UNIVERSIDAD AUTÓNOMA DE MADRID

CENTRO DE BIOTECNOLOGÍA Y GENÓMICA DE PLANTAS (CBGP)

MASTER'S THESIS

LIBRARIES AND SUPPORT TOOLING FOR FAIRIFICATION OF GERMPLASM DATABASES

Master's Degree in Bioinformatics and Computational Biology

Author: Aguayo Jara, Elena

Advisor: Wilkinson, Mark

Department of Biological Informatics

Tutor: Sánchez-Montañés, Manuel

Department of Computer Engineering

Course: 2023/24

Date: May 2024

Abstract

The origins of contemporary plant biodiversity conservation are largely linked to the practice of *in situ* conservation - protection of the environment within which the plant lives - and the practice of collecting, storing, and using seeds *ex situ*, known as germplasm banks. However, digitalisation of the data associated with these collections has been only partial. The databases of these banks have become outdated and disparate, collecting information in a wide range of data formats, storage systems, following different standards and with different levels of data accessibility and lack of documentation on how to interact with the data. As a result, it is extremely difficult to develop conservation strategies due to the barriers of combining what is known about a given species across all sites that include that species in their collections. To address this, this thesis aims to apply for the first time the FAIR Principles -Findable, Accessible, Interoperable, Reusable - to the data of the César Gómez-Campo Germplasm Bank (BGV-UPM) with the future aim of creating a network of germplasm databases to facilitate the use and re-use of the data. To start the FAIR transformation, a number of steps must be followed, some of them consist of mapping existing concepts into ontologies and using YARRRML - a templating language - to transform the new database into RDF and build the FAIR data model. This could help people to know what information are in seed banks and increase the usage of what they store.

Keywords

FAIR Principles, FAIR Data Point, Germplasm Bank, Seed Bank, Semantic Web, RDF, YARRRML, Ontology, Protégé, MIAPPE

Acknowledgements

The FLAIR-GG: Proyect TED2021-130788B-I00 is financed by MCIN/AEI /10.13039/501100011033 and by the European Union Next Generation EU/ PRTR. The FLAIR-GG computational infrastructure is supported by the "Severo Ochoa Program for Centres of Excellence in R&D 2022-2025” from the Agencia Estatal de Investigación of Spain, awarded to the Center for Plant Biotechnology and Genomics (CBGP-UPM-INIA/CSIC). Last, thanks to Universidad Autónoma de Madrid.

The development of this thesis was possible thanks to the members of the Wilkinson Laboratory for Biological Informatics of the CBGP, my advisor Mark D. Wilkinson and lab members Alberto Cámara and Pablo Alarcón. Also, I would like to express my gratitude to Santiago Moreno, director of the César Gómez Campo Germplasm Bank (BGV-UPM). Last but not least, I want to thank Paloma, Jose and Miguel for being by my side these last two years.

Abbreviations

AGRO	The Agronomy Ontology
ANAEETHES	AnaEE Thesaurus
API	Application Programming Interface
BCIO	Behaviour Change Intervention Ontology
BCO	The Biological Collections Ontology
CAO	Chemical Analysis Metadata Platform Ontology
CDAO	Comparative Data Analysis Ontology
CSV	Comma Separated Values
DOI	Digital Object Identifier
dwg	Darwin Core
ECO	Evidence and Conclusion Ontology
ECTO	Environment Exposure Ontology
EFO	Experimental Factor Ontology
ENM	eNanoMapper Ontology
EUPATH	VEuPathDB Ontology
ExO	Exposure Ontology
FLOPO	Flora Phenotype Ontology
FOODON	Food Ontology
GAZ	Gazeetter
GACS	Global Agricultural Concept Scheme
GEMET	GEneral Multilingual Environmental Thesaurus
GENEPIO	Genomic Epidemiology Ontology

GEO	Geographical Entity Ontology
GSSO	The Gender, Sex, and Sexual Orientation ontology.
HP	Human Phenotype Ontology
IAO	Information Artifact Ontology
ICO	Informed Consent Ontology
IDO	Infectious Disease Ontology
IDOMAL	Malaria Ontology
IRI	Internationalized Resource Identifier
MI	Minimum Information
MIAPPE	Minimum Information About a Plant Phenotyping Experiment
MICRO	Ontology of Prokaryotic Phenotypic and Metabolic Characters
MONDO	Mondo Disease Ontology
NBO	Neuro Behaviour Ontology
NGBO	Next Generation Biobanking Ontology
PATO	The Phenotype and Trait Ontology
PCO	Population and Community Ontology
PECO	Plant Experimental Conditions Ontology
PROV	The Provenance Ontology
RBO	Radiation Biology Ontology
RDF	Resource Description Framework
TO	Plant Trait Ontology
OBA	Ontology-Based APIs framework
OBI	Ontology for Biomedical Investigations

OBO	Open Biomedical Ontologies
OMIT	Ontology for MicroRNA Target
ONE	Ontology for Nutritional Epidemiology
ORNASEQ	Ontology for RNA Sequencing
ORG	The Organization Ontology
OWL	Web Ontology Language
SDGIO	Sustainable Development Goals Interface Ontology
SIO	Semanticscience Integrated Ontology
SQL	Structured Query Language
URI	Uniform Resource Identifier
v	Version
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
XSD	XML Schema Datatypes
ZP	Zebrafish Phenotype Ontology

Contents

1. Introduction

1.1. Data integration across de the web

The substantial volume of data generated in the life sciences underscores the need for effective data standards to facilitate reusability and sharing, which is becoming increasingly crucial in modern science (1). Standards have the potential to provide uniformity and consistency in the data produced by diverse researchers, institutions, and technologies. They are used to describe the study and how the data were generated, enabling not only enhanced comprehension of a particular set of experimental findings but also improved capacity to compare studies conducted by different scientists and organizations (2,3).

Data integration is an ongoing challenge for biological informaticians, often constituting a dedicated area of study, with numerous research groups worldwide approaching the problem from diverse perspectives. Integration is difficult for a variety of reasons, generally categorized into three core issues: syntax, structure, and semantics. Additionally, assigning and utilizing unique identifiers for data items and concepts is an essential requirement in biology. Syntactic barriers include aspects such as binary or textual format, and free-text or structured text; structural barriers involve elements like flat-file formats, and XML Schema; semantic barriers include inconsistent naming, naming conflicts (multiple things with the same name, or multiple names for the same thing) or insufficiently defined names; and finally identification issues involve non-unique identifiers, identifiers that can only be interpreted within a particular scope (e.g. in the context of a given database), non-opaque identifiers, and unstable or unpredictable identifiers (3).

While it is a prevalent practice to publish standardized sequencing data in public repositories, various other types of data are frequently exclusively disseminated through scientific publications. This task is particularly challenging in the field of plant phenotyping, given its inherent complexity and heterogeneity (1). To explain the implementations that have been carried out to standardize data, we will start by talking about the creation of the World Wide Web (WWW) and the Semantic Web.

By the close of the 1980s, Tim Berners-Lee introduced a groundbreaking concept: an architecture for disseminating hypermedia on the Internet, now recognized as the World Wide Web (4). The Web establishes a hypertext framework that interlinks documents throughout the Internet, facilitating connections between documents located on different machines. Integral to the Web's concept is the notion of an open community, where anyone can contribute their ideas to the collective knowledge for public access. Consequently, it has become a diverse amalgamation of various analyses, presentations, and summaries across subjects, created by individuals with the motivation to publish web pages. However, the Web is often criticized for being broad but shallow, featuring numerous gigabytes of information (5).

1.2. The FLAIR-GG Project

The loss of agrobiodiversity began to be noticeable from the 1960s partly because local varieties were replaced by genetically uniform high-yield varieties. This caused around 75% of plant genetic diversity to be lost by the 1990s (6). But it's not just that, we're also facing climate change, with a series of extreme weather events (e.g. major fires, droughts, floods, etc.) happening in our country - and globally - in recent years. This current situation is speeding up the loss of biodiversity and causing huge human pressure on habitats (changes in land use, pollution, overuse, etc.), impacting how ecosystems work (7).

This highlighted the need to take measures to conserve plant biodiversity. So, in 1966, the César Gómez-Campo Germplasm Bank at the Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas of the Universidad Politécnica de Madrid (BGV-UPM), which is part of the Department of Biotechnology and Plant Biology, was created. It was the first seed bank in Spain and also the first in the world to specialise in wild plant species. It currently conserves around 10,000 seed accessions from 3,500 different species. By accessions we mean the samples of collected material that are brought into the seed banks for conservation. Each accession represents a batch of genetic material from a single collection for a given species or defined biological population that is added to the bank (8). The BGV-UPM is fundamentally made up of two main collections, one dedicated to endemic and threatened species from the Iberian Peninsula, Balearic Islands and Macaronesian region, which includes the Canary Islands archipelago, and the other to species from the *Brassicaceae* family. More than 20% of Spain's endemic flora is conserved in the BGV-UPM (8,9).

However, digitalizing data associated with germplasm collections has only been partially done. Seed banks face multiple barriers that make it hard for computers to work together with their data. Some main problems include outdated or unmaintained databases, different data formats and underlying storage systems/software, lack of agreed standards for naming taxonomy, and lack of documentation on how to interact with the data. To address this, a collaborative project between BGV-UPM and CBGP-UPM-INIA-CSIC was proposed: FLAIR-GG-TED2021-130788B-I00. It is funded by MCIN/AEI /10.13039/501100011033 and by European Union Next Generation EU/ PRTR. This project aims to create a FAIR - Findable, Accessible, Interoperable, Reusable - way to represent seedbank data and improve the public metadata describing all participating seedbanks. All germplasm resources will be linked through a shared portal called the "Virtual Platform", enabling combined searches across all participants within the FLAIR-GG network.

1.3. Objectives of the study

It is important to point-out that this Master’s project is in large part aimed at developing the underlying tools, vocabularies, and models that will be used to execute the FLAIR-GG Network integration, with a focus on the César Gómez Campo Germplasm Bank. As such, most activities in this thesis are described in the Methodology section, and only a few example cases are described in the Results section. Therefore, a few goals have been pursued in this thesis:

- Harmonization of vocabularies used to describe fields in the germplasm database mapping them into ontologies and contribute to the maintenance of the FAO ontology.
- Build models and YARRRML mapping templates to represent the semantic relations of the germplasm data.
- Explore and evaluate relevant APIs that can be integrated into the FLAIR-GG virtual platform, enabling the platform to access and combine data from various external sources.

2. Background

2.1. The Semantic Web

The Semantic Web is an extension of the existing web to improve collaboration between computers and people. The Semantic Web is poised to impart structure to the meaningful content of web pages, establishing an environment where software agents traversing from page to page can seamlessly execute intricate tasks for users (4). The term "semantic" inherently suggests meaning or comprehension. The Semantic Web establishes connections between facts, allowing for references to specific pieces of information within documents or applications instead of linking directly to the entire document or application importantly, these connections are “labelled” with their meaning in a machine-readable manner – e.g. the purpose/meaning of the connection is explicit. This differs from the legacy Web, where hyperlinks have no meaning beyond the interpretation of the human reading the Web page and clicking the link. This, therefore, is the key distinction between Semantic Web technologies and other data-related technologies, such as relational databases or the World Wide Web itself, the Semantic Web focuses on the meaning of data rather than its structural or syntactic aspects (10).

Moreover, the familiar Web architecture currently sustains a distributed network of hypertext pages that can reference one another through global links known as Uniform Resource Locators (URLs). The Semantic Web takes this notion a step further by generalizing identifier to the Uniform Resource Identifier (URI) schema, and more importantly, allowing identifiers to represent “things” beyond just Web page. Unlike URLs, URIs can identify any entity existing in the world, not limited to just documents or the address of a web page (11,12). Because URIs are globally unique, and shared, any two web applications can refer to the same entity by using the same URI. Moreover, many URIs can also be “dereferenced” - that is, to utilize the structure of the URI (details such as server name, protocol, port number, file name, etc.) to locate a file (or a position in a file) on the web (4). The framework that is used to carry data and knowledge on the Semantic Web is Resource Description Framework (RDF), and this will be described in detail now (5).

2.2. Resource Description Framework

The World Wide Web Consortium (W3C) functions as a governing body for the internet, providing recommendations and creating approaches to facilitate the exchange of data and semantics. The Semantic Web relies on fundamental representation languages such as Resource Description Framework (RDF), Resource Description Framework Schema (RDFS), and Web Ontology Language (OWL), with RDF serving as the centre of everything (4). In 2004, the RDF gained approval as the central framework for the Semantic Web, paving the way for the development of "Linked Data" (13). In other words, RDF is a framework for representing information on the web.

The foundational model for all Semantic Web languages is the graph, in contrast to the traditionally used relational model. A graph can be seen as a collection of statements in the form of Subject-Predicate-Object (a collection of triples). These graphs are stored in specialized data management systems known as triplestores. The query language SPARQL (SPARQL Protocol and RDF Query Language) was developed to retrieve information from these resources (11). Visualization of an RDF graph can be achieved through a node and directed-arc diagram, where each triple is depicted as a node-arc-node link (Figure 1).

Subject
Predicate
Object

Figure 1. RDF triple and an RDF Graph. Generated using BioRender (https://www.biorender.com).

Generally, the subject and object are entities, people, or ideas, and the predicate represents the relationship between them. For instance, the statement "Madrid is part of Spain" has Madrid as the subject, part of as the predicate, and Spain as the object. Therefore, it could be expressed as the following simplified triple: "Madrid partOf Spain".

RDF applies the concept of the URI to resolve the identity issue in graph merging. The application is as follows: a node from one graph is merged with a node from another graph only if they have the same URI. It is used a simplified version of a URI abbreviation scheme called CURIEs (compact URI). In its most basic form, a URI expressed as a CURIE has two components: a namespace and an identifier. So the CURIE representation for the identifier Spain in the namespace geo is geo:Spain. It is important to know that URIs contain no embedded spaces. For example, an identifier "part of" is typically not used on the Web. Instead, we follow the InterCap convention (sometimes called CamelCase), whereby names that are made up of multiple words are transformed into identifiers without spaces by capitalizing each word (5). Thus, the previous triple example would look like "geo:Madrid geo:partOf geo:Spain". This property of the URI provides a straightforward way for the W3C to define the meaning of certain terms in the standard.

A few years later, RDF Schema (RDFS) emerged, it is a framework that helps define the expressivity of RDF and is an extension of the RDF Vocabulary being an extension of the RDF vocabulary (14). Specifically, it defines categories of resources (e.g. Classes, Datatypes), and adds features such as subclasses and labels (11). Built on top of RDF and RDF Schema is the Web Ontology Language (OWL). OWL is a Description Logic that emerged from a fusion of two preceding Description Logic frameworks Agent Markup Language (DAML) and Ontology Inference Layer (OIL). OWL is used to create ontologies that can be used to automatically (or manually) classify data, and OWL can be represented in RDF for the purpose of sharing these ontologies on the Web (15).

The term *ontology* is used to mean a certain kind of computational artifact – e.g. something akin to a program, an XML schema, or a web page – generally presented as a document. An ontology is a set of precise descriptive statements about some part of the world (usually referred to as the *domain of interest* or the *subject matter* of the ontology). Precise descriptions satisfy several purposes: most notably, they prevent misunderstandings in human communication, and they ensure that software behaves in a uniform, predictable way and works well with other software (15).

To sum up, the role of the Semantic Web is as follows:

- **Publish data sets:** publishing datasets as RDF is a data format that allows representing information in a more flexible and semantic way compared to other formats such as XML, this is because in XML the data structure is organized in a tree schema, which means that each element has a single parent and can contain multiple children. This can be limiting when it comes to representing complex relationships between different entities. On the other hand, RDF uses a graph-based structure. This means that data is represented as nodes (entities) and relationships (edges) between those nodes. This structure is more flexible and allows modelling more complex relationships between different entities. Additionally, RDF does not require a predefined schema, which means that we are not limited by a fixed data structure. This makes RDF more adaptable to changes in data and enables greater interoperability between different systems (16).

- **Linked data sets:** by making pieces of data accessible, shareable, and interconnected on the Semantic Web, it becomes easier to navigate through a "Web of Data" by following connections from one data item in one source to related data items in other sources. The key to this navigation lies in publishing data in the RDF format and establishing RDF links between data items across different sources (16).

- **Mapping models:** using the linking and reasoning features of the Semantic Web, we can connect ontologies with data entities (16).

- **Supportive metadata:** using RDF we can represent the provenance of integration systems in a flexible and extensible way, which means the origin or source of the data used in integration systems. For example, it could include details about the database from which the data was extracted, the extraction date, who performed the extraction, etc (16). RDF enables this representation in a flexible manner, meaning there is no predefined rigid data structure for describing provenance. Instead, it can be adapted to the specific needs of the integration system and can evolve over time. The extensibility capability of RDF means that we can add more information about provenance as needed without fundamentally changing the way data is represented. This is useful in environments where data provenance is critical for proper use and understanding in integration systems.

2.3. The FAIR Data Principles

Current data publishing paradigms are strongly dependent on the website or repository that will host that published data, resulting in a wide range of different formats, interfaces, and descriptive metadata depending on the source. In other words, given the rapid growth and evolution of the data environment, with new technologies and new and more complex data types, and the growth of general-purpose repositories, where the data types are not the same in all repositories, creating tools, in all computer languages, for all data types and all analytical tools required by those data types, is not a sustainable activity. The focus must therefore be on helping machines to discover and explore data through the implementation of more globally pervasive interoperability technologies and standards, which becomes the first priority for good data management (17,18).

The primary aim of adhering to FAIR principles (Box 1) is to standardize the requirements and expectations of data publishers, particularly focusing on their metadata, to enhancing the potential for data reuse by other researchers. It's noteworthy that these principles extend beyond just data and can be applied to algorithms, workflows, tools, websites, and ontologies. This ensures that all shared information is not only reproducible but also easily accessible. Importantly, these principles are not confined to human users but also pertain to the computational applications managing this data and their capacity to interpret the content they analyse. The interpretation process involves extracting the originally attributed meaning (semantics) from the document's content (17). With this objective in mind, a group of researchers convened in Leiden (Netherlands) in 2014 to address these issues comprehensively. The outcome of this conference was the formulation of the FAIR principles, emphasizing that all information must be Findable, Accessible, Interoperable, and Reusable (FAIR) for both computer systems and human users (19).

Failure to implement these principles with our data can give rise to a range of problems. For instance, a computer may be capable of processing a certain block of data, but lacks the ability to determine its license, thus cannot know if it has the permission to do so. Or a computer may recognize the data content (e.g. gene expression) but struggle to process them due to an unfamiliar format (19).

To be Findable:

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

F2. Data are described with rich metadata (defined by R1 below).

F3. Metadata clearly and explicitly include the identifier of the data it describes.

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

To be Accessible:

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

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

A1.2 The protocol allows for an authentication and authorization procedure, where necessary.

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

To be Interoperable:

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

I2. (Meta)data use vocabularies that follow FAIR principles.

I3. (Meta)data include qualified references to other (meta)data.

To be Reusable:

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

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

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

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

Box 1. The FAIR Principles (20).

Finally, in 2016, the Data Fair Principles were published for the first time (Box 1) by Wilkinson et al., which are based on the following points:

- **Findable:** the initial step in (re)utilizing data is to find them. Metadata and data should be readily accessible for both people and machines. Machine-readable metadata are crucial for the automatic discovery of datasets and services.

- **Accessible:** after the user locates the necessary data, they need to understand how the data can be accessed, potentially including authentication and authorization requirements.

- **Interoperable:** once the user has found the required data, they typically need to integrate it with other data. Additionally, the data need to be interoperable with applications or workflows for analysis, storage, and processing.

- **Reusable:** the goal of FAIR is to optimize the reuse of data. To accomplish this, metadata and data should be thoroughly described so that they can be reproduced and/or combined in various contexts.

It is important to note that the term "metadata" is used in cases where the principle should apply to both metadata and data and, also, it's essential to highlight that the FAIR principles draw a distinction between data and metadata. Notably, metadata plays a vital role, for instance, Principle F2 asserts that 'data are described with rich metadata, and the reusability principle emphasizes that metadata includes details about usage licenses (17).

Expanding on this, the concept of a Fair Data Point (FDP) is introduced. The FDP aims to establish a standardized approach to presenting metadata in accordance with the FAIR principles, making it easier for computers to process. Additionally, it aids information consumers in identifying the metadata associated with the provided resources (19). Commonly, the FAIR Data Point is used to expose metadata of datasets but metadata of other types of digital objects can also be exposed such as ontologies, repositories, analysis algorithms, websites, etc (21).

In the FLAIR-GG Project, a virtual platform has been designed to provide access to the PDFs of the César Gómez-Campo Germplasm Bank. The BGV FDP is hosted at the CBGP-UPM-INIA-CSIC (https://fdp.bgv.cbgp.upm.es/). The ultimate goal is to create the FDPs of all the associated seed banks and be able to make queries depending on the information you want to obtain.

2.5. Minimum Information About Plant Phenotyping Experiments

Plant phenotyping data is characterized by its diversity, dispersion, and increasing abundance due to the adoption of advanced automated techniques. While various initiatives have tackled the description of plant traits (e.g., the Plant Trait Ontology) and phenotypes, there is a lack of standardization in describing the experimental design that contextualizes phenotypic observations. A fundamental depiction of the materials employed, and the conducted experiment is crucial to support data discovery and data mining applications. MIAPPE, which stands for Minimum Information About a Plant Phenotyping Experiment, is currently under development to address these requirements (22). MIAPPE serves as a Minimum Information (MI) standard specifically focused in plant phenotyping and was initially developed within the context of the transPLANT, European Plant Phenotyping Network, and

ELIXIR-EXCELERATE projects. It outlines a set of attributes deemed essential for comprehensive documentation of a phenotyping experiment, following the model initially established for microarray data. Essentially, MIAPPE is an open, community-driven data standard intended to align data from various plant phenotyping experiments and computational sources, facilitating sharing, publication, and reuse. MIAPPE has a GitHub repository: https://github.com/MIAPPE.

The aim of MIAPPE is:

- To reduce the possibility of a researcher to lose crucial information in the documentation of an experiment.

- To facilitate the annotation of content with vocabularies relevant to the community.

- To promote an implementation of data format standards. MIAPPE represented an important step to achieve FAIRness for plant phenotyping data (1).

It is designed to handle datasets encompassing crops and woody plants cultivated in greenhouses, single fields, or experimental networks over periods ranging from one to several years. The primary components of MIAPPE datasets are the Dataset itself (or Investigation following the ISA framework, which is explained below), the Study, the Biological Material, and the Observation Variables (Figure 2). MIAPPE provides a specification comprising a checklist and a data model of metadata necessary for adequately describing plant phenotyping experiments (1,22). It's important to note that not all elements listed in an MI must be reported in every case. An MI document should be viewed as a checklist, consulted by individuals describing or depositing data to ensure the inclusion of all critical data characteristics relevant for the interpretation and potential replication of the research.

Before going any further, it is important to explain what the ISA framework is, which is an open source that help to manage (provide rich description of the experimental metadata) and reuse different life science datasets and follows the FAIR Principles. ISA stands for “**I**nvestigation” (the project context), “**S**tudy” (a unit of research) and “**A**ssay” (analytical measurement) (23). Therefore, Investigation and Study are homonymous in ISA and MIAPPE model, while Observed Variable corresponds to the ISA concept Assay (23,24).

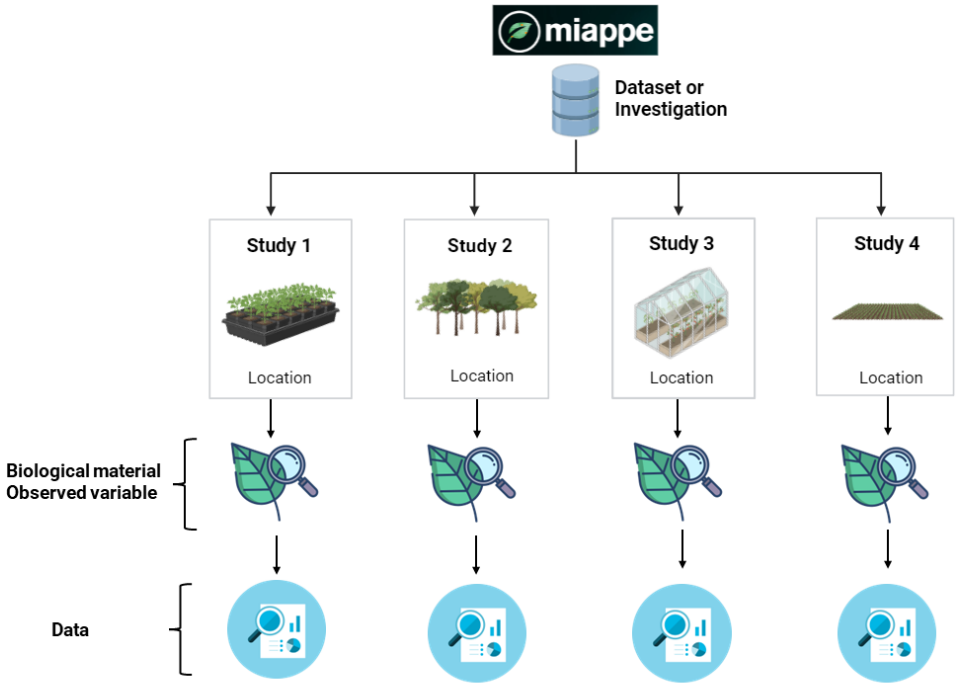

Figure 2. Primary components of the MIAPPE data model v1.1. Generated using Biorender (https://www.biorender.com)

In Figure 2, we can see four primary components, which are:

- **Dataset/Investigation:** the dataset serves as the foundational entity, covering essential information shared across all types of datasets. This includes details such as the title, list of authors, description, publication dates, DOI, licence, etc (these two latter fields are very important in the FAIRness process).

- **Study:** corresponds to one study at a specific location (must be single per study) and duration. This multiannual capability is particularly valuable for handling perennial plants like trees or grapevines. Additionally, it contains geographical coordinates providing the location context for the study. An investigation can have one or more studies.

- **Biological material:** identifies the biological material used in the studies, such as plants cultivated from a specific bag or seed, or those grown in a particular field. This description includes identification and traceability of the biological material. These materials must be identified with a unique identifier and is recommended to follow the FAO recommendations - Multi-Crop Passport Descriptors (MCPD) -, such as geographical coordinates, the origin of the biological material (e.g. gene bank accession or *in situ* material like orchard, forest, etc) and a identifier of a standard taxonomy such as NCBI (enabling interoperability and data linking through the use of identifiers).

- **Observed variable:** provides a detailed account of how a measurement was conducted. Documents a phenotypic or environment parameter that was observed and recorded as part of the study. It is linked to the method and unit of measurement employed (1,22).

3. Methodology

3.1. MAP concepts in FAO to existing ontologies

To proceed with the FAIRification of the BGV we need a set of descriptors, which have been obtained from the list of descriptors of the Food and Agriculture Organization of the United States (FAO) and the Millennium Seed Bank (25). The FAO list of descriptors was developed in collaboration with Bioversity International and is known as Multi-Crop Passport Descriptors (MCPD V 2.1) (26). These descriptors are a widely used international standard to facilitate germplasm passport information exchange, e.g. it helps to characterize plant-derived resources such as seeds, varieties, species, etc. In this way, researchers can communicate more efficiently about plant characteristics, describing what aspects are needed to be known in a study. In addition, the descriptors of the application used by the César Gómez-Campo Germplasm Bank to collect data, named CollApp, have also been included in this mapping.

The aim is to find terms that match across ontologies, so we stick with that definition, this is the way to ensure that computers can understand the semantics of terms, because machine-readable ontologies play a fundamental role in Semantic Web development.

These terms from FAO and Millennium Seed Bank are classified according to:

- Habitat description.
- Soil type (these descriptors come from the Millennium Seed Bank).
- Biological status of accession.
- Type of germplasm storage.
- Landform.

The descriptors not mentioned in this list come from CollApp. Once we have identified these descriptors, we proceed to search for them in existing ontologies through online repositories. We should note whether these terms we find are classes or properties and only keep the fields that are not individuals because ontologies try to define the semantic structure of a domain by modelling classes, relations, and properties between concepts. In addition, it is important to annotate their unique identifier code for each concept, usually these codes are URIs. The tools used to search for ontology terms are Ontobee, Ontology Lookup Service (OLS), Linked Open Vocabularies (LOV) and Agroportal.

OLS is a repository for biomedical ontologies, offering a convenient gateway to access the most up-to-date versions of these ontologies. Users have the option to explore the ontologies directly on the website or access them programmatically using the OLS API. This repository is developed and maintained by the Samples, Phenotypes, and Ontologies Team at EMBL-EBI (27).

On the other hand, LOV is a repository that covers a wide variety of ontologies in different fields, not only biomedicine like OLS, so it encourages the use of ontologies in different disciplines with the aim of improving interoperability and data integration on the semantic web. A noteworthy feature is that the information is not curated, as is the case in OLS and Ontobee (28).

The Agroportal project provides a central repository of reference ontologies for agriculture and food, using the OntoPortal technology (originally developed by the Stanford Center for Biomedical Informatics Research) (29).

Ontobee is a web page dedicated to managing and exploring ontology terms through linked data. It offers tools for visualizing, querying, and developing ontologies. With Ontobee, users can navigate and explore the details and hierarchical structure of specific ontology terms Additionally, Ontobee provides RDF source code for each web page, facilitating remote querying of ontology terms and enabling integration with the Semantic Web. Ontobee is developed by OBO Foundry library ontologies (30).

3.2. Edit OWL ontologies in Protégé

Once we have searched all the ontological terms, we use an application called Protégé (v.5.6.3). This app was developed by the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine and is an open-source ontology editor (31). To do this, we download the application and the OWL file that corresponds to the FAO ontology - since we are going to update this ontology because it was incomplete - in this GitHub repository and the file can be found in "*multi-crop-passport-descriptor-ontology/multi-passport-descriptor.owl*".

The first thing we are going to analyze is how many terms have their exact synonym annotated, since it is important to establish exact semantic relations. To carry out this task, we have executed a SPARQL query using the Yasgui web-based SPARQL editor and client (version December 2023) (32).

Endpoint: http://138.4.139.18:8890/sparql

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?s ?label ?o WHERE {

GRAPH <http://localhost:8890/DAV/home/LDP/Ontologies/multi-passport-descriptor.ttl>

        {?s <http://www.geneontology.org/formats/oboInOwl#hasExactSynonym> ?o .

         ?s rdfs:label ?label

        }

        } LIMIT 100
```

Box 2. The SPARQL Query used to analyze how many terms have their exact synonym annotated.

Thanks to this query (Box 2) we can quickly analyze which are the terms without their exact annotated synonyms. It is important to mention that the exact synonym of these terms is in the file containing the FAO list of descriptors. In addition, we have also checked that the ontology we are updating contains all the terms described in the MCPD file and also new fields have been added.

3.3. Build YARRRML mapping models for FAIR data transformation

YARRRML is a human readable text-based representation for declarative rules for generating RDF from a variety of native data formats including CSV, XML, and Relational/SQL. It is a subset of YAML, a widely used data serialization language designed to be human-friendly (33,34).

In this thesis we use YARRRML to convert the new seed bank database to RDF. A simpler way to do this is to use a Python package (v3.11) called EMbuilder. This package serves as a template for writing YARRRML and defines three main objects: prefixes (Box 3), triples, and configuration, that is it defines what is the source of input information.

```
data = {
  "prefixes": {
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "sio": "http://semanticscience.org/resource/",
    "foaf": "http://xmlns.com/foaf/0.1/",
    "prov": "https://www.w3.org/ns/prov#",
    "org": "https://www.w3.org/ns/org#",
    "dwc": "http://rs.tdwg.org/dwc/terms",
     "geo": "http://www.w3.org/2003/01/geo/wgs84_pos#",
     "efo": "https://www.ebi.ac.uk/efo/",
    "xsd": "http://www.w3.org/2001/XMLSchema#",
    "this": "http://my_example.com/"},
```

Box 3. Set of prefixes and namespaces used in this thesis (28).

To build these YARRRML mapping models, the seed bank database has been divided into three semantic models on which the YARRRML code is based, these are:

- Location model: the site of collection, type of soil, dates, etc.
- Administration model: who collected it, where is it stored, species-at-risk, etc.
- Germplasm model: species, taxon, phenotypic information, etc.

These models provide a visual representation of key concepts and their relationships, which facilitates the understanding and handling of the corresponding data. Each semantic model is explained in detail in this section, and each model can also be found in the GitHub repository under the path "*FAIR_GermplasmBank/Semantic Models/Images*".

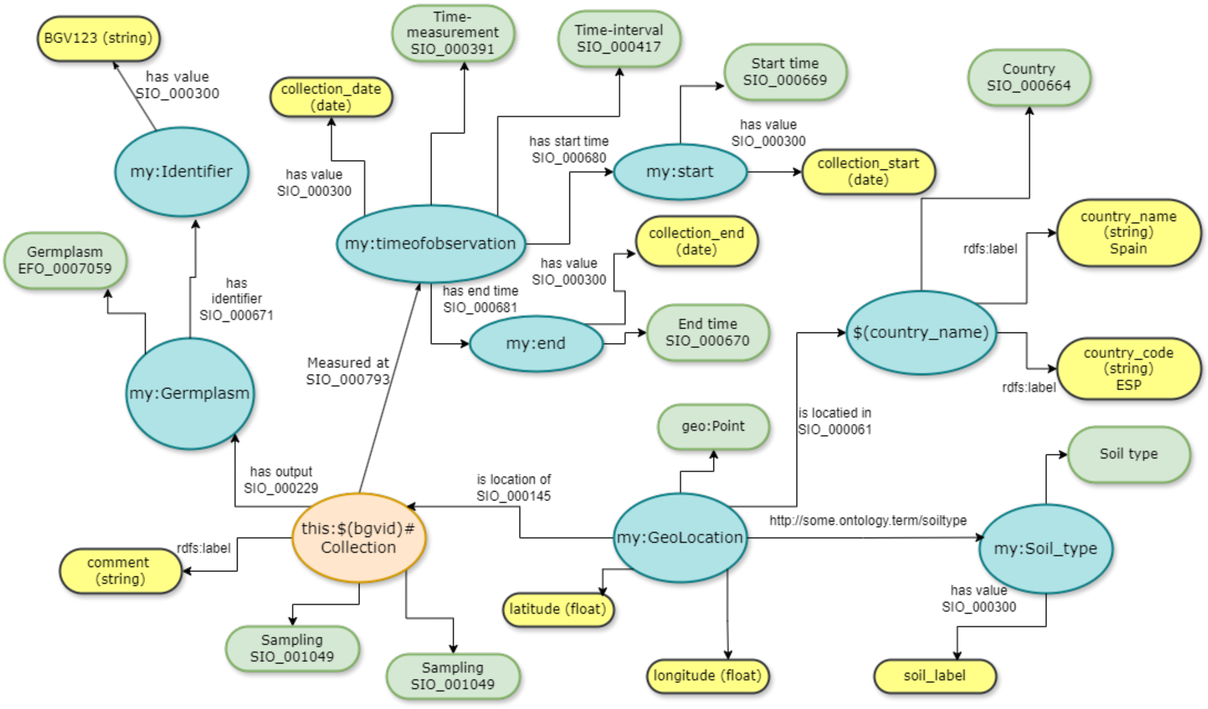

Figure 3. Location model. Generated using draw.io (https://www.drawio.com) and edited manually. Nodes in blue are URIs, nodes in yellow are literal values, and nodes in green are ontology terms. All unlabelled edges are rdf:type connections to ontological classes.

The main classes and properties represented in the location model (Figure 3) are:

- Collection: represents the event of collecting the germplasm (my:Collection), with properties like rdfs:label, and has as input the geolocation of this event (my:GeoLocation) and ouput the act of sampling (SIO_001049) and the time of observation the event (my:timeofobservation) and the germplasm collected (my:germplasm). This class has a unique identifier related to the sample collected (bgvid) and is shared in all three models.

- Germplasm: represents the germplasm collected (my:Germplasm). This has a identifier (SIO_000671) represented with the class my:Identifier, and is related with the ontological term germplasm (EFO_0007059).

- Time of observation: represents the observation period (my:timeofobservation), which has as output the start (SIO_000680) of this period (my:start) and the end (SIO_000680) of the observation period (my:end). Also has a literal value (SIO_000300) which is the date of the observation.

- GeoLocation: represents the geographical location of the collection, which hs as output the type of soil (my:soil_type) and geopraphical coordinates (geo:Point) properties such as ‘is_located_in’ (country) and ‘geo:Point’ (geographical coordinates) and is located in (SIO_000061) in some country (SIO_000664).

- Identifier: represents the unique identifier of the germplasm (my:Identifier), which has a value (SIO_000300).

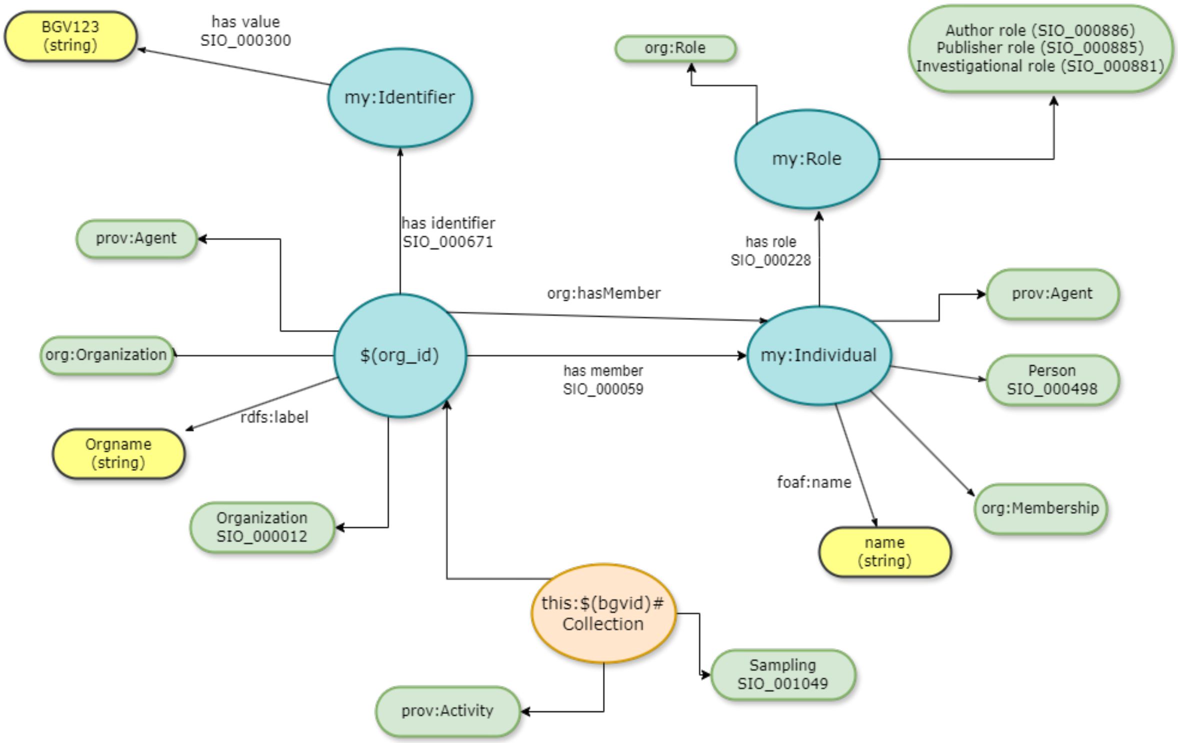

Figure 4. Administration model. Generated using draw.io (https://www.drawio.com) and edited manually. Nodes in blue are URIs, nodes in yellow are literal values, and nodes in green are ontology terms. All unlabelled edges are rdf:type connections to ontological classes.

The main classes and properties represented in the administration model (Figure 4) are:

- org_id: this class represents the name and identifier (URI) of the organisation involved in the collection or sampling activity. For example, the organisation could be “UBGV-UPM” with the URI “http://www.bancodegermoplasma.upm.es/”. This information would map to the “org:Organization” ontological term in the model.

- Individual: this class represents a unique identifier for a member of the organization (my:Individual). The class map with the ontological term “org:Membership” which indicates that a person is a member of the organization with no indication of the role played, and that individual is a person (SIO_000498).

- Role: this class (my:Role) would be a URI, which, according to the SIO ontology, represents the role (org:Role) of author role (SIO_000886), publisher role (SIO_000885) or investigational role (SIO_000881).

- Identifier: represents the unique identifier of the germplasm (my:Identifier), which has a value (SIO_000300).

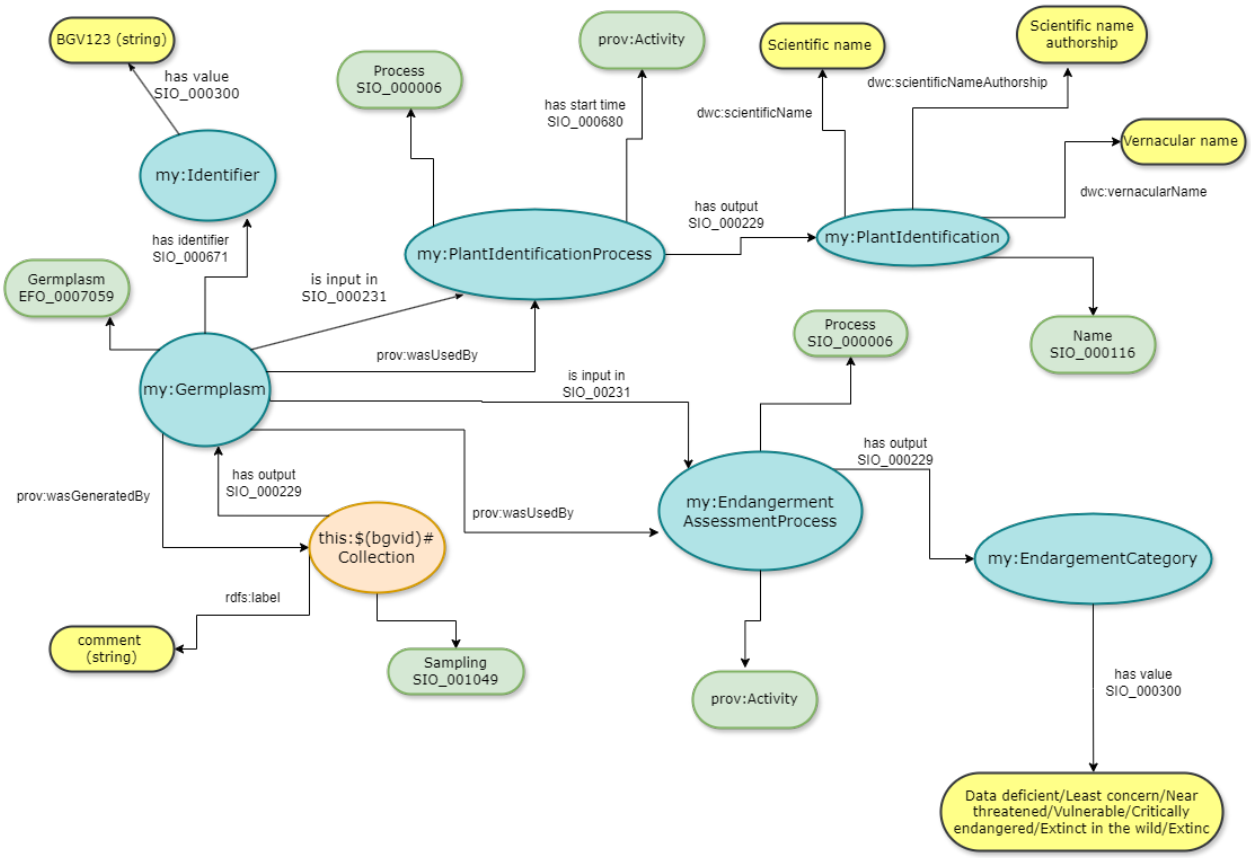

Figure 5. Germplasm model. Generated using draw.io (https://www.drawio.com) and edited manually. Nodes in blue are URIs, nodes in yellow are literal values, and nodes in green are ontology terms. All unlabelled edges are rdf:type connections to ontological classes.

This model (Figure 5) consists of the following main classes and properties:

- Plant identification: this class (my:PlantIdentification) represents the identification of a plant, which may have a scientific name (dwc:scientificName) and a vernacular name (dwc:vernacularName) associated with it.

- Plant identification process: this class represents the identification process of a plant, which has as input (SIO_000231) the germplasm (my:Germplasm) and as output (SIO_000229) the plant identification (my:PlantIdentification).

- my:Germplasm: this class represents the germplasm collected, which is related to Plant identification through the "prov:wasGeneratedBy" and "prov:was_used_by" properties. It is related with the ontological term germplasm (EFO_0007059), entity (SIO_000000) and prov:Entity.

- Collection: this class represents the event of collected the germplasm, and has as ouput (SIO_000229) the relationship with Germplasm (my:Germplasm).

- Endangerment assessment process: this class represents the process (SIO_000006, prov:Activity) of assessing the conservation status of a species (my:EndangermentAssessmentProcess).

- Endangerment category: this class represents the conservation status category assigned to a species (my:Endangerment category), which can take one of the values listed in the box, such as "Data deficient", "Least concern", "Near threatened", "Vulnerable", "Endangered", among others.

- Identifier: represents the unique identifier of the germplasm (my:Identifier), which has a value (SIO_000300).

3.4. Standardized CSV file

In order to migrate the database to the new database manager, in this case, MySQL, it is necessary to flatten the CSV file that contains all the information. The part carried out in this thesis has focused on homogenising the configuration of the date column - called COLLDATE - of the file by a Python script (v3.11) using the pandas library (v2.2.2).

The objective is to get the date to be yyyy-mm-dd. We may face different scenarios if we do not know a piece of information:

- In the case of not knowing the year, month, and the day: XX-XX-XX should appear in the excel cell .

- In the case of not knowing the month and the day: yyyy-XX-XX should appear in the excel cell.

- In the case of not knowing the day: yyyy-mm-XX should appear in the excel cell.

In addition, the script must remove any data that does not correspond to a date in that column, e.g. seasons of the year or person who entered the date.

3.5. APIs search

An API, or application programming interface, is a set of rules or protocols that enables software applications to communicate with each other to exchange data, features, and functionality (35). In this project we want to use APIs to draw information into the virtual platform to integrate with data available over the FLAIR-GG network; therefore, in this thesis we have made a search on the internet to find out which (relevant) APIs are available.

3.6. Code availability

All the code developed during this project is available in this GitHub repository: https://github.com/elena-aguayo-jara/FAIR_GermplasmBank. Also, this is the GitHub from the FLAIR-GG project: https://github.com/wilkinsonlab/FLAIR-GG. The path to the scripts used are described as *"Folder/File"*.

4. Results and Discussion

As mentioned above, the goal of this project is to apply the FAIR principles to the new Germplasm Bank database by following a series of steps.

4.1. MAP concepts into ontologies

First, we mapped the different database fields into ontologies by searching against OLS, Ontobee, LOV and Agroportal, taking note of whether these terms are classes or properties. Given that on many occasions these terms have been found with different meanings in ontologies, we also had to select the best match given our meaning. In this section we are going to show a series of examples that best represent the different situations we have had to face. The complete mapping file can be found in the GitHub repository, due to the extension of the file, under the path "*FAIR_GermplasmBank/MAP concepts into ontologies"*.

First case

The term we are going to show is *forest*, belonging to the wild habitat class.

1) First, we have searched for this term in OLS and Ontobee, where we have found two related terms (Box 4), one of them is *forest* as such and the other is *forest ecosysten*. Both terms are classes and we have also noted their identification code: ENVO_00000111 and ENVO_01001243 belong to the Environment Ontology (ENVO). In addition, OLS allows us to know in which other ontologies these terms are found, which are annotated in Box 4.

Class forest ENVO_00000111: An area with a high density of trees. A small forest may be called a wood. The definitions of forest can vary greatly, and different classes will be needed to support the major categories. Tree cover alone is not enough to distinguish between forests and plantations. The international definition proposed by the 2010 FAO Forestry Resource Assessment: "land spanning more than 0.5 ha with trees higher than 5 metres and canopy cover of more than 10 percent, or trees able to reach these thresholds in situ. It does not include land that is predominantly under agricultural or urban land use." - FAO. 2010.

- Ontologies: ENVO, ECTO, EUPATH, MICRO, AGRO, GENEPIO, RBO, MONDO, GAZ, FOODON.

Class forest ecosystem ENVO_01001243: An ecosystem which is determined by communities of plants with a tree growth form and in which members of those communities form continuous or discontinuous regions of canopy cover.

- Ontologies: ENVO, ECTO, AGRO, GENEPIO, RBO, MONDO, GAZ, FOODON.

Box 4. Field *forest* found in OLS and Ontobee.

2) Next, we have used the LOV search engine, where we found three classes (Box 5):

- The first class belongs to an ontology called Ontomedia, whose prefix is osr. We have noted the link leading to this term in the ontology, however, this link does not work.

- The second class belongs to the ontology LinkedGeoData (prefix lgdo), in this case it is similar, the link returns error 404.

- The third link corresponds to the Proton ontology (prefix pext). In this case, the link does work and leads us to a definition annotated in Box 5.

Class osr:Forest (osr).

- Ontology: Ontomedia.
  - URI: http://contextus.net/ontology/ontomedia/core/space#Forest
    This link returns the following message "That page can't be found".

Class lgdo:Forest (lgdo).

- Ontology: LinkedGeoData.
  - URI: http://linkedgeodata.org/ontology/Forest
    This link seems to correspond to a GitHub page but returns a 404-error message.

Class pext:Forest (pext): Dense growths of trees, plants, and underbrush covering a large area.

- Ontology: Proton.
  - URI: http://www.ontotext.com/proton/protonext#Forest

Box 5. The term *forest* found in LOV.

3) Finally, we use Agroportal, where we find the term forest with the same definition in two thesauri (Box 6), but in no ontology so this does not provide us with the logical relationships between concepts, e.g. they do not express classes or properties.

Forest: A vegetation community dominated by trees and other woody shrubs, growing close enough together that the tree tops touch or overlap, creating various degrees of shade on the forest floor. It may produce benefits such as timber, recreation, wildlife habitat, etc.

- Thesaurus:
  a. AnaEE Thesaurus (ANAEETHES).

     Link: http://opendata.inra.fr/anaeeThes/c3_2626

  b. GEneral Multilingual Environmental Thesaurus (GEMET).

     Link: https://www.eionet.europa.eu/gemet/en/concept/3406

Box 6. The term *forest* found in Agroportal.

4) Finally, we select the best match and, in this case, the class forest ENVO_01001243 (Box 4). The hierarchical classification of this term also supports its selection.

Second case

As we can see, there are ontologies that no longer work, something we have come across on numerous occasions. The term we are going to talk about this time is *seed company* (Box 7). We only found this term in Agroportal, and one of the results is the Multi-Crop Passport Ontology (MCPDO), e.g., the FAO ontology, which in turn shows the link to Crop-Ontology (CO) (36), an ontology created by the union of Bioversity International and the International Center for Tropical Agriculture (CIAT); however, all the links to Crop Ontology only return an RDF/XML file. In addition, when we enter the identifier of this term (CO_020:0000061) in the tool of Crop Ontology we found nothing (the identifier appears in the link to the RDF/XML file of Box 7). This is because Crop Ontology is no longer maintained.

Ontology: Multi-Crop Passport Ontology.

- Agroportal link: https://agroportal.lirmm.fr/ontologies/CO_020?p=classes&conceptid=http%3A%2F%2Fwww.cropontology.org%2Frdf%2FCO_020%3A0000061
- Link of the RDF/XML file: https://cropontology.org/rdf/CO_020:0000061

Box 7. The term *seed company* found in Agroportal.

Because the FAO ontology is going to be of enormous utility to us, we made the decision to restore it, and publish it in a domain that we control (the FLAIR-GG Project GitHub). However, it is undesirable to "brand" a widely used ontology with a project name in the URL, and so we use the W3ID redirection system (https://w3id.org) to create permanent, generic identifiers for all the concepts, which are served by the OWL file stored in our GitHub. In these cases where the term appears only in the FAO ontology and without definition, the domain experts will be responsible for providing it at a later stage.

4.2. OWL file edited with Protégé

As mentioned above, the seed bank is going to use the FAO ontology, once the new ontology terms are mapped, we use Protégé to edit the OWL file of the FAO ontology. In this thesis we check which terms do not have their exact synonym annotated:

- Specific terms from *Biological status of accession*.
- Specific terms from *Type of germplasm storage*.
- Specific terms from *Collecting/Adquisition source*.

These terms have no annotated synonyms, but neither do they have synonyms in the FAO descriptor file, both because they are more specific and because fewer people are likely to use them, and because there are fewer synonyms per se. In addition, we noticed that in the OWL file there were repeated terms with a small change in the name, e.g. *Biological status of accession code* and *Biological* *status of accession* (the latter term is the one that appears only in the MCPD list). This could be because they were added because someone needed to perform a task and did not realise that these terms were already added. Furthermore, we have checked that there are a number of classes that do not appear in the OWL file and should be added: persistent unique identifier, collecting institute name, collecting institute address, MSL status of the accession, donor institute name, collecting mission identifier, coordinate uncertainty, coordinate datum, georeferencing method, breeding institute name, institute maintaining safety duplicates, in addition to all the terms from CollApp (accession codes, germplasm bank codes, location, original taxonomy, common name, dates, motives and permits, *in situ* conservation status or conservation actions, environment, population, sample, photographs, duplicates of BGV-UPM accessions, donations to the BGV-UPM, archaeological fields, management fields in bansem).

To sum up, the function of annotating exact synonyms is that regardless of, for example, the language used to search for a term, we will be able to find it because we would search for it by its synonym. An example is to search for SAMPSTAT instead of *Biological status of accession*.

4.2. YARRRML mapping models

Using the ontology terms we have identified and/or added to FAO, we can now build the YARRRML templates representing our semantic models. This allows us to automatically transform the data from a CSV file (Table 1) to RDF that is compliant with the structure and semantics of those models. Before continuing, it is important to make it clear that these models are not final as more classes and properties are still being added but creating them already serves to have an initial skeleton on which to build the FLAIR-GG Network.

As mentioned above, all the code used and the CSV files can be found in the GitHub repository under the path "*FAIR_GermplasmBank/Semantic Models*". In this section we are only going to show a simple example of the code (Box 8). In this case, we have focused on the location model (Figure 3).

```
"triplets": [

["this:$(uniqid)#Collection","sio:SIO_000229","this:$(uniqid)#Germplasm","iri"],

["this:$(uniqid)#Collection","sio:SIO_000793","this:$(uniqid)#timeofobservation","iri"]

["this:$(uniqid)#Collection","rdfs:label","$(comment)","xsd:string"],

["this:$(uniqid)#timeofobservation","sio:SIO_000300","$(collection_date)","xsd:date"],
```

Box 8. Abbreviated example of the YARRRML template of the location model. Fields that match the $( ) pattern are dynamically filled by finding the associated column in the CSV file, and iteratively pulling data in from each line in that file.

uniqid	comment	collection_date
12345	“Some data”	2023-01-01

Table 1. Abbreviated example of the CSV file of the location model.

- The first triplet in box 8 establishes a relationship between the class “my:Collection” (represented by the class instance 'this:$(uniqid)#Collection', which is the subject) and the class 'my:Germplasm' (represented by the class instance 'this:$(uniqid)#Germplasm', which is the object) using the property 'sio:SIO_000229' (meaning 'has output' in SIO).

- The second triple relates the class “my:Collection” to the class “my:timeofobservation” (represented by “this:$(uniqid)#timeofobservation”) using the property 'sio:SIO_000793' (meaning 'has temporal observational component' in SIO). In other words, this triple defines that the collection class has an associated time component.

- The third triplet relates the class “my:Collection” (represented by “this:$(uniqid)#Collection”) to the object ‘comment’ which is a column of the CSV file and is a string (represented by “xsd:string”). The predicate is the property “rdfs:label”, this indicates that a label or textual description (in this case the object ‘comment’) is being assigned to the class instance ‘this:$(uniqid)#Collection”.

- The fourth triplet relates the class “my:timeofobservation” (represented by the class instance “this:$(uniqid)#timeofobservation”) to the object “collection_date”. It maps the value of the “collection_date” column of the CSV as the value (SIO_000300' – “has value”) of an instance of the class “my:timeofobservation”, representing the date of the observation or sample collection (xsd:date).

In the triples explained above, the class instances “this:$(uniqid)#Germplasm” and “this:$(uniqid)#timeofobservation” are described as “iri”. IRI stands for “Internationalized Resource Identifier”. The IRI is a generalization of the URI that uses all the characters from different languages, can include characters with accents and symbols, all the characters that are part of the UNICODE character set (4). Also, in these lines of code, we often see the term 'this' appear, which refers to the prefix defined for the base URI of the triples and must be added to that base URI to build the full IRI, so that resources can be linked more easily on the web. This base URI is defined as basicURI in the code configuration shown below (Box 9):

```
"config": {
      "source_name": "source_cde_test",
      "configuration": "default",
      "csv_name": "location.csv",
      "basicURI": "this"
      }
```

Box 9. Configuration of the location model.

4.3. Standarized CSV file

After performing the YARRRML transformation, one of the steps to follow was to collaborate with the standardisation of the CSV file that contains all the information of the seed bank, specifically we focused on the column that stores the dates since not all of them followed the same pattern. We can see the results in table 2 below:

Unformatted dates	Formatted dates
1960----	1960-XX-XX
1960-verano	1960-XX-XX
1960-09-00	1960-09-XX
13/09/1960	1960-09-13
	XX-XX-XX

Table 2. Abbreviated example of formatted dates of the Location model CSV file.

One problem that arose was that the script would fill empty cells with the next cell, instead of adding XX-XX-XX directly, this caused the data to be out of order. This is because the “skip_blank_lines” parameter of the Pandas library's “pd.read_csv()” function is set to true by default, causing it to skip blank lines. By changing this value to false, all blank cells in the CSV file are read as null values (NaN) and filled with XX-XX-XX.

4.4. Relevant APIs available

Finally, the APIs found that will help to integrate with the data of the FLAIR-GG network are as follows:

a. **IUCN Red List of Threatened Species (v3)**: provides information about global extinction risk status of animal, fungus and plant species (37):

- Data about species: We find taxonomic data on species, endangered species, threat categories, habitats, regions, etc.
- Geographical data: We have distribution data of threatened species from the IUCN. It provides maps and layers of species distribution ranges in different coordinate systems and formats.

  Note: The Red List API v3 will be replaced soon by v4.

b. **Ministerio de Transformación Digital y Servicio Público:** API v0.0.1 of the *Portal de Datos Abiertos datos.gob.es* allows access to information such as:

- Different datasets by applying various filters such as ID, title, format, word class, modification date, etc.
- Taxonomy of geographical coverage: provinces, autonomous communities, etc.
- Taxonomy of primary public sectors.
- List of publications, topics and geographical locations (38).

c. **FAOSTAT:** This API allows access to the FAOSTAT database, that is the statistical database from FAO (39).

d. **FAODATA:** This is a Python libray (v1.1) to download data from FAO (40).

e. **GAEZ Data Portal:** This API (v4) provides access to satellite images, area photographs, soil maps, topography, etc. GAEZ stands for "Global-Agro Ecological Zones" (41), is a project developed by FAO and the International Centre for Tropical Agriculture (CIAT). GAEZ provides information on agro-ecological conditions in different parts of the world. In addition, FAO has recently developed a Python package called PyAEZ, which provides user-friendly, and intuitive ways to input data and output results.

f. **Node Swagger API:** FAO releases this API (version April 2024) which provides access to a range of different data, including weather data for countries, cities, etc (42).

g. **AGROVOC:** This a multilingual vocabulary designed to include concepts and terminology from FAO's areas of interest (43). It is a relevant set of open data related to agriculture, available for public use, and its major impact is to facilitate access and visibility of data across domains and languages. To access this vocabulary, there is an API (v1).

h. **Centro Nacional de Recursos Fitogenéticos y Agricultura Sostenible (CRF):** It provides information on more than 75,000 entries of cultivated germplasm that are conserved *ex-situ* in the "Programa Nacional de Conservación y Utilización Sostenible de los Recursos Fitogenéticos para la Agricultura y la Alimentación" (PNRF) (44). To access their information there is also an API (version April 2024).

i. **Genesys:** This is the database of the Grin Global Project (45). In this system, genebanks store and manage plant genetic resources data. It has an API (version 2024.1) to access their information.

j. **Global Biodiversity Information Facility (GBIF):** Its main purpose is to make it easier to publish, find, access, and use data on biodiversity, species distribution, scientific names, and so on (46). To access all this information there is also an API. Most of the GBIF API is at v1 and Maps API at v2. In addition, GBIF has also developed a Python library to access its data, called pygbif o rgif if the R programming language is preferred.

k. **World Flora Online (WFO):** It is an international collaborative project that aims to create a comprehensive and constantly updated checklist of all known plant names in the world, along with their distribution, habitats, conservation and other relevant information (47). To access this information, WFO provides a list of APIs.

4.5. Future work

There are some tools that we can use to improve this project. First, we have the FAIR Cookbook, which is an open source developed by the community of Life Sciences professionals, including members of the ELIXIR community. This tool serves as a guide for creating and maintaining FAIR data by offering cases - in this case called recipes - of FAIRification in different domains (48). We could contribute to the expansion of this tool with new recipes of our own by publishing the process followed to implement the FAIR Principles in the Germplasm Bank and the development of the FLAIR-GG Network. There is also another tool, ELIXIR Research Data Management Kit (RDMkit), this is an open source software that allows us to apply different techniques on our data to evaluate its FAIRification level and how to convert the data into FAIR (49). RDMKit could be used to study the FAIRness level of the data and metadata of seed banks wishing to join the FLAIR-GG Network, and to help control the process of applying the FAIR Principles, thus perhaps reducing the time spent on studying the FAIRness level of the data; RDMKit also recommends specific measures to FAIRify the data, such as FAIR Cookbook, so these tools could also be recommended to seed banks to know how to start the FAIRification of their own data. Futhermore, there is a software framework called **e**lectronic **D**ata **A**rchive **L**ibrary (e!DAL) developed by the Leibniz Institute of Plant Genetics and Crop Plant Research and serves to maintain, publish and share data in a FAIR way in the cloud assigning them persistent identifiers, such as DOI or URI, making the data more accessible by storing primary datasets and facilitating their transformation into RDF data. (50). In addition, e!DAL also offers suggestions on how to improve data standardisation and provides an API that would facilitate its integration with data from the FLAIR-GG Network.

On the other hand, it could be interesting to work in the future with an application called PhenoApp, which integrates MIAPPE and the FAIR Principles. PhenoApp is an open-source Android mobile application for collecting plant phenotypic data in the field and in greenhouses. The tool offers the possibility to customize the list of descriptors for any possible scenario, and the data can be saved and exported as Excel file (51). The idea would be to integrate this app into the application used by the César Gómez-Campo Germplasm Bank to collect data, named CollApp. For more information, the source code of PhenoApp is available here: https://gitea.julius-kuehn.de/JKI/pheno-app

Focusing more on the FLAIR-GG Network, it might be interesting to train a Large Language Model, like a chatbot, with the database to help non-technical users create their SPARQL queries, thus facilitating the search for information. Moreover, it could be useful to create machine-understandable representations of international treaties that regulate the exchange of seeds using Open Digital Rights Language (ODRL), which is used to represent access conditions into a machine readable-way (52). This would be created to facilitate the work for the people in charge of accepting whether or not a person can obtain seeds from a certain seed bank, but always the final decision would be taken by a person, not the machine. Later, an algorithm will be created that will match queries with their corresponding databases, this will be done by applying Shapes Constraint Language (SCHACL) - a W3C recommendation -, which is a language used to validate that RDF data satisfies a set of conditions, where these conditions are called "shapes" (53); in this way, the virtual platform will indicate which databases could solve a given query, since right now you decide which database to search in and then perform the query.

5. Conclusion

In this paper we proceed to apply the FAIR Principles to the César Gómez-Campo Germplasm Bank database with the aim of making the data more accessible and reusable. This is the first time that the FAIR Principles have been applied to a germplasm bank. The aim is to be able to expand the FLAIR-GG Network and have many more seed banks join and thus have a network of interconnected banks.

Adherence to the FAIR Principles would significantly benefit the objectives of seed banks, aiding integrated conservation strategies, rapid and extensive germplasm searches through seed collections based on several variables: taxonomy, geography, climatic ranges at the collection site (of utmost importance in a world where the climate is changing), soil type, etc. FAIR data makes the germplasm collection more discoverable, increasing the use of the seeds they store. It also makes the database more interoperable with public databases such as those of weather and climate agencies, geographic information agencies, publicly available annotated collections of nucleotide sequences and their translations into proteins, and geological and soil agencies.

It should be noted that in this thesis the FAIRification process is not finished as the development of the harmonisation of the vocabulary and the subsequent design of the semantic models has taken a long-time, all-in order to have a network as organised as possible, but it has served to lay the foundations of this project. This will allow the FLAIR-GG Network to expand and have more seed banks incorporated, thus creating an interconnected network. This thesis lays the foundation for greater accessibility, interoperability, and reuse of germplasm data, which in turn improve and increase research and conservation strategies in the field of plant biodiversity. As the implementation of the FAIR Principles is completed, it is expected that the FLAIR-GG network will become a valuable resource for the scientific community.

References

1. Papoutsoglou EA, Faria D, Arend D, Arnaud E, Athanasiadis IN, Chaves I, et al. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol. 2020 Jul;227(1):260–73.

2. Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, et al. Data Standards for Omics Data: The Basis of Data Sharing and Reuse. Methods Mol Biol Clifton NJ. 2011;719:31–69. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152841/

3. González AR, Callahan A, Cruz-Toledo J, Garcia A, Egaña Aranguren M, Dumontier M, et al. Automatically exposing OpenLifeData via SADI semantic Web Services. J Biomed Semant. 5(1):46. Available from: https://doi.org/10.1186/2041-1480-5-46

4. Berners-Lee T, Hendler J, Lassila O. The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Available from: https://www.researchgate.net/publication/225070375_The_Semantic_Web_A_New_Form_of_Web_Content_That_is_Meaningful_to_Computers_Will_Unleash_a_Revolution_of_New_Possibilities

5. Allemang D, Hendler J, Gandon F. Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL. Third Edition. 2020. 510 p. Available from: https://doi.org/10.1145/3382097

6. FAO. What is Agrobiodiversity?. Available from: https://www.fao.org/3/y5609e/y5609e02.htm

7. Torres E, Iriondo JM. La conservación de los parientes silvestres de los cultivos y la necesidad de publicar datos según los principios FAIR. Conserv Veg. 2022 Dec 19;(26):3–6. Available from: https://revistas.uam.es/conservacionvegetal/article/view/16403

8. Bacchetta G, Bueno Sánchez A, Fenu G, Jiménez-Alfaro B, Mattana E, Piotto E, et al. Conservación ex situ de plantas silvestres. Principado de Asturias / La Caixa; 2008. 378 p.

9. Banco de Germoplasma Vegetal 'César Gómez Campo' (BGV-UPM). Available from: http://www.bancodegermoplasma.upm.es/

10. Cambridge Semantics. Cambridge Semantics. Introduction to the Semantic Web. Available from: https://cambridgesemantics.com/blog/semantic-university/intro-semantic-web/

11. Antezana E, Mironov V, Kuiper M. The emergence of Semantic Systems Biology. New Biotechnol. 2013 Mar 25 [cited 2024 Feb 20];30(3):286–90. Available from: https://www.sciencedirect.com/science/article/pii/S1871678412008655

12. Ruttenberg A, Rees JA, Samwald M, Marshall MS. Life sciences on the Semantic Web: the Neurocommons and beyond*. Brief Bioinform. 2009 Mar 1;10(2):193–204. Available from: https://doi.org/10.1093/bib/bbp004

13. W3C. RDF Concepts and Abstract Syntax. Available from: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#resources-and-statements

14. W3C. RDF Schema. Available from: https://www.w3.org/TR/rdf-schema/

15. W3C. OWL 2 Web Ontology Language. Available from: https://www.w3.org/TR/owl2-primer/

16. Goble C, Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform. 2008 Oct 1;41(5):687–93. Available from: https://www.sciencedirect.com/science/article/pii/S1532046408000178

17. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018. Available from: https://www.nature.com/articles/sdata201618

18. Papoutsoglou EA, Athanasiadis IN, Visser RGF, Finkers R. The benefits and struggles of FAIR data: the case of reusing plant phenotyping data. Sci Data. 2023 Jul 13;10(1):457. Available from: https://www.nature.com/articles/s41597-023-02364-z

19. da Silva Santos LOB, Burger K, Kaliyaperumal R, Wilkinson MD. FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. Data Intell. 2023 Mar 8;5(1):163–83. Available from: https://doi.org/10.1162/dint_a_00160

20. GO FAIR. FAIR Principles. Available from: https://www.go-fair.org/fair-principles/

21. FAIR Data Point. Available from: https://www.fairdatapoint.org/

22. MIAPPE (EMBL-EBI). Available from: https://www.miappe.org/

23. ISA tools | Standardizing metadata for scientific experiments. Available from: https://isa-tools.org/index.html

24. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 2010 Sep 15;26(18):2354–6. Available from: https://doi.org/10.1093/bioinformatics/btq415

25. Millennium Seed Bank. 2007. Available from: https://www.kew.org/wakehurst/whats-at-wakehurst/millennium-seed-bank

26. Alercia A, International B, Nations F and AO of the U, Diulgheroff S, Mackay M. FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 [MCPD V.2.1] - December 2015. 2015; Available from: https://hdl.handle.net/10568/69166

27. Côté RG, Jones P, Martens L, Apweiler R, Hermjakob H. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W372–6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447739/

28. Vandenbussche PI, Atemezing GA, Poveda-Villalón M, Vatant B. Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web - IOS Press. Semantic J. 2016 Dec 6;8(3):437–52. Available from: https://content.iospress.com/articles/semantic-web/sw213

29. Jonquet C, Toulet A, Arnaud E, Aubin S, Dzalé Yeumo E, Emonet V, et al. AgroPortal: A vocabulary and ontology repository for agronomy. Comput Electron Agric. 2018 Jan 1;144:126–43. Available from: https://www.sciencedirect.com/science/article/pii/S0168169916309541

30. Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D347–52. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210626/

31. Musen MA. The Protégé Project: A Look Back and a Look Forward. AI Matters. 2015 Jun;1(4):4–12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/

32. Rietveld L, Hoekstra R. YASGUI: Not Just Another SPARQL Client. In: Cimiano P, Fernández M, Lopez V, Schlobach S, Völker J, editors. The Semantic Web: ESWC 2013 Satellite Events. Berlin, Heidelberg: Springer; 2013. p. 78–86.

33. Van Assche D, De Meester B, Heyvaert P, Dimou A. W3C. YARRRML. Available from: https://rml.io/yarrrml/spec/#bib-YAML

34. YAML Ain't Markup Language (YAML^{TM}). Available from: https://yaml.org/spec/1.2.2/

35. ¿Qué es una interfaz de programación de aplicaciones (API)? | IBM. Available from: https://www.ibm.com/es-es/topics/api

36. Alliance Bioversity International - CIAT. Environment Protection. Available from: https://alliancebioversityciat.org/alliance-accelerated-change-preserve-and-protect-our-environment

37. IUCN Red List of Threatened Species. IUCN. 2023. The IUCN Red List of Threatened Species. Version 2023-1. Available from: https://www.iucnredlist.org/en

38. API - Ministerio para la Transformación Digital y Servicio Público. Available from: https://datos.gob.es/es/apidata#!/dataset/findDatasetById

39. FAOSTAT. Available from: https://www.fao.org/faostat/en/#home

40. FAODATA - PyPI. Available from: https://pypi.org/project/faodata/

41. Gaez Data Portal. Available from: https://gaez.fao.org/pages/data-access-download

42. API - FAO. Available from: https://api.finto.fi/doc/

43. AGROVOC. Available from: https://www.fao.org/agrovoc/es/sobre-agrovoc

44. Centro de Recursos Fitogenéticos y Agricultura Sostenible (CRF). Available from: https://www.inia.es/unidades/Institutos%20y%20Centros/CRF/Paginas/Home.aspx

45. GRIN Project. Available from: https://www.grin-global.org/

46. GBIF. Available from: https://www.gbif.org/

47. World Flora Online. Available from: https://www.worldfloraonline.org/

48. Rocca-Serra P, Gu W, Ioannidis V, Abbassi-Daloii T, Capella-Gutierrez S, Chandramouliswaran I, et al. The FAIR Cookbook - the essential resource for and by FAIR doers. Sci Data [Internet]. 2023 May 19;10(1):292. Available from: https://www.nature.com/articles/s41597-023-02166-3

49. ELIXIR (2021) Research Data Management Kit. A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075). Available from: https://rdmkit.elixir-europe.org/

50. Arend D, Lange M, Chen J, Colmsee C, Flemming S, Hecht D, et al. e!DAL - a framework to store, share and publish research data. BMC Bioinformatics. 2014 Jun 24;15(1):214. Available from: https://doi.org/10.1186/1471-2105-15-214

51. Röckel F, Schreiber T, Schüler D, Braun U, Krukenberg I, Schwander F, et al. PhenoApp: A mobile tool for plant phenotyping to record field and greenhouse observations. F1000Research. 2022 Nov 28;11:12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9813448/

52. W3C. ODRL. Available from: https://www.w3.org/TR/odrl-model/

53. W3C. Shapes Constraint Language (SHACL). 2017. Available from: https://www.w3.org/TR/shacl/

Supplementary material

A MAP concepts into ontologies

All the mapping can find in this GitHub repository: "*FAIR_GermplasmBank/MAP concepts into ontologies"*.

B Edit OWL ontology in Protégé

The OWL file edited is in this GitHub repository: **"***wilkinsonlab/multi-crop-passport-descriptor-ontology/multi-passport-descriptor.ow"l.*

C YARRRML mapping models

The Python scripts, CSV files and pictures of the semantic models are in this GitHub repository: "*FAIR_Germplasm/Semantic Models"*.

D Standardized CSV file

The Python script is in this GitHub repository: "*FAIR_GermplasmBank/Standardized CSV file"*.