CBGP
CENTRO DE BIOTECNOLOGÍA
Y GENÓMICA DE PLANTAS
UPM-INIA/CSIC

EXCELENCIA
SEVERO
OCHOA

RCN NSF funding: Award Abstract # 2126334

Grant TED2021-130788B-I00 funded by

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Financiado por
la Unión Europea
NextGenerationEU

Plan de Recuperación,
Transformación y
Resiliencia

AGENCIA
ESTATAL DE
INVESTIGACIÓN

FLAIR-GG

Building the infrastructure for a network of FAIR Germplasm Resources

Alberto Cámara

On behalf of the Wilkinson and Moreno Vazquez Laboratories

Centro de Biotecnología y Genómica de Plantas
(CBGP, UPM-INIA/CSIC)
Universidad Politécnica de Madrid

POLITÉCNICA

INIA
Instituto Nacional de Investigación
y Tecnología Agraria y Alimentaria

CSIC
Consejo Superior de Investigaciones Científicas

www.cbgp.upm.es

Introduction

- Banco de germoplasma vegetal César Gómez Campo
  - Founded in 1966
  - First germplasm bank in Spain.
  - First germplasm bank focused on wild crop relatives in the world.

Prof. César Gómez Campo

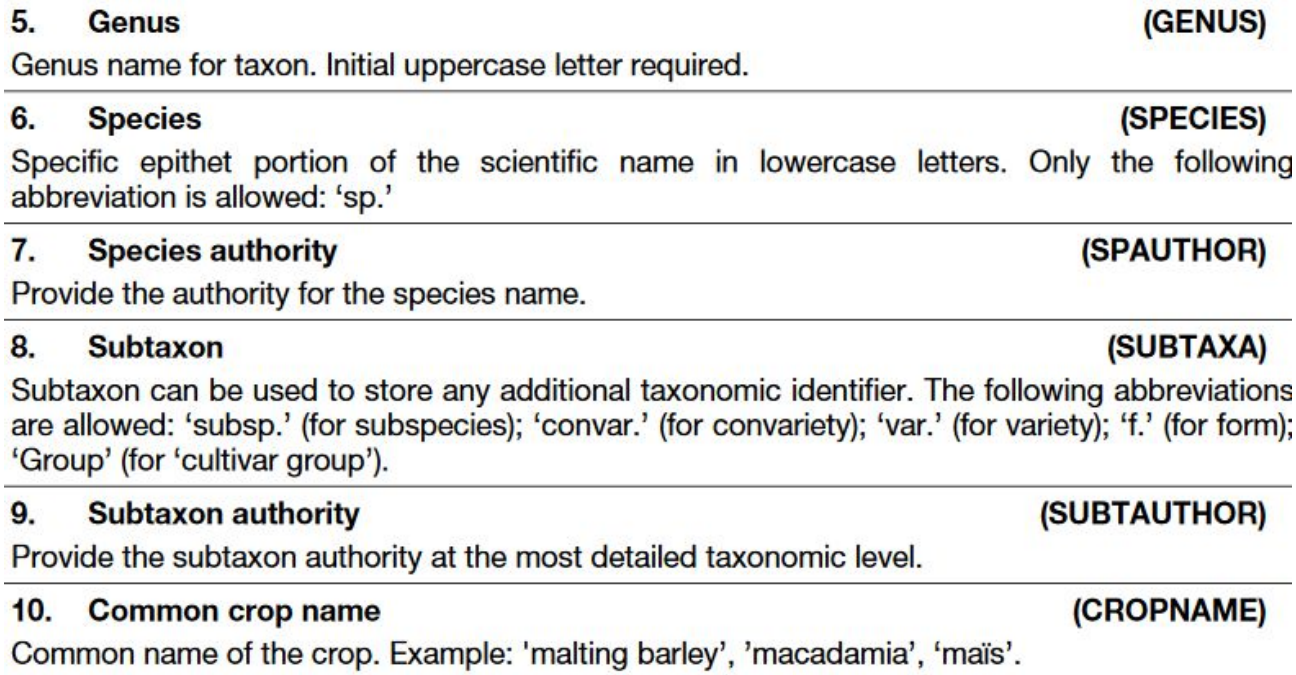
Source for both images: http://www.bancodegermoplasma.upm.es

Germplasm bank’s data

- Plant data: scientific name and authorship, vernacular name, etc.
- Collection data: geolocation, date, soil type, etc.
- Administrative data: collector(s), breeding institution, storage institution, etc.

More than 10.500 accessions, ~4.400 species

Data standardization: FAO multi-crop passport descriptors



5. Genus (GENUS)

Genus name for taxon. Initial uppercase letter required.

6. Species (SPECIES)

Specific epithet portion of the scientific name in lowercase letters. Only the following abbreviation is allowed: ‘sp.’

7. Species authority (SPAUTHOR)

Provide the authority for the species name.

8. Subtaxon (SUBTAXA)

Subtaxon can be used to store any additional taxonomic identifier. The following abbreviations are allowed: ‘subsp.’ (for subspecies); ‘convar.’ (for convariety); ‘var.’ (for variety); ‘f.’ (for form); ‘Group’ (for ‘cultivar group’).

9. Subtaxon authority (SUBTAUTHOR)

Provide the subtaxon authority at the most detailed taxonomic level.

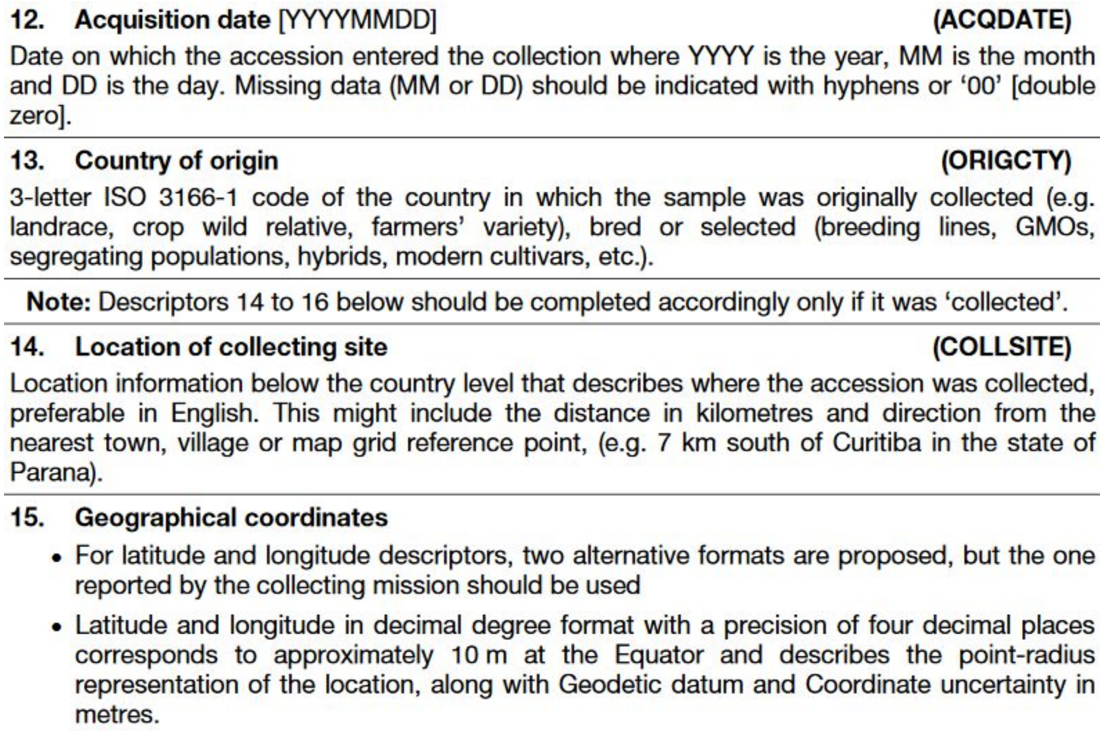
10. Common crop name (CROPNAME)

Common name of the crop. Example: 'malting barley', ’macadamia’, ‘maïs’.

Plant data

Source: https://www.fao.org/plant-treaty/tools/toolbox-for-sustainable-use/details/en/c/1367915/

Data standardization: FAO multi-crop passport descriptors



12. Acquisition date [YYYYMMDD] **(ACQDATE)**

Date on which the accession entered the collection where YYYY is the year, MM is the month and DD is the day. Missing data (MM or DD) should be indicated with hyphens or '00' [double zero].

13. Country of origin **(ORIGCTY)**

3-letter ISO 3166-1 code of the country in which the sample was originally collected (e.g. landrace, crop wild relative, farmers' variety), bred or selected (breeding lines, GMOs, segregating populations, hybrids, modern cultivars, etc.).

Note: Descriptors 14 to 16 below should be completed accordingly only if it was 'collected'.

14. Location of collecting site **(COLLSITE)**

Location information below the country level that describes where the accession was collected, preferable in English. This might include the distance in kilometres and direction from the nearest town, village or map grid reference point, (e.g. 7 km south of Curitiba in the state of Parana).

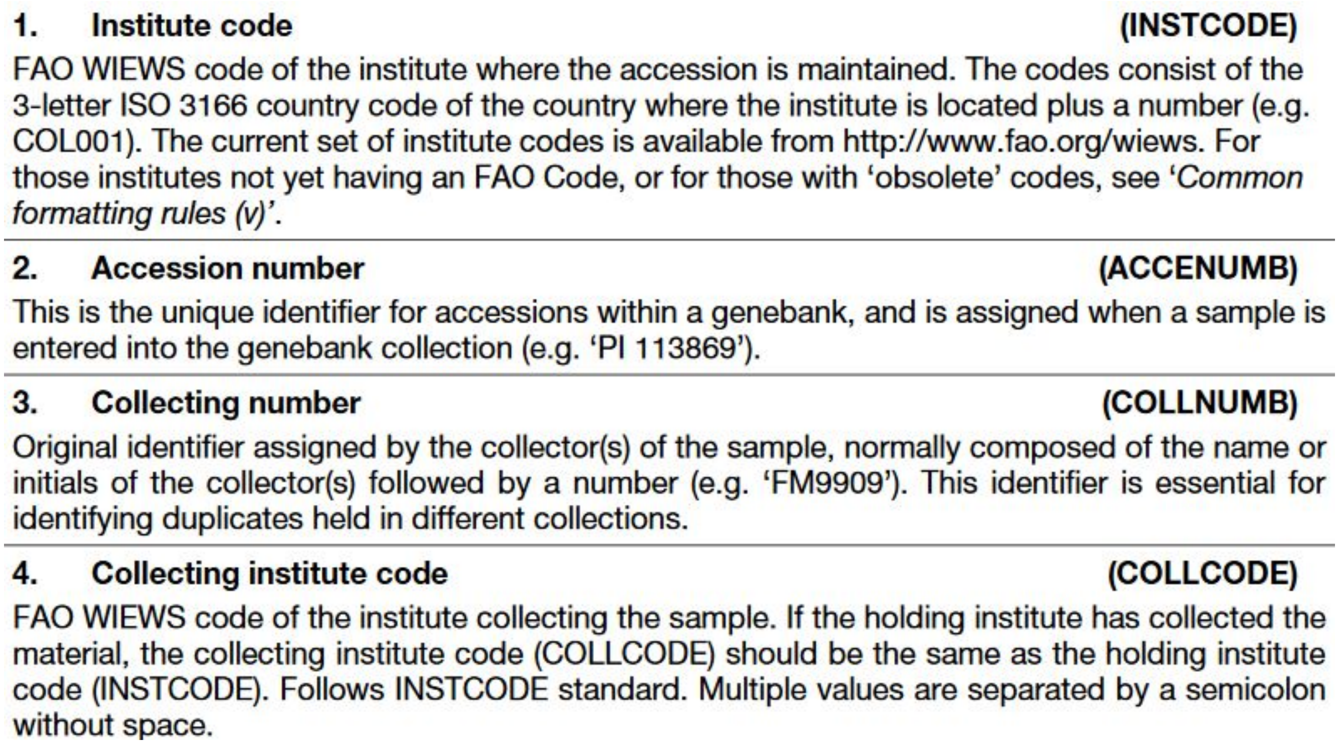
15. Geographical coordinates

- For latitude and longitude descriptors, two alternative formats are proposed, but the one reported by the collecting mission should be used
- Latitude and longitude in decimal degree format with a precision of four decimal places corresponds to approximately 10 m at the Equator and describes the point-radius representation of the location, along with Geodetic datum and Coordinate uncertainty in metres.

Collection data

Source: https://www.fao.org/plant-treaty/tools/toolbox-for-sustainable-use/details/en/c/1367915/

Data standardization: FAO multi-crop passport descriptors



1. Institute code (INSTCODE)

FAO WIEWS code of the institute where the accession is maintained. The codes consist of the 3-letter ISO 3166 country code of the country where the institute is located plus a number (e.g. COL001). The current set of institute codes is available from http://www.fao.org/wiews. For those institutes not yet having an FAO Code, or for those with 'obsolete' codes, see '*Common formatting rules (v)*'.

2. Accession number (ACCENUMB)

This is the unique identifier for accessions within a genebank, and is assigned when a sample is entered into the genebank collection (e.g. 'PI 113869').

3. Collecting number (COLLNUMB)

Original identifier assigned by the collector(s) of the sample, normally composed of the name or initials of the collector(s) followed by a number (e.g. 'FM9909'). This identifier is essential for identifying duplicates held in different collections.

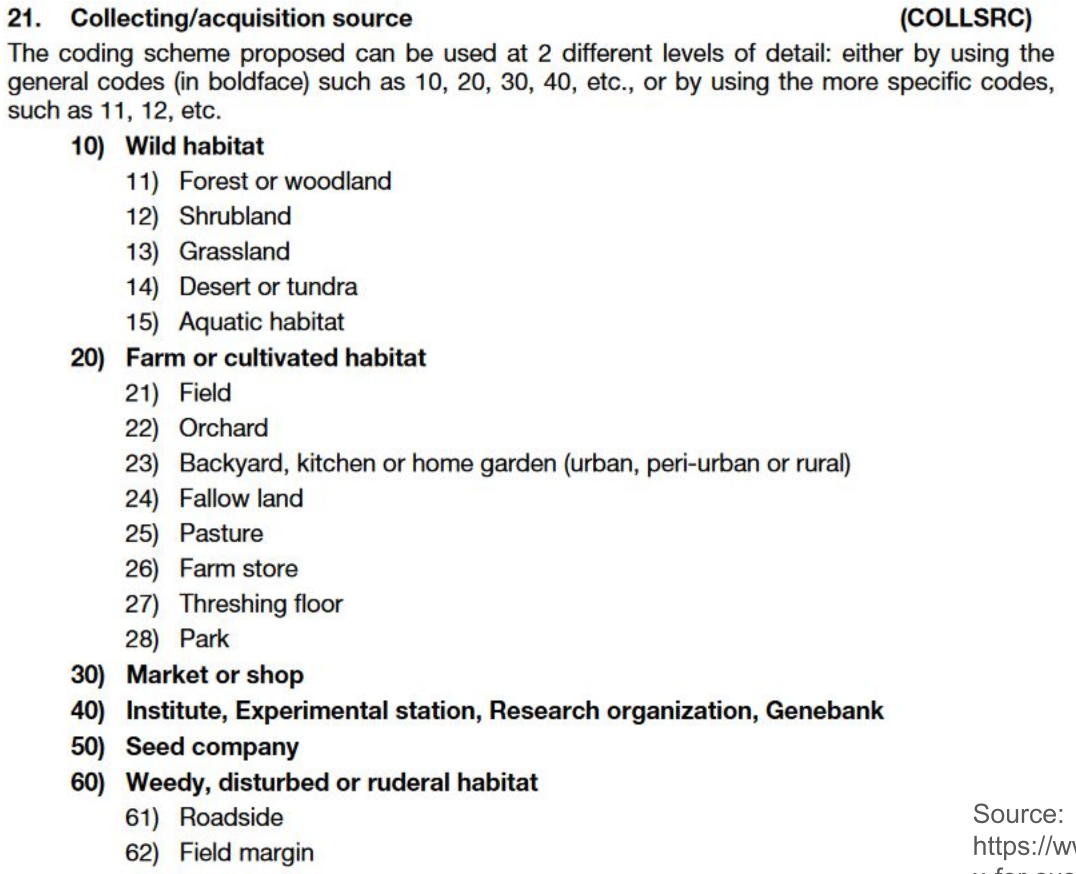
4. Collecting institute code (COLLCODE)

FAO WIEWS code of the institute collecting the sample. If the holding institute has collected the material, the collecting institute code (COLLCODE) should be the same as the holding institute code (INSTCODE). Follows INSTCODE standard. Multiple values are separated by a semicolon without space.

Administrative data

Source:
https://www.fao.org/plant-treaty/tools/toolbox-for-sustainable-use/details/en/c/1367915/

Data standardization: FAO multi-crop passport descriptors



21. Collecting/acquisition source (COLLSRC)

The coding scheme proposed can be used at 2 different levels of detail: either by using the general codes (in boldface) such as 10, 20, 30, 40, etc., or by using the more specific codes, such as 11, 12, etc.

10) Wild habitat
- 11) Forest or woodland
- 12) Shrubland
- 13) Grassland
- 14) Desert or tundra
- 15) Aquatic habitat

20) Farm or cultivated habitat
- 21) Field
- 22) Orchard
- 23) Backyard, kitchen or home garden (urban, peri-urban or rural)
- 24) Fallow land
- 25) Pasture
- 26) Farm store
- 27) Threshing floor
- 28) Park

30) Market or shop

40) Institute, Experimental station, Research organization, Genebank

50) Seed company

60) Weedy, disturbed or ruderal habitat
- 61) Roadside
- 62) Field margin

Source: https://www.fao.org/plant-treaty/tools/toolbox-for-sustainable-use/details/en/c/1367915/

Resource Description Framework (RDF)

Like human language, RDF statements take the form:

Subject	Predicate	Object
Alberto	likes	pasta

These are known as “triples”.

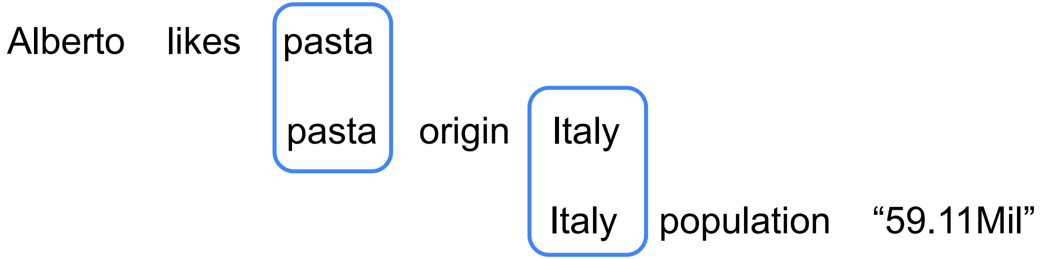
Resource Description Framework (RDF)

The Object of one triple becomes the Subject of another:

S	P	O		
Alberto	likes	pasta		
		S	P	O
		pasta	origin	Italy

Resource Description Framework (RDF)

The Object of one triple becomes the Subject of another:



Alberto likes pasta

pasta origin Italy

Italy population “59.11Mil”

Allowing for the representation of complex concepts, creating what is known as Linked Data.

Resource Description Framework (RDF)

A triplet that is closer to reality:

this:A123 dwc:scientificName ‘Arabidopsis thaliana’

Signifying that accesion number A123’s scientific name is Arabidopsis thaliana.

Resource Description Framework (RDF)

A triplet that is closer to reality:

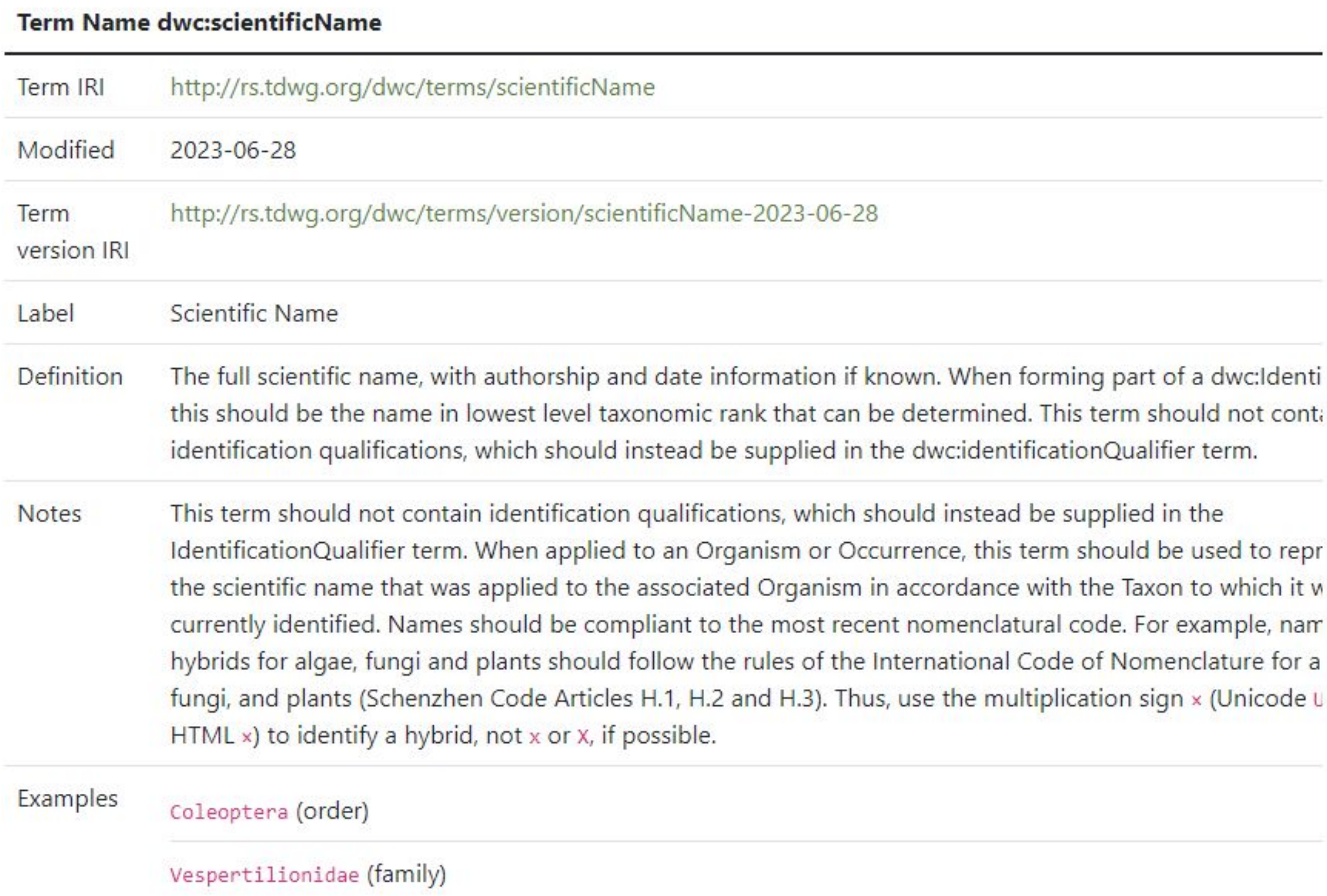
this:A123 dwc:scientificName ‘Arabidopsis thaliana’

Signifying that accesion number A123’s scientific name is Arabidopsis thaliana.

From the last slide to this one I’ve added ontologies.

this: https://my.exampleurl.com/
dwc: https://dwc.tdwg.org/list/#

Ontologies!



Term Name dwc:scientificName	
Term IRI	http://rs.tdwg.org/dwc/terms/scientificName
Modified	2023-06-28
Term version IRI	http://rs.tdwg.org/dwc/terms/version/scientificName-2023-06-28
Label	Scientific Name
Definition	The full scientific name, with authorship and date information if known. When forming part of a dwc:Identi this should be the name in lowest level taxonomic rank that can be determined. This term should not cont identification qualifications, which should instead be supplied in the dwc:identificationQualifier term.
Notes	This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term. When applied to an Organism or Occurrence, this term should be used to repr the scientific name that was applied to the associated Organism in accordance with the Taxon to which it w currently identified. Names should be compliant to the most recent nomenclatural code. For example, nam hybrids for algae, fungi and plants should follow the rules of the International Code of Nomenclature for a fungi, and plants (Schenzhen Code Articles H.1, H.2 and H.3). Thus, use the multiplication sign × (Unicode U HTML ×) to identify a hybrid, not x or X, if possible.
Examples	`Coleoptera` (order) `Vespertilionidae` (family)

~Half of the FAIR Principles are addressed by RDF!

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

RDF generally requires all entities to have a URL, therefore, everything has a globally unique identifier

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

URLs all use the Web as a mechanism for retrieval of the data they identify. The Web (HTTP Protocol) is open, free, and universally implementable, and allows for authentication.

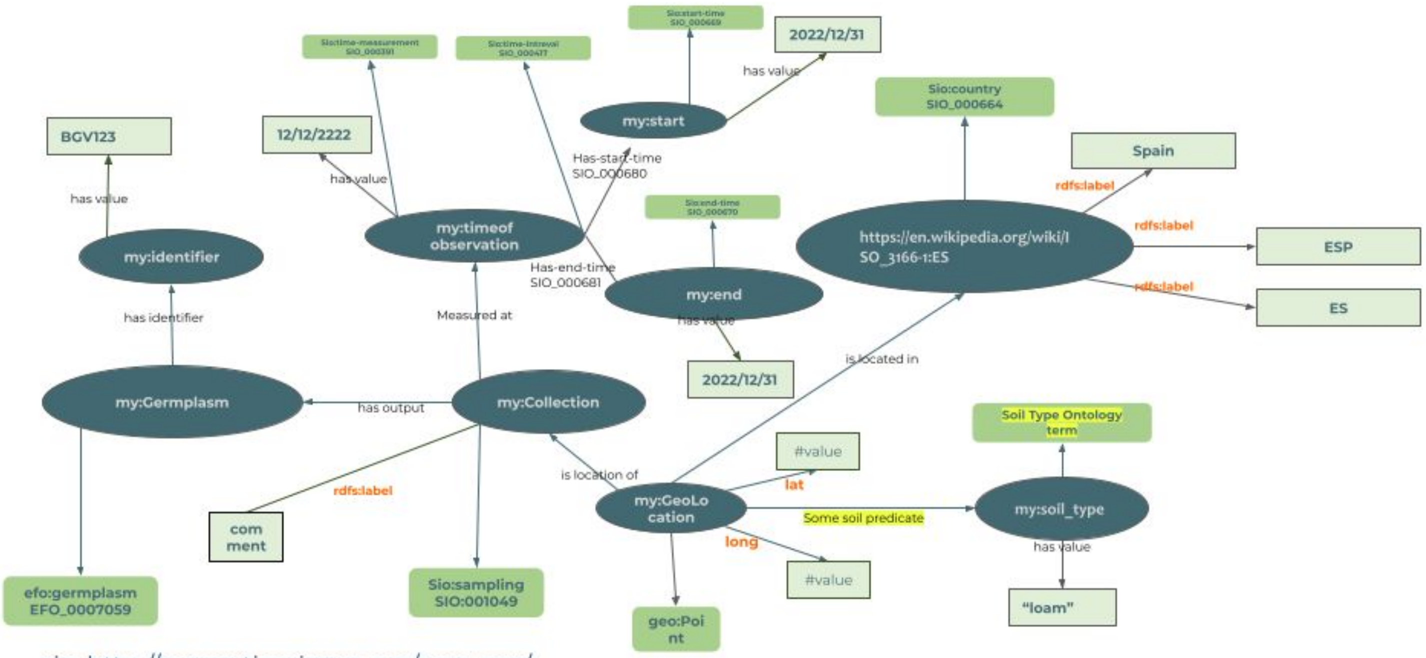
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

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

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

RDF was invented to be a formal, broadly applicable language for knowledge representation, and encourages the use of shared formal vocabularies to create qualified references.

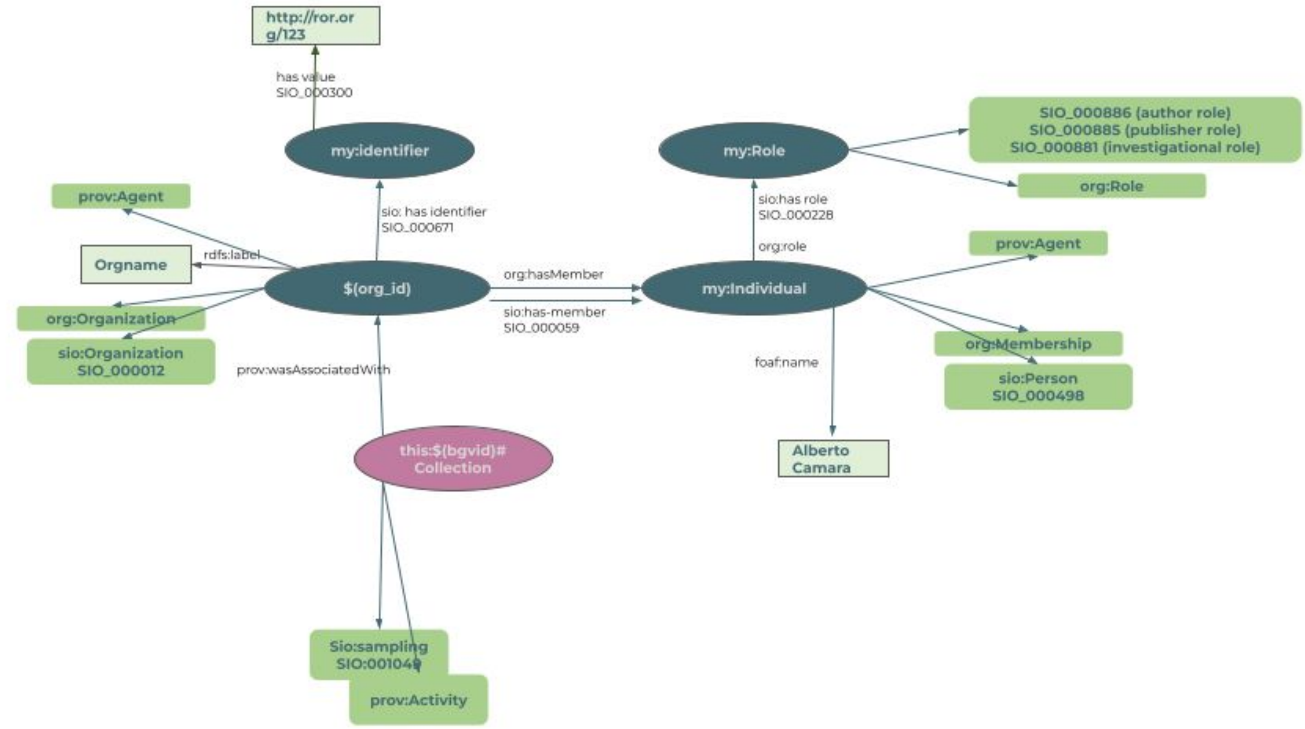
Semantic Models: Collection data



sio: http://semanticscience.org/resource/
geo: http://www.w3.org/2003/01/geo/wgs84_pos#
Efo: http://www.ebi.ac.uk/efo/

Source :https://github.com/wilkinsonlab/FLAIR-GG/tree/main/SemanticModel

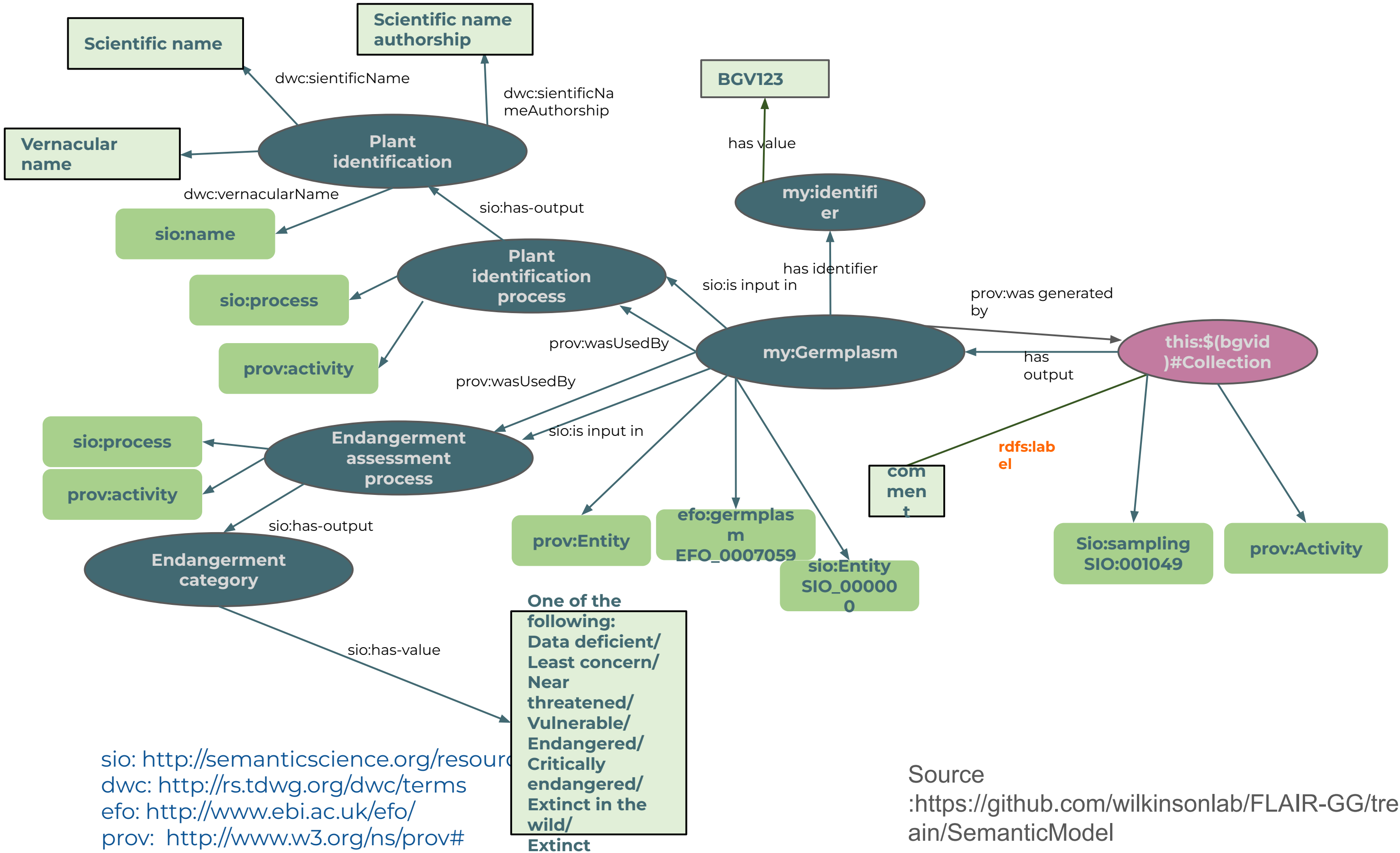
Semantic Models: Administrative information



sio: http://semanticscience.org/resource/
org: http://www.w3.org/ns/org#
foaf: http://xmlns.com/foaf/0.1/

Source :https://github.com/wilkinsonlab/FLAIR-GG/tree/main/SemanticModel

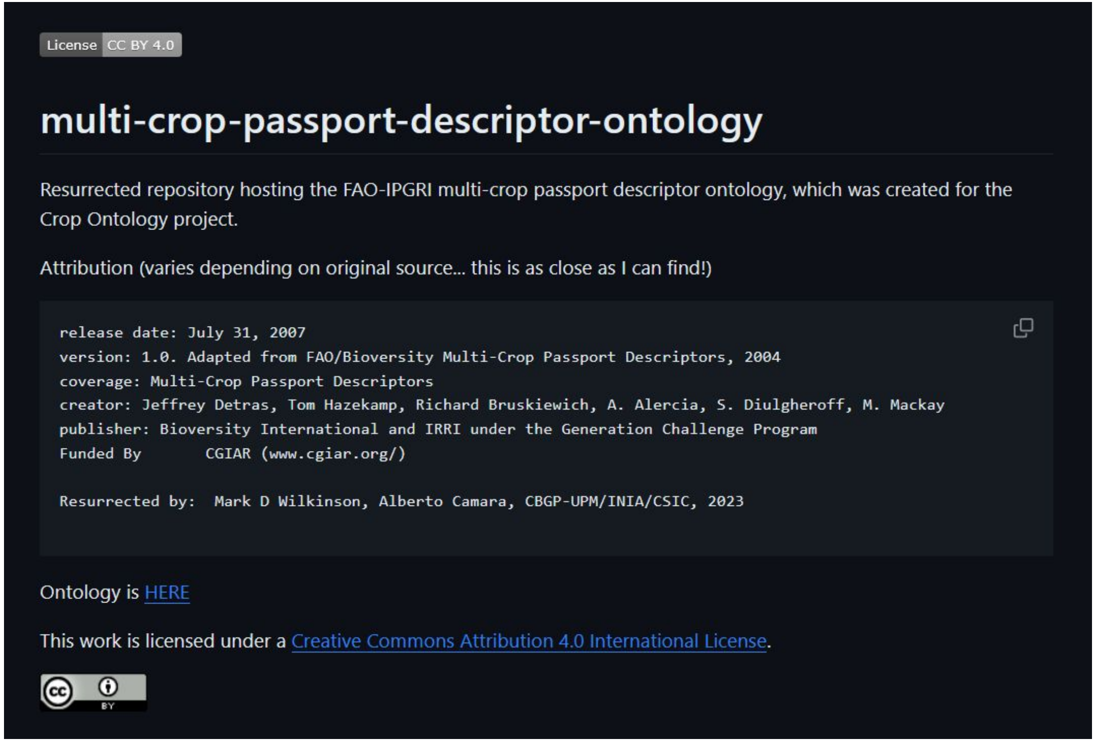
Semantic models: germplasm data



sio: http://semanticscience.org/resourc
dwc: http://rs.tdwg.org/dwc/terms
efo: http://www.ebi.ac.uk/efo/
prov: http://www.w3.org/ns/prov#

Source :https://github.com/wilkinsonlab/FLAIR-GG/tree/main/SemanticModel

FAO’s multi-crop passport ontology



License CC BY 4.0

multi-crop-passport-descriptor-ontology

Resurrected repository hosting the FAO-IPGRI multi-crop passport descriptor ontology, which was created for the Crop Ontology project.

Attribution (varies depending on original source... this is as close as I can find!)

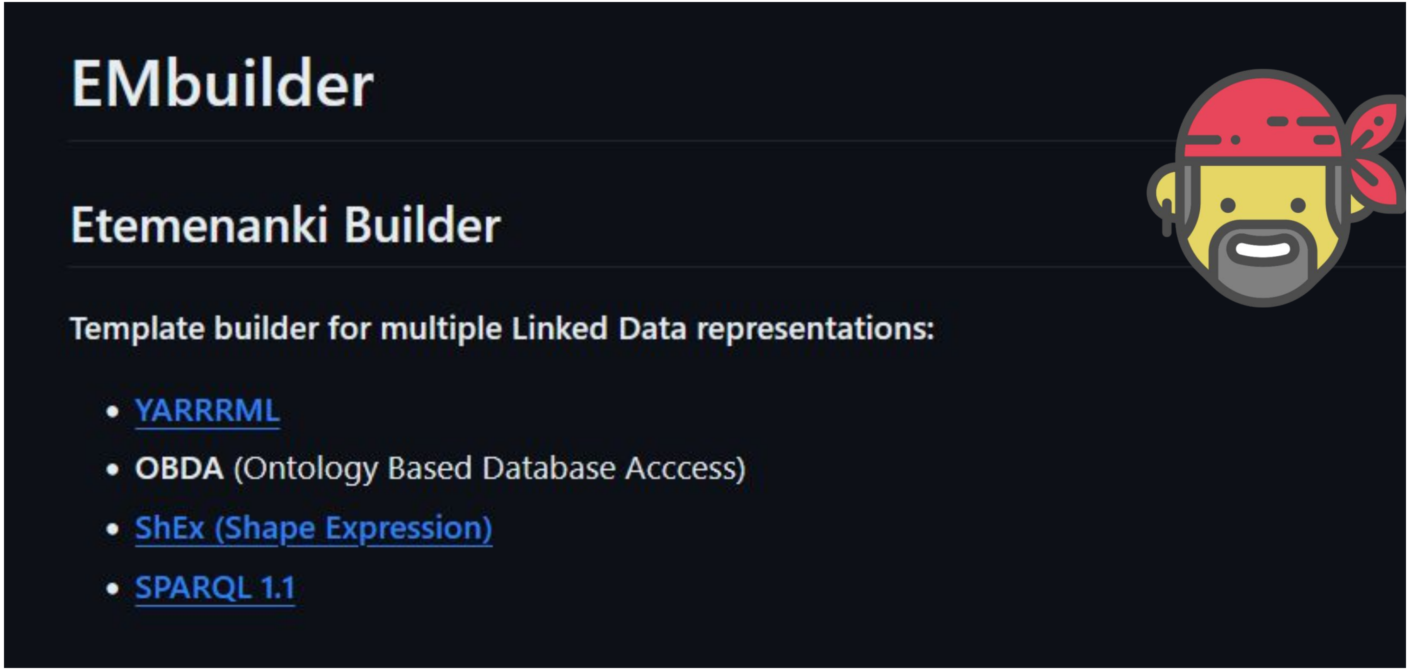
```
release date: July 31, 2007
version: 1.0. Adapted from FAO/Bioversity Multi-Crop Passport Descriptors, 2004
coverage: Multi-Crop Passport Descriptors
creator: Jeffrey Detras, Tom Hazekamp, Richard Bruskiewich, A. Alercia, S. Diulgheroff, M. Mackay
publisher: Bioversity International and IRRI under the Generation Challenge Program
Funded By       CGIAR (www.cgiar.org/)

Resurrected by:  Mark D Wilkinson, Alberto Camara, CBGP-UPM/INIA/CSIC, 2023
```

Ontology is HERE

This work is licensed under a Creative Commons Attribution 4.0 International License.

Source: https://github.com/wilkinsonlab/multi-crop-passport-descriptor-ontology

YARRRML transformation



EMbuilder

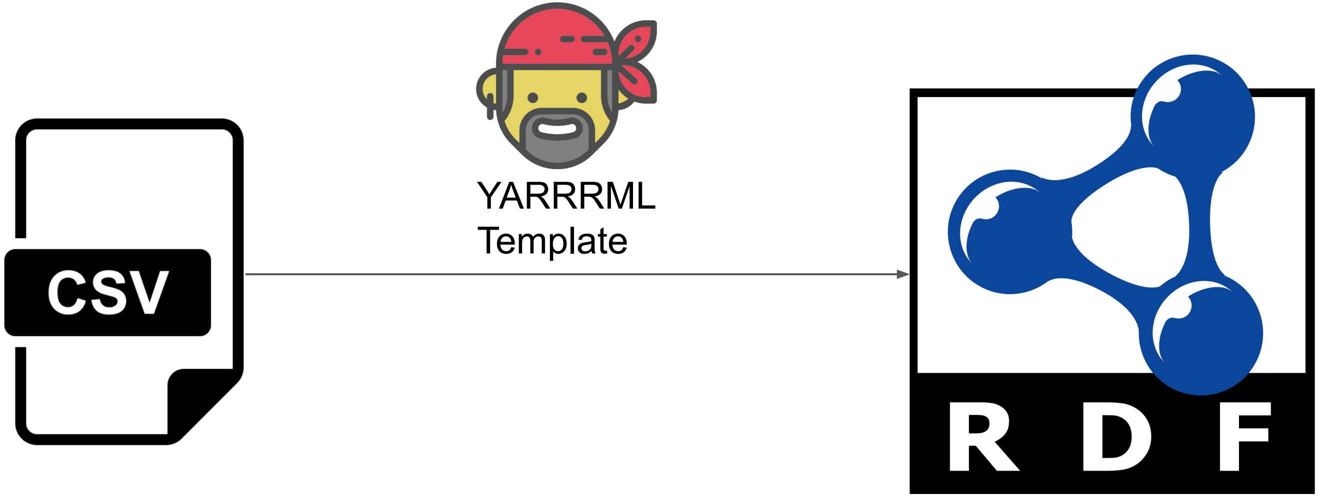
Etemenanki Builder

Template builder for multiple Linked Data representations:

- YARRRML
- **OBDA** (Ontology Based Database Acccess)
- ShEx (Shape Expression)
- SPARQL 1.1

Built by: Pablo Alarcón Moreno
https://github.com/pabloalarconm/EMbuilder

(Simplified) YARRRML Transformation pipeline



YARRRML Template

Sources:
https://icon-icons.com/icon/file-csv-format-type/134696
https://rml.io/yarrrml/
https://es.m.wikipedia.org/wiki/Archivo:Rdf_logo.svg

(Simplified) YARRRML Transformation pipeline

YARRRML Template

```
["this:$(uniqid)#Plant_identification", "dwc:sientificName", "$(Scientific_name)", "iri"],
["this:$(uniqid)#Plant_identification", "dwc:sientificNameAuthorship", "$(Scientific_name_authorship)", "iri"],
```

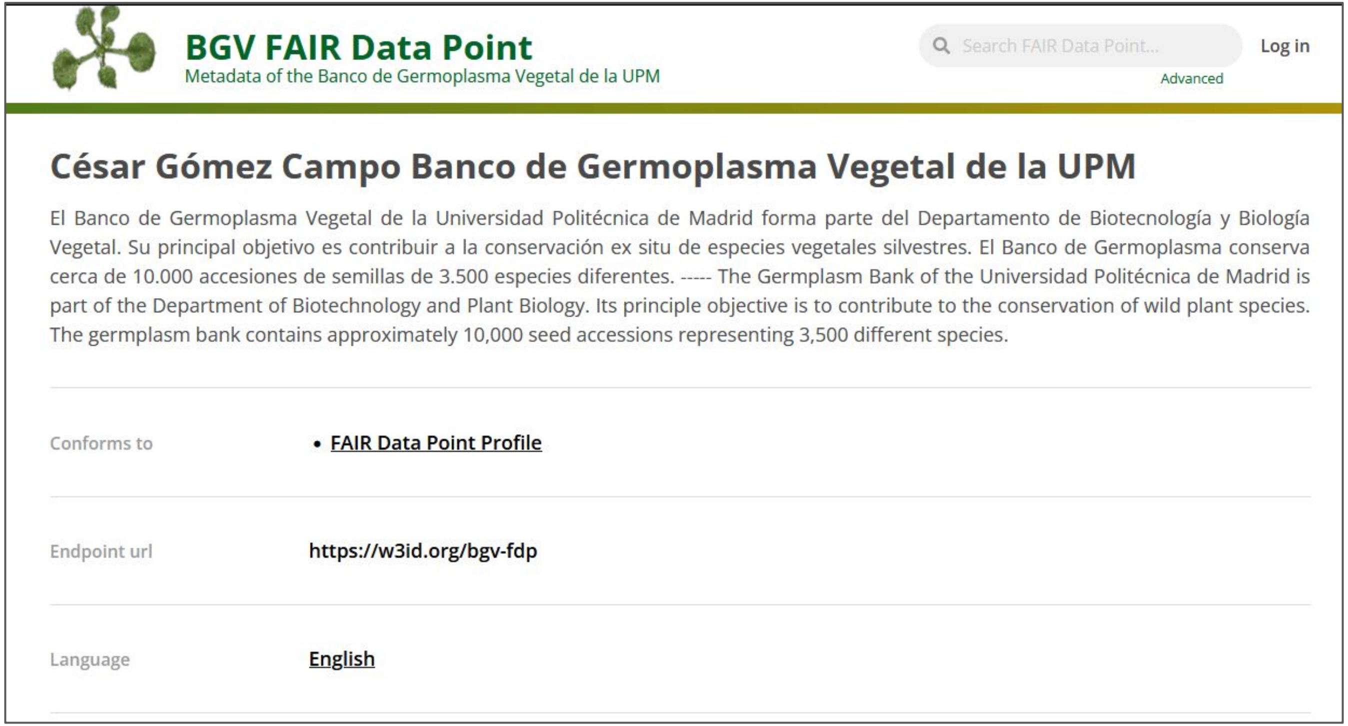
The whole transformation pipeline exists as a layer on top of your pre-existing database!

Sources:
https://rml.io/yarrrml/

CBGP CENTRO DE BIOTECNOLOGÍA Y GENÓMICA DE PLANTAS UPM-INIA/CSIC

FAIR Data Point (FDP)

- Metadata record of the germplasm database
- Follows the Data Catalog (DCAT) ontological model
- Provides a REST interface and a Web interface for building DCAT records



BGV FAIR Data Point

Metadata of the Banco de Germoplasma Vegetal de la UPM

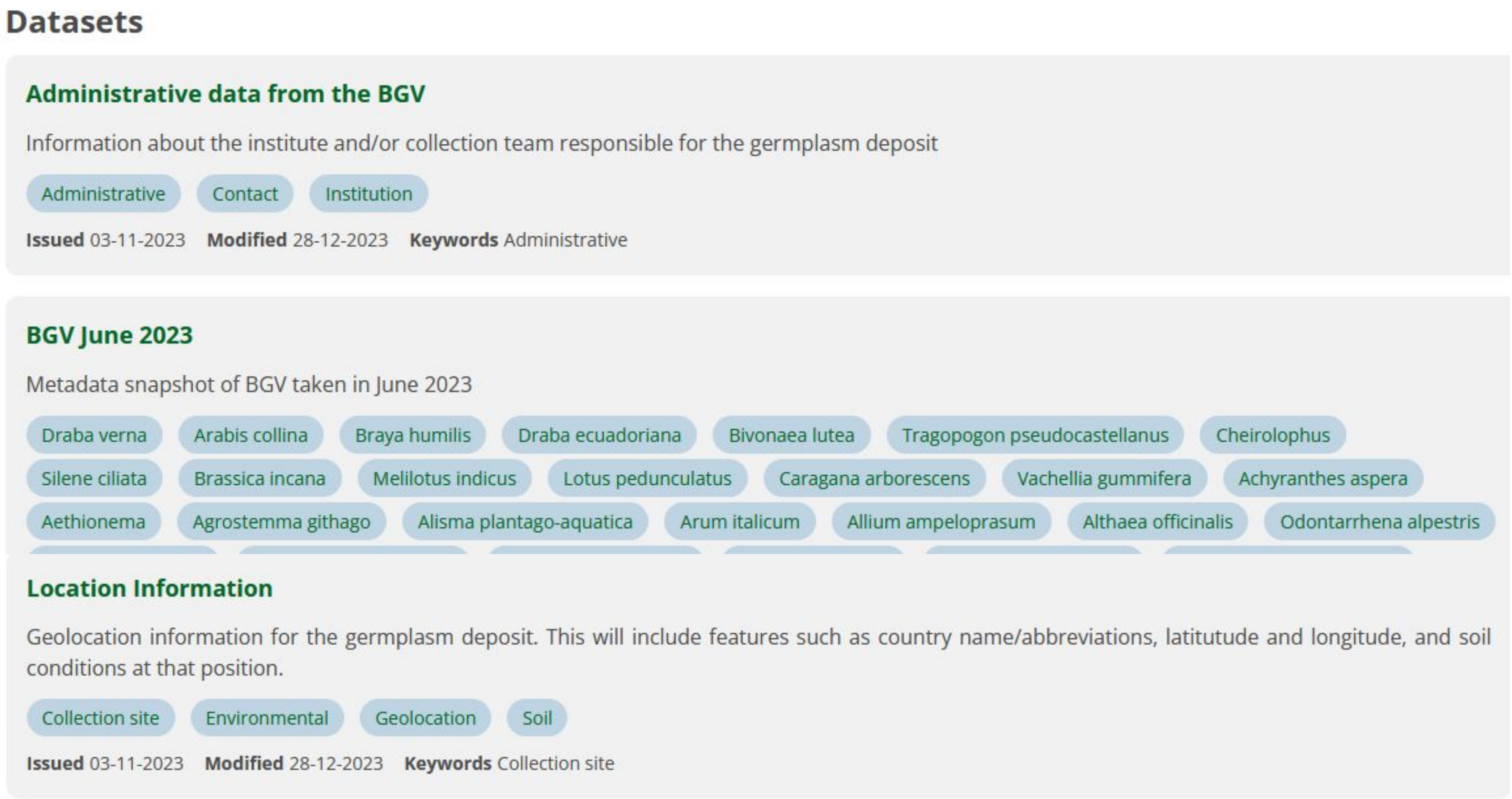
César Gómez Campo Banco de Germoplasma Vegetal de la UPM

El Banco de Germoplasma Vegetal de la Universidad Politécnica de Madrid forma parte del Departamento de Biotecnología y Biología Vegetal. Su principal objetivo es contribuir a la conservación ex situ de especies vegetales silvestres. El Banco de Germoplasma conserva cerca de 10.000 accesiones de semillas de 3.500 especies diferentes. ----- The Germplasm Bank of the Universidad Politécnica de Madrid is part of the Department of Biotechnology and Plant Biology. Its principle objective is to contribute to the conservation of wild plant species. The germplasm bank contains approximately 10,000 seed accessions representing 3,500 different species.

Conforms to	• FAIR Data Point Profile
Endpoint url	https://w3id.org/bgv-fdp
Language	English

Source: https://fdp.bgv.cbgp.upm.es/

FDP dcat:Dataset records



Datasets

Administrative data from the BGV

Information about the institute and/or collection team responsible for the germplasm deposit

Administrative · Contact · Institution

Issued 03-11-2023 **Modified** 28-12-2023 **Keywords** Administrative

BGV June 2023

Metadata snapshot of BGV taken in June 2023

Draba verna · Arabis collina · Braya humilis · Draba ecuadoriana · Bivonaea lutea · Tragopogon pseudocastellanus · Cheirolophus · Silene ciliata · Brassica incana · Melilotus indicus · Lotus pedunculatus · Caragana arborescens · Vachellia gummifera · Achyranthes aspera · Aethionema · Agrostemma githago · Alisma plantago-aquatica · Arum italicum · Allium ampeloprasum · Althaea officinalis · Odontarrhena alpestris

Location Information

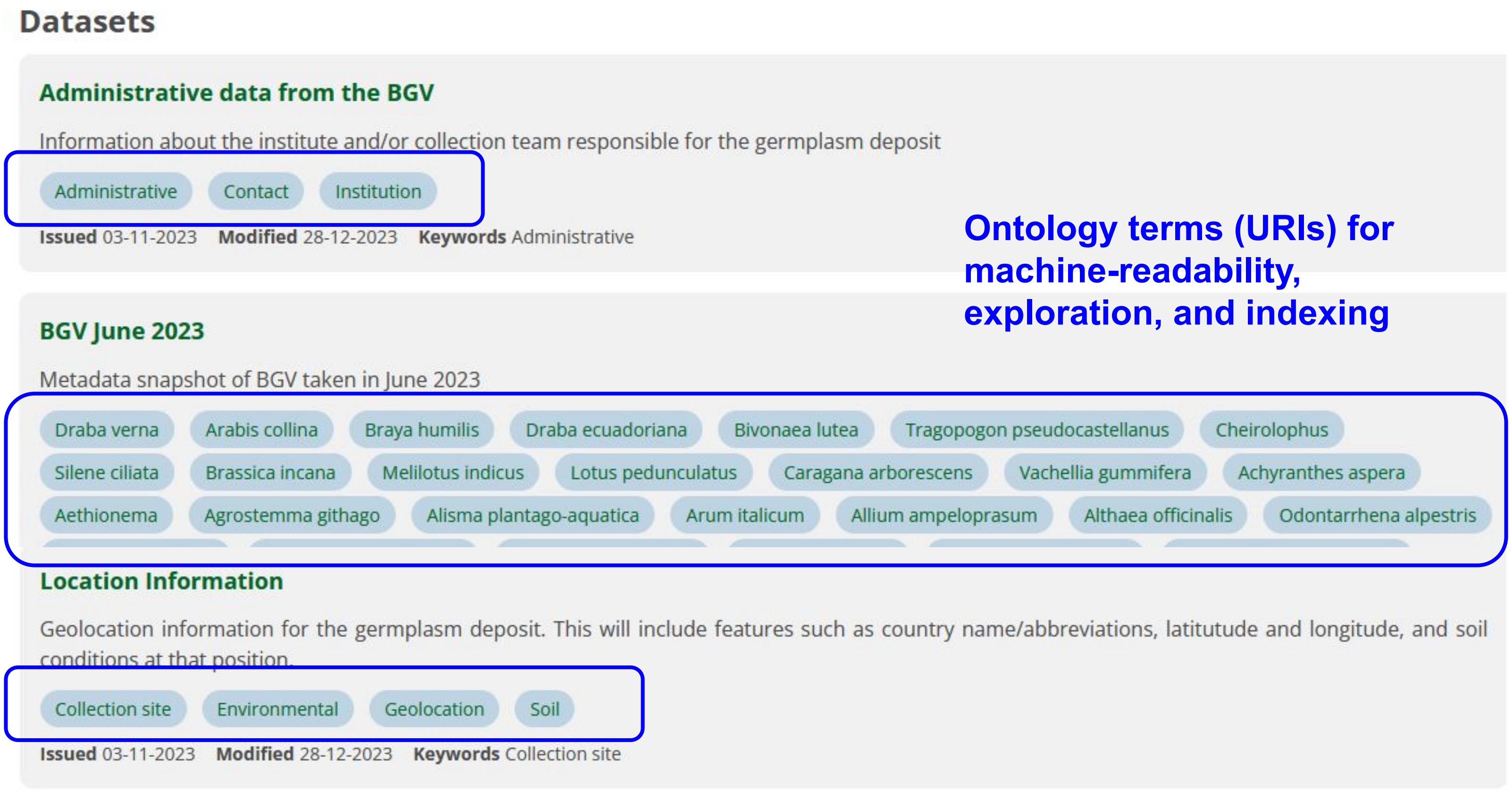
Geolocation information for the germplasm deposit. This will include features such as country name/abbreviations, latitutude and longitude, and soil conditions at that position.

Collection site · Environmental · Geolocation · Soil

Issued 03-11-2023 **Modified** 28-12-2023 **Keywords** Collection site

Source: https://fdp.bgv.cbgp.upm.es/catalog/3e699f66-6b8a-4c6a-9d06-d8685718cc33

FDP dcat:Dataset records



Datasets

Administrative data from the BGV

Information about the institute and/or collection team responsible for the germplasm deposit

Administrative · Contact · Institution

Issued 03-11-2023 **Modified** 28-12-2023 **Keywords** Administrative

BGV June 2023

Metadata snapshot of BGV taken in June 2023

Draba verna · Arabis collina · Braya humilis · Draba ecuadoriana · Bivonaea lutea · Tragopogon pseudocastellanus · Cheirolophus · Silene ciliata · Brassica incana · Melilotus indicus · Lotus pedunculatus · Caragana arborescens · Vachellia gummifera · Achyranthes aspera · Aethionema · Agrostemma githago · Alisma plantago-aquatica · Arum italicum · Allium ampeloprasum · Althaea officinalis · Odontarrhena alpestris

Location Information

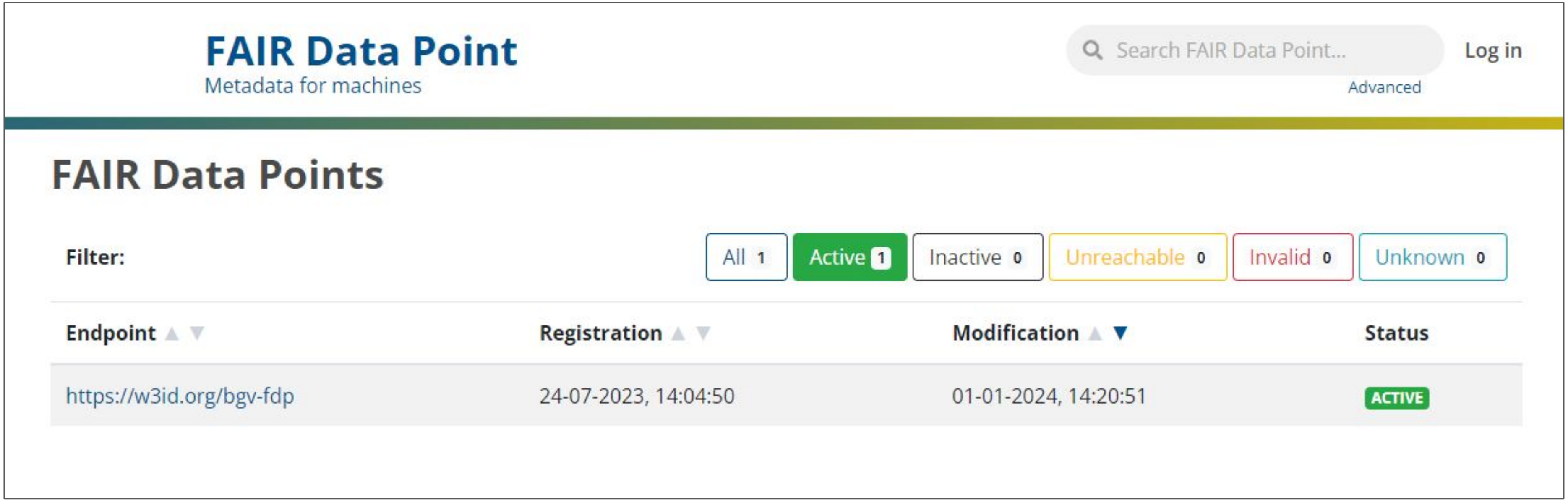
Geolocation information for the germplasm deposit. This will include features such as country name/abbreviations, latitutude and longitude, and soil conditions at that position.

Collection site · Environmental · Geolocation · Soil

Issued 03-11-2023 **Modified** 28-12-2023 **Keywords** Collection site

Ontology terms (URIs) for machine-readability, exploration, and indexing

FAIR Data Point Index



FAIR Data Points

Filter: All 1 | Active 1 | Inactive 0 | Unreachable 0 | Invalid 0 | Unknown 0

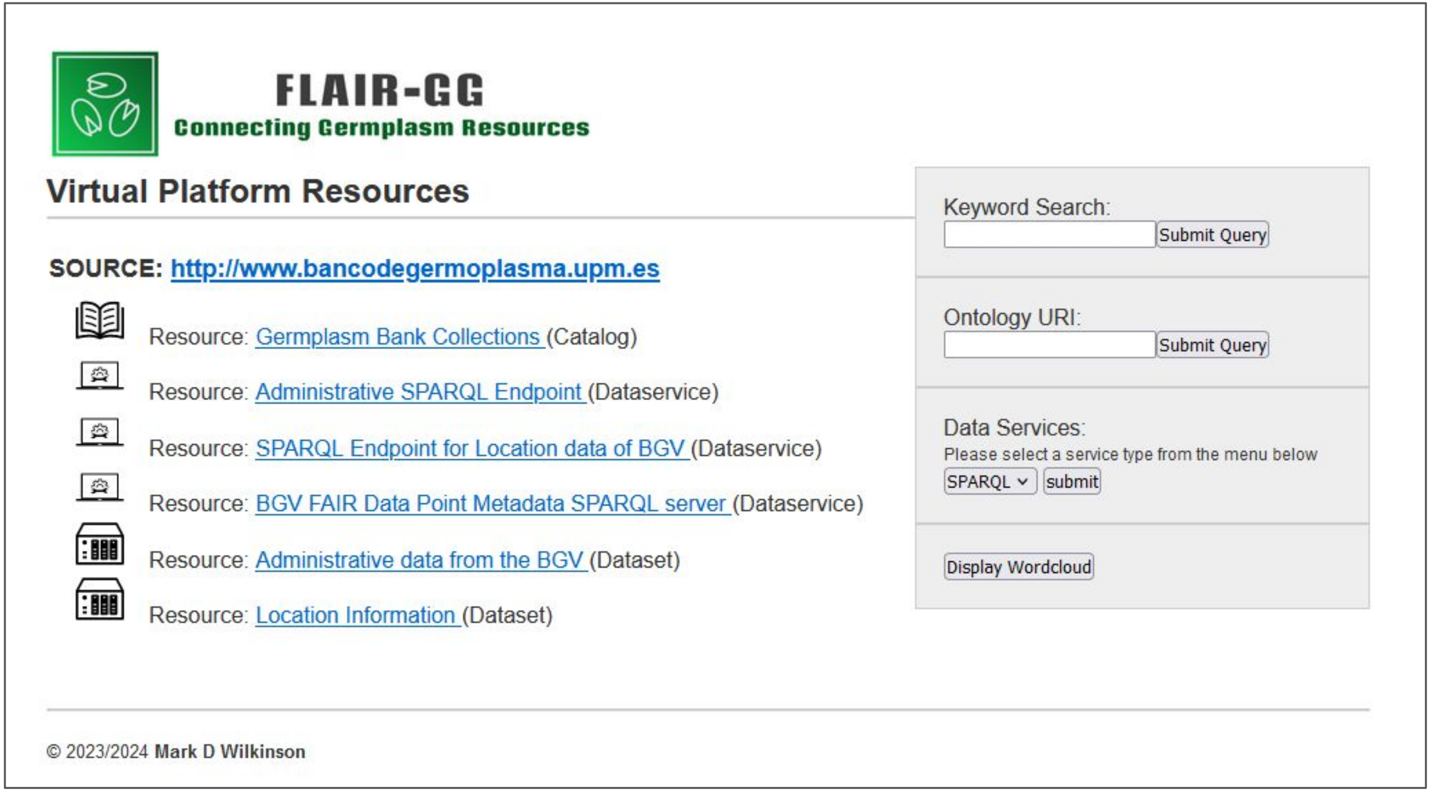
Endpoint	Registration	Modification	Status
https://w3id.org/bgv-fdp	24-07-2023, 14:04:50	01-01-2024, 14:20:51	ACTIVE

A record of all participants in the FLAIR-GG Network

(currently only us… but soon we will grow!)

FLAIR-GG “Virtual Platform”

A place to do federated exploration over the entire network of participants



FLAIR-GG

Connecting Germplasm Resources

Virtual Platform Resources

SOURCE: http://www.bancodegermoplasma.upm.es

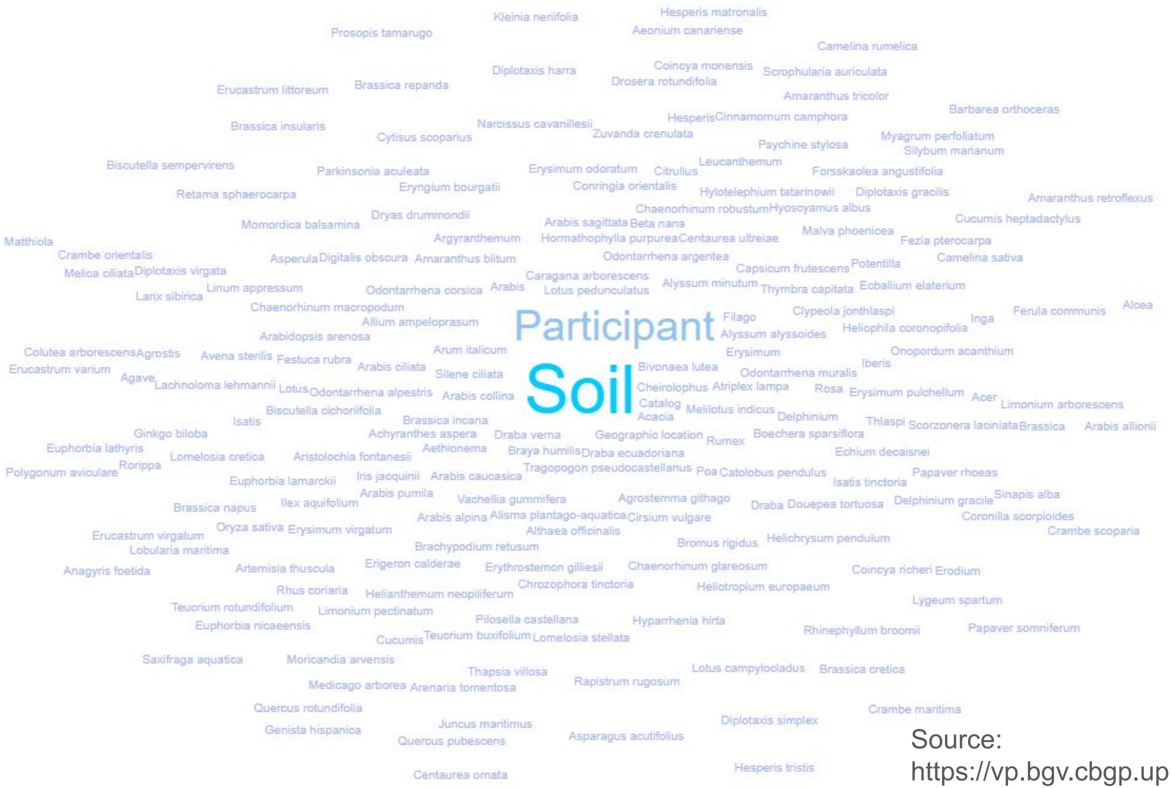
- Resource: Germplasm Bank Collections (Catalog)
- Resource: Administrative SPARQL Endpoint (Dataservice)
- Resource: SPARQL Endpoint for Location data of BGV (Dataservice)
- Resource: BGV FAIR Data Point Metadata SPARQL server (Dataservice)
- Resource: Administrative data from the BGV (Dataset)
- Resource: Location Information (Dataset)

© 2023/2024 **Mark D Wilkinson**

Source:https://vp.bgv.cbgp.upm.es/flair-gg-vp-server/resources

Network at-a-glance: The FLAIR-GG Word Cloud

FDP metadata ontology terms and keywords are automatically harvested from all network participants, and weighted by frequency in the network



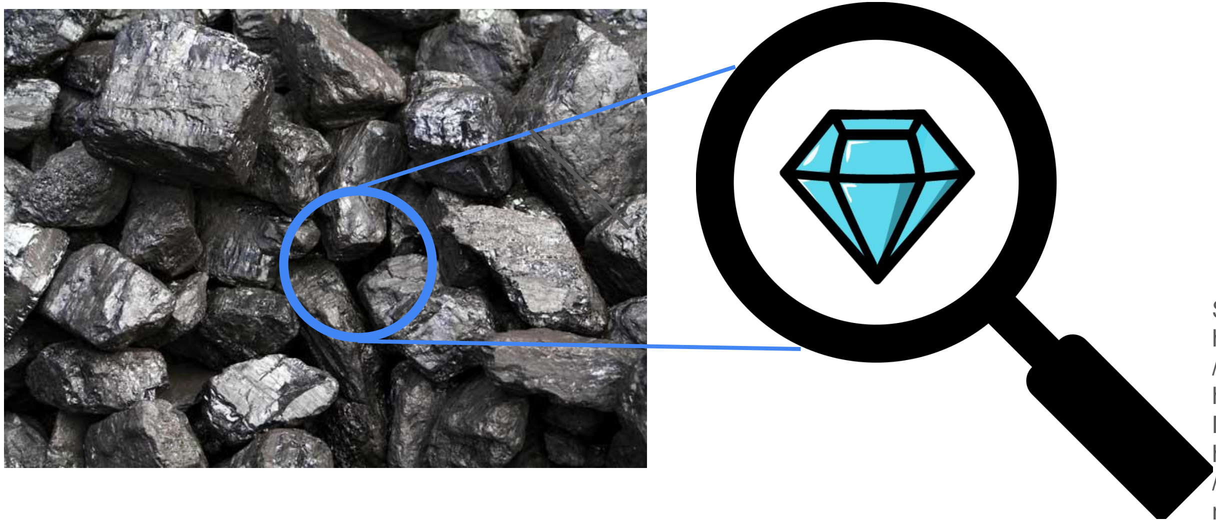
Source:
https://vp.bgv.cbgp.upm.es/flair-gg-vp-server/wordcloud

Benefits of having a network of germplasm resources

1. Finding out the relative value of your germplasm in the context of the whole network.

1. Find all the sources of a particular species.

2. Cross-reference between duplicates.



Sources:
https://commons.wikimedia.org/wiki/File:Liquefaction-charbon.jpg
https://en.m.wikipedia.org/wiki/File:Magnifying_glass_icon.svg
https://commons.wikimedia.org/wiki/File:Diamond_Icon_Transparent.png

You can join the FLAIR-GG Network whenever you want!

The pathway for joining FLAIR-GG:

1) Create a FAIR Data Point metadata record that has certain required metadata facets

2) Inform our FAIR Data Point index that “you exist”

3) It will then automatically index you and ensure that you are “compliant”

4) The Virtual Platform uses the Index to harvest metadata from all participants, so once you are in the Index, you are part of the network!

We are currently working on the documentation for this process, so in the meantime, just email us if you want to join!

alberto.camara-ballesteros@ejprd-project.eu

Future plans

- Open Digital Rights Language (ODRL) representation of international treaties regarding germplasm data.
- Authorization/Authentication.
- Query-endpoint matching.

Acknowledgements

Grant TED2021-130788B-I00 funded by

Wilkinson Lab Team

Alberto Cámara (alberto.camara-ballesteros@ejprd-project.eu)
Pablo Alarcón
Oussama Mohammed Benhamed
Mark Wilkinson

RCN NSF funding: Award Abstract # 2126334

Moreno Vazquez Team

Santiago Moreno Vazquez
German Pastor
Elena Torres

Center for Plant Biotechnology and Genomics, UPM-INIA-CSIC
Severo Ochoa Center of Excellence, Universidad Politécnica de Madrid

Proyectos Estratégicos Orientados a la Transición Ecológica y a la Transición Digital,
Government of Spain, Ministry of Science and Innovation