bookmark, the researcher can use it to export all the project, by clicking the *My repository bookmarks* dropdown menu, selecting the bookmark that was created and hitting the *Send* button, as shown in Figure 4.10 (1). B2Share only demands four metadata fields to be specified, namely domain, open access policy, title and description. Dendro already takes care of these four: the domain is *generic*—by the time this was written, it was the only suitable domain, as others were for very particular communities—, open access is *true* and title and description were already provided by the time the project was created, as they were mandatory fields. However, only this information by itself is scarce, so a few more descriptors accepted by B2Share are sent, which are:

- Creator - the project's creator URI from Dendro;
- Publisher - the project's publisher, if defined by the researcher. Otherwise, it is set as *http://dendro.fe.up.pt*. If this is not sent, EUDAT will set B2Share as the publisher by default;
- Subject - project's subjects (can be many), if made available by researcher;
- Language - project's language, if made available by researcher;
- Contributors - other possible contributors than the creator himself.

All these fields are sent as part of the request body (POST request), then mapped to Dublin Core terms by B2Share. Still, domain metadata, which is the most important, is not yet part of that description within B2Share, so all domain information from Dendro would be useless. To

compensate this, all metadata from the project (domain and generic) is sent along with data itself, in the form of a plain text file and another with the RDF syntax. This way, any researcher that retrieves the project can then interpret it correctly using domain information that was created by other researchers. Regarding project's data, it is sent as a zip file containing all project's files with the same structure as in Dendro, which means the same folders will be present in that zip with the exactly same contents. By only sending the zip file and two files containing metadata, we're also reducing the amount of requests done to B2Share: one for creating the deposit object, three for adding the files—zip, text and rdf—and a final one to commit the object.

After the package is sent, the record's URL from B2Share is shown to the user and sent via email (2). The record does not become accessible immediately, as it takes some minutes for it to be processed.

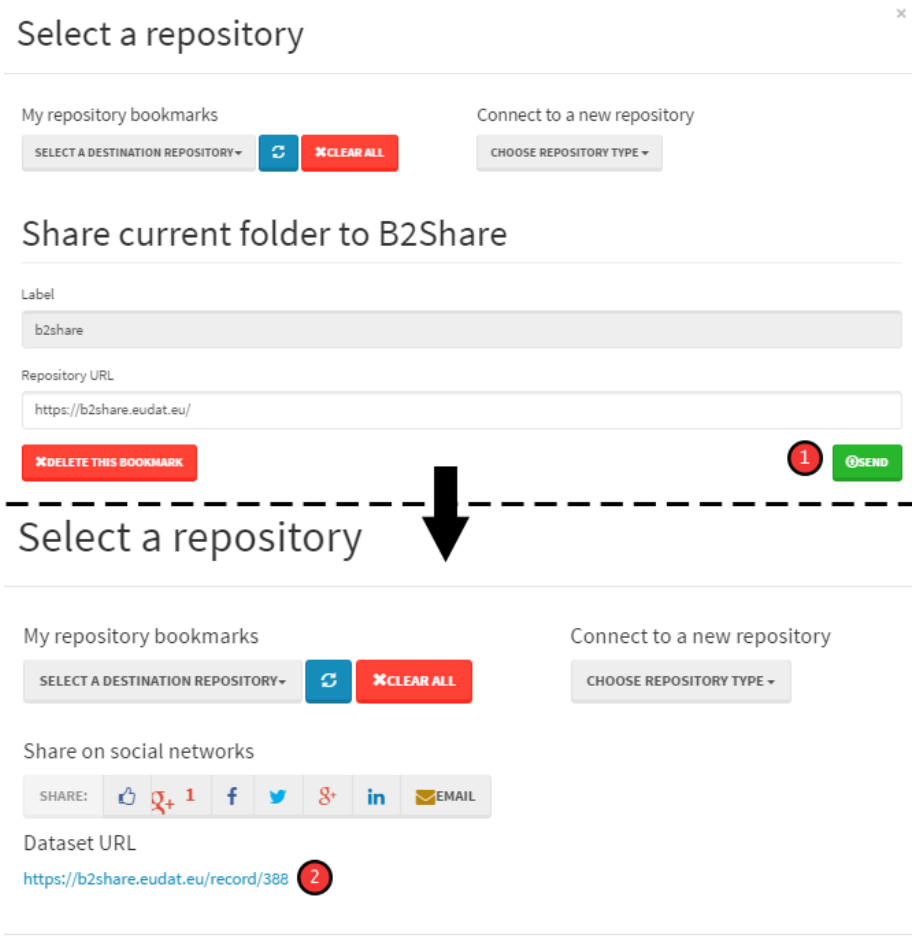

Figure 4.10: A successful exportation of a project to B2Share

Chapter 5

Testing EUDAT Link

In order to evaluate this integration module and its acceptance, the best approach is to test the solution with real researchers that may have actual datasets ready to be described in Dendro and then exported. After some contacts with researchers, only one was actively looking for publishing final data that could reach a final repository—from this point on, we'll call this researcher the Tester.

5.1 Data Context

Our Tester is currently a researcher at INESC TEC (Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência)[1], an institute focused on Research and Development of innovative technological projects. This particular researcher is working on a decision support system for *rent-a-car* companies, allowing them to effectively allocate vehicles to reservations and thus maximizing profits.

As for data outputs, those mostly consist of spreadsheets filled with values related to this particular problem. Those values are identified within each table by their respective variable name, which can be incomprehensible to any third party if no description and short explanation are given. In order to solve this description problem, the Tester produced an auxiliary pdf file containing a short description for each variable in the spreadsheets.

When asked about the methodology used to compare this same results against results from other researchers, the Tester stated to never have found any published values of this nature, making results validation an hard task, which is why the researcher wanted to export this data. Therefore, publishing these results can be seen as a good contribute for other people that may seek validation for related problems and projects. It was the researcher who actively was looking for publishing this data. Thus, the researcher naturally had no idea about Data Management Plans and has no regular practices towards Open Access.

The researcher sends these files via email to another colleague, and the only way to refer these datasets is by citing a scientific paper from where the data is available, which is not a very

[1] https://www.inesctec.pt/

convenient and collaborative way to work. Furthermore, all data is stored in the researcher's personal university webpage. As we can check, a tool like Dendro can be very convenient to the Tester at all levels, from storage to data export.

5.2 Evaluation Tasks Until Export

There was a simple workflow that the Tester should follow towards the final deposit:

Table 5.1: Test tasks before export

Task	Description
Create project	Set appropriate project's title and description (both mandatory), publisher, language, coverage (all optional) and a privacy level
Create project structure	Create the appropriate folders and subfolders
Upload files	Upload all data, according to the previous defined structure
Add adequate descriptors	Set the desired descriptors from the available ontologies
Create B2Share bookmark	Defining a new B2Share bookmark that can later be used to export the project. Researcher can insert a personal B2Share token
Export	Send the project to B2Share using the previously defined bookmark

The initial creation of the project was not an hard task for the Tester. All the fields were filled in with the information about the project, using the paper's abstract as the description. Since there were no privacy constraints, the project was set as public.

As the researcher only had one Excel file with all the tables, only one folder was created for this test. After that, both the Excel file and the pdf with explanations for the variables were uploaded.

Although there's no ontology in Dendro with metadata for this very specific kind of data, there were some keywords that could be used. After choosing the Dublin Core ontology, the Tester picked the *Subject* descriptor—which corresponds to *dc:subject*—three times, setting the keywords *Empty repositions*, *Car rental* and *Assignment*. After these, the researcher found no more adequate descriptors that could be added to the project.

As the project already had a good level of metadata, the Tester was ready to deposit the project, and the next step was to create the B2Share bookmark. The default Dendro's B2Share account was used here, so creating the bookmark was straight forward.

After the creation, the bookmark was chosen and the Tester hit the *Send* button. After a while, they got a message stating all data was sent, as well as the record's URL. An email was also immediately received with this same URL.

5.3 Evaluation Tasks After Export

Now that all data was exported, the researcher can use the record's URL to check everything is fine on EUDAT's side, according to the guidelines we have previously prepared:

Table 5.2: Test tasks after export

Task	Description
Validate B2Share record	Follow the record's URL that was shown (and sent to the email) and verify all the information is the same as Dendro's
Verify handle	Check if the unique handle provided by B2Share is correct and redirects to B2Share record's page
Verify if information was indexed	Using B2Share's interface, do some text search using terms from the project's metadata and check if the record is retrieved

The record didn't become immediately available after the export. After some minutes, the researcher checked back again and the page showed up correctly. Afterwards, the Tester checked the record's information as follows, from B2Share:

Table 5.3: Fields' correctness on B2Share's side

Dendro's information	Correct?
Title	✓
Description	✓
Publisher	✓
Creator	✓
Subjects	×
Language	✓
Project's zipped data	✓
Text file with all metadata	✓
RDF file with all metadata	✓

Table 5.3 shows that the only missing information were the subjects that our Tester added before. This happened because those descriptors were added to the folder and not to the project's root. Dendro only sends information from the project itself and not from the inner folders or contents to B2Share, as it wouldn't be right to send metadata from particular folders or contents to describe the project globally. Being so, this issue could be easily solved by adding those subjects to the root.

The record's page also provided an unique handle that identifies this data, in the form *http://hdl.handle.net/«handle»*. By following this identifier, the researcher reached the record's page again, meaning that the handle is functional and ready to be cited.

Now that the project was properly stored on EUDAT's side, it was time to check if all the metadata sent by Dendro was indexed and being disseminated. It took some hours for the metadata to be searchable, and these were the terms the Tester looked for:

Table 5.4: Fields' correctness on B2Find's side

Term from	**Project retrieved**
Title	✓
Description	✓
Dendro's username (creator)	✓
Publisher (i.e. *Dendro*)	✓

As stated previously, B2Share provides all its metadata through OAI-PMH to B2Find, which means our information should be present there as well. By going to B2Find search interface and doing the exact same searches, we got the same results but one: when using the researcher's name, the project wasn't retrieved. When analyzing the project's information from B2Find, we found that the project's creator is there defined as the email with which we've registered Dendro in B2Share and not as the creator that's defined there. This is probably a bad design issue from EUDAT which will be reported. Nevertheless, it is good that all information that was just in Dendro is now disseminated by EUDAT, and the researcher had almost no trouble and lost a reasonable time to achieve that.

When asked about Dendro and all the tasks involved from creating the project to deposit data, the researcher left a good feedback about the overall process, stating that the Dendro phase could be easily used as part of the daily routine of data management. Moreover, they felt this integration is fundamental, specially when considering that there is no need to re-upload all data to B2Share and do the same descriptions that were previously available in Dendro.

5.3.1 Conclusions

While preforming this evaluation with the researcher, we could check that they had few knowledge on activities related to research data management and all the practices we try to encourage with the use of Dendro and EUDAT Link. All the data was being exchanged by email, and this data had lots of variables with no meaning or description for whoever tried to read them, so it is almost impossible to find other results to compare against. Nevertheless, they had no trouble working with the platform during all phases, which means our solution can be used to achieve the practices we look for.

During the test, a set of tasks was predefined, so the user could be guided during the process. From creating the project and its structure to deposit all of it in B2Share, the researcher had contact with all the important aspects of the solution and its functions. They had the opportunity to check that all the information in B2Share was correct afterwards. Even more important, they got an unique international handle that can be immediately used to cite their data.

The only issue during the test was related to the missing keywords that were not sent. This is due to the level at which metadata is added, and it is normal that only metadata from the project's root is sent to characterize it as a whole.

Although only one researcher was available to publish data at this point, this evaluation can be seen as a proof of concept when it comes to test the Dendro and the integration module as a joint tool to help researchers in their daily activities of data description, collaboration and dissemination. The difficulty to find available researchers wiling to deposit their data shows one of the main barriers to research data management, and it was part of this project to try to raise awareness among researchers for these issues.

Our Tester felt that the solution was very easy and intuitive to use and they are even willing to use it as part of his daily routine for data management, meeting then one of our main goals, which is to promote the Open Access for the Long Tail of science.

Chapter 6

Conclusions

EUDAT Link is an integration module between the Dendro platform—an early stage platform for data description—and EUDAT—an European set of services that covers the research data management workflow. By using metadata that was previously made available by researchers, EUDAT Link saves researchers valuable time, also simplifying the export process. Using two different approaches for this integration added value to the overall implementation. First, using the B2Find module and allowing researchers to define the privacy level for their projects, which can result in an early exposure of metadata and an enhanced visibility for their data. Then, by letting researchers export their project when they feel their data is described with enough and accurate metadata. This project showed us that there is a long way before Open Access and research data management achieve their full potential, so one of its main goals was to sensitize researchers of the existent approaches that can help them in their daily routines.

EUDAT Link's evaluation involved contacting both researchers and students from diverse domains, allowing us to get in touch with their environments and perspectives over their data's visibility. Although it is still common to find that researchers are not aware of the importance of research data management, the cases where data is expected to be deposited and disclosed are becoming frequent. After evaluating EUDAT Link, results showed that our integration module was important to the research data management workflow and our testers. Integrating EUDAT with the Dendro platform allows researchers to directly deposit their data as soon as the project comes to an end. Staging areas and data repositories are an essential part of a research data management workflow and bring great advantage in working together. On one hand, staging areas such as Dendro enable researchers to actively contribute to the quality of their data, encouraging timely data description, structured metadata and compliance with existing standards for both data description and dissemination. On the other hand, repositories such as EUDAT assume increasing importance, as they provide means for ensuring data dissemination and improving their visibility. Although challenging, the adoption of these platforms has been receiving support from funding institutions that consistently require data management plans to be in place when approving funding for specific projects. Nevertheless, specific outcomes such as being able to cite research data or receive credit for their publications are mostly responsible for researchers getting motivated to rethink their data

management workflow.

6.1 EUDAT as a Final Stage

EUDAT and its modules—B2Find and B2Share, in this particular case—are still in a very early stage of development and can't handle domain metadata as Dendro does. Despite the fact that all data is correctly sent, the metadata supported is still at a basic Dublin Core level. This is not enough for all domain metadata supported by Dendro, which is why we had to devise an alternative and send metadata as separate files along the project's data. However, the sent metadata proved to be valuable for the data retrieval.

When evaluating aspects like B2Share's API, which was directly used in this project, some limitations emerged. If researchers want to change some information about a particular project, all data and metadata must be re-sent, which will lead to a new record with a different handle—this is equivalent to have two completely different records. This is a common characteristic in long-term repositories, so we must raise awareness among researchers for this fact.

By indexing metadata and enabling users to look for datasets through its search engine, B2Share is key for dissemination of all projects sent from Dendro. B2Share also exposes all metadata through an OAI-PMH endpoint, making it available for B2Find, which does the harvesting and enables users to do the same searches as in B2Share. Our tests showed that data retrieval was working, which encouraged our researchers to rethink how they handle data and how valuable Dendro and EUDAT Link can be, seeing them as possible daily tools for data management.

EUDAT is financed by the Horizon2020 program and it has some support from big institutions like CERN, which can help it promoting its adoption by other research institutions. Furthermore, organizing a set of pilot programs in order to gather feedback from communities and letting them integrate their services can help the platform to take the right path towards a solution that can reach that level of adoption.

6.2 Future Work and Improvements

In order to keep up with the pace towards better strategies for Open Science and all it involves, some improvements can be made to EUDAT Link. As noticed during the evaluation, metadata from inner folders is not sent to describe the project, so we could change Dendro in order to gather even more general metadata—like subjects—during the project creation.

The Data Pilot associated with this integration, which aims at integrating EUDAT with Dendro while developing both platforms and getting proper support, has a duration of one and a half years, so EUDAT Link will evolve as EUDAT itself, as more requirements are set by our partners during the remaining period. The pilot allowed EUDAT to gather feedback from us; in return, we raised researchers' expectations when it comes to data deposit and research data management.

B2Share and its API are under development, so this project allowed us to detect flaws that can be corrected by EUDAT—a new version will be launched soon. For example, when using the

B2Find search interface and looking for our data, we noticed that the *creator* field is different from the one on B2Share's side, which is not correct. Furthermore, a new module—B2Note—is being developed by EUDAT, which will enable researchers to deposit data in a triple store. This can be an opportunity to fully capture the metadata created in Dendro, which is stored in a triple store. By using triples as format for metadata, EUDAT will benefit from a more meaningful information representation, as the *subject-predicate-object* model handles data at several structuring levels and allows the representation of domain information using formal structures like ontologies. This leads to a more human-readable data and accounts for a better interoperability among data sources.

Conclusions

References

[ACRdSR16] Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, pages 1–12, 2016.

[ACRR14] Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. Labtablet: Semantic metadata collection on a multi-domain laboratory notebook. In *Metadata and Semantics Research Conference Proceedings*, volume 478, pages 193–205. Springer International Publishing, 2014.

[AH11] Dean Allemang and Jim Hendler. *Semantic Web for the Working Ontologist*. 2011.

[AR12] Chris Armbruster and Laurent Romary. Comparing Repository Types. *International Journal of Digital Library Systems*, 1(4):61–73, 2012.

[Bas12] Steve Baskauf. Beginner's Guide to RDF. https://code.google.com/archive/p/tdwg-rdf/wikis/Beginners.wiki, 2012. [Online; accessed on 1-February-2016].

[Bel09] Gordon Bell. Foreword. In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pages xi–xv. 2009.

[BJ15] Carolyn Bishoff and Lisa Johnston. Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University. *Journal of Librarianship & Scholarly Communication*, 3(2):1–27, 2015.

[Bl06] Tim Berners-lee. Linked Data - Design Issues. *Design Issues*, pages 1–6, 2006.

[Can05] Linda Cantara. METS: the metadata encoding and transmission standard. *Cataloging & classification quarterly*, 40(3/4):237–253, 2005.

[Ccs02] Ccsds. Reference Model for an Open Archival Information System (OAIS). *forSpace Data Systems*, (January):1–148, 2002.

[Com16] European Commission. Guidelines on Data Management in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, February 2016.

[CRR14] João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. Creating lightweight ontologies for dataset description. Practical applications in a cross-domain research data management workflow. In *IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 2014.

[Dek07] Dipen Deka. OAI-PMH : A Tool for Metadata Harvesting and Federated Search. *Planner*, pages 358–366, 2007.

[DHSW02] Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L. Weibel. Metadata principles and practicalities. *D-Lib Magazine*, 8(4), 2002.

[Dig13] Digital Curation Centre. What is digital curation? | Digital Curation Centre, 2013.

[Fay10] E Fay. Repository software comparison: building digital library infrastructure at LSE. *Ariadne*, (2009):1–11, 2010.

[Fur09] Betty Furrie. Understanding MARC Bibliographic: Machine-readable cataloging. In *Library of Congress website*. 2009.

[Gil08] Tony Gill. Metadata and the Web. *Introduction to Metadata*, pages 1–18, 2008.

[Gre03] Jane Greenberg. Metadata and the World Wide Web. In *Encyclopedia of library and information science*, volume 26, pages 1876–1888. 2003.

[Gue03] Rebecca S. Guenther. MODS: The Metadata Object Description Schema. *portal: Libraries and the Academy*, 3(1):137–150, 2003.

[Har10] Corey A Harper. Dublin Core Metadata Initiative: Beyond the element set. *Information Standards Quarterly*, 22(1):20–28, 2010.

[HB04] S. Harnad and T. Brody. Comparing the Impact of Open Access vs Non OA Articles in the Same Journals. *D-Lib Magazine*, Volume 10(6):1–6, 2004.

[ISO09] Information and documentation – The Dublin Core metadata element set, February 2009.

[ISO12] Space data and information transfer systems – Open archival information system (OAIS) – Reference model, September 2012.

[KNB] KNB. EML Specification. https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html. [Online; accessed on 1-February-2016].

[MBH+97] William K. Michener, James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1):330–342, 1997.

[NAS14] NASA. NASA Plan: Increasing Access To The Results of Scientific Research. http://science.nasa.gov/media/medialibrary/2014/12/05/NASA_Plan_for_increasing_access_to_results_of_federally_funded_research.pdf, November 2014.

[NOR08a] Michael Norris, Charles Oppenheim, and Fytton Rowland. Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR, 2008.

[NOR08b] Michael Norris, Charles Oppenheim, and Fytton Rowland. The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, 59(12):1963–1972, 2008.

REFERENCES

[Obi07] Marek Obitko. *Translations between Ontologies in Multi-Agent Systems*. dissertation, Czech Technical University in Prague Faculty of Electrical Engineering, 2007.

[PCHS07] Carole L. Palmer, Melissa H. Cragin, P. Bryan Heidorn, and Linda C. Smith. Data curation for the long tail of science: The case of environmental sciences. *3rd International Digital Curation Conference*, 2007.

[RR14] Robin Rice and Laine Ruus. Creating a Data Management Plan for your Research. In *Open research data: the future of science*, 2014.

[RRC14] João Rocha da Silva, Cristina Ribeiro, and João Correia Lopes. The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. *iPres Conference Proceedings*, 2014.

[SG08] Bjørn Sandvik and Bruce Gittings. Using KML for Thematic Mapping. *Institute of Geography\nSchool of GeoSciences*, MSc.:22, 2008.

[SR14] Guus Schreiber and Yves Raimond. RDF 1.1 Primer, 2014.

[SRGM04] Donald Sawyer, Lou Reich, David Giaretta, and Patrick Mazal. The Open Archival Information System (OAIS) Reference Model and its Usage. *SpaceOps*, pages 1–10, 2004.

[TTH11] Kristin M. Tolle, D. Stewart W Tansley, and Anthony J G Hey. The fourth Paradigm: Data-intensive scientific discovery. In *Proceedings of the IEEE*, volume 99, pages 1334–1337, 2011.

[VNLW04] Herbert Van De Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10:1–19, 2004.