scientific data

Check for updates

OPEN

COMMENT

The FLAIR-GG federated network of FAIR germplasm data resources

Alberto Cámara Ballesteros[1], Elena Aguayo Jara[2], Evrykleia Sofia Verykaki[3], Germán Pastor del Olmo[3], Santiago Moreno Vázquez[3], Elena Torres[3] & Mark D. Wilkinson[1]✉

A key source of biodiversity preservation is in the *ex situ* storage of seed in what are known as germplasm banks (GBs). Unfortunately, wild species germplasm bank databases, often maintained by resource-limited botanical gardens, are highly disparate and capture information about their collections in a wide range of underlying data formats, storage platforms, following different standards, and with varying degrees of data accessibility. Thus, it is extremely difficult to build conservation strategies for wild species via integrating data from these GBs. Here, we envisage that the application of the FAIR Principles to wild species and crop wild relatives information, through the creation of a federated network of FAIR GB databases, would greatly facilitate cross-resource discovery and exploration, thus assisting with the design of more efficient conservation strategies for wild species, and bringing more attention to these key data providers.

There are a number of unresolved issues related to the availability of information on plant genetic resources (PGR) conserved *ex situ* in Europe[1]. First, it is not clear how PGR-holders maintain the germplasm and to what extent materials are made available to users. Most banks are not aware of the extent to which their accessions are important, are unique or, on the contrary, are perhaps heavily duplicated in other banks within their country or more broadly. The difficulty in gathering information from peer seed banks makes it difficult to develop collaborative projects between banks. The situation is more dire when considering the specific case of wild germplasm conserved *ex situ*. Quite often, diversity among or within species conserved *ex situ* is not well known, and environmental information about the collecting sites is scarce or not available[2]. Similarly, integrating *ex situ* data with data on *in situ* conserved germplasm is even more problematic. Finally, beyond problems with the discovery and retrieval of these data, tooling for their integration, and support for analytics is scarce. Unfortunately, existing efforts to support wild species germplasm collections are not fully addressing these limitations.

The worldwide information systems for plant genetic resource for food and agriculture conserved *ex situ*, Genesys PGR (https://www.genesys-pgr.org/), and its European partner EURISCO (https://eurisco.ipk-gatersleben.de/apex/eurisco_ws/r/eurisco/home) include in their databases elite genotypes, breeding lines, landraces of common crops and neglected and underused crops (NUS), crop wild relatives (CWR) and wild food plants (WFP). In Europe, there are around 400 PGR-holders from 43 countries informing EURISCO and accounting for about 2 million accessions[3]. Because CWR and WFP have been recognized as cornerstone of food security under climate change scenarios, in 2022 EURISCO decided also to integrate information on CWR and WFP conserved *in situ*[4]. *Ex situ* and *in situ* native wild germplasm, including CWR and WFP, have also been monitored and documented for conservation purposes mainly by seed banks within botanical gardens having diverse origin, structures and funding sources[5]. Since 2011, the European Native Seed Conservation Consortium (ENSCONET) coordinates and integrates European seed conservation practice, policy and research for native plant species[6]. ENSCONET has attempted to organize dispersed information from botanical gardens and other institutions through a central data platform, ENSCOBASE. ENSCOBASE contains entries from 35 seed banks

[1]Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas (CBGP). Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC). Pozuelo de Alarcón (Madrid), Madrid, Spain. [2]Facultad de Medicina, Universidad Autónoma de Madrid (UAM); CBGP UPM-INIA/CSIC, Madrid, Spain. [3]Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid (UPM), Madrid, Spain. ✉e-mail: mark.wilkinson@upm.es

from 19 countries across Europe, representing almost 67,000 accessions from 11,792 taxa native to Europe[2]. ENSCOBASE and EURISCO have grown in parallel but independently for many years, mainly because the goals and plant materials of these databases only partially overlapped. Conservation and utilization of CWR and WFP have become a major meeting point for ENSCOBASE and EURISCO and the main reason why these two information systems should connect. We propose that the FAIR Principles could provide a path to this objective.

The FAIR Principles[7] are a set of guidelines for enhancing the Findability, Accessibility, Interoperability, and Reusability (FAIR, or "FAIRness") of data resources on the Web. While the FAIR Principles acknowledge that humans are a key stakeholder in data discovery and reuse, they focus mostly on the ability of a machine to automatically discover, access, and integrate data, in a world of increasingly large and complex datasets. FAIR can be implemented in a variety of ways, using a variety of technologies; however, it is most commonly implemented using Semantic Web technologies such as Resource Description Framework (RDF) (https://www.w3.org/RDF/) and Web Ontology Language (OWL) (https://www.w3.org/TR/owl2-overview/) — based ontologies, and generally will separate the resource into two primary concerns: The metadata, and the data they describe. In some, but not all cases, the data itself may be transformed to follow the FAIR Principles, particularly in situations where there is significant benefit beyond discoverability, including data integration and/or federated analytics.

Most *ex situ* germplasm in Europe is not FAIR. This makes it a challenge to reuse EURISCO's 2 million accessions. With the goal of improving the situation, the ECPGR (European Cooperative Programme for Plant Genetic Resources) created AEGIS '*A European Genebank Integrated System*'[1] in 2011. AEGIS aimed to increase reusability by selecting "high quality" accessions conserved under standardized procedures. Because of the lack of financial support for PGR-holders to maintain AEGIS standards and derived obligations with an increased number of users, and the lack of a transboundary political vision of conservation by many countries, most PGR-holders in Europe were reluctant to incorporate accessions to AEGIS. Ten years after its creation, only 3% of the EURISCO accessions are present in AEGIS[1]. ENSCOBASE and associated genebanks face similar challenges. Compared to EURISCO, the number of contributing genebanks and number of accessions documented in ENSCOBASE is considerably smaller, no National Focal Points are participating, and no AEGIS-like initiatives to share responsibilities across European countries are being proposed. Moreover, many European seed banks specialized in wild native species are contributing neither to ENSCOBASE nor to EURISCO (data not shown).

The interoperability of germplasm information in Europe also needs to be addressed. Almost a decade after the creation of ENSCOBASE, Rivière and Müller[8] tried to determine, using ENSCOBASE, the contribution of seed banks across Europe towards the 2020 Global Strategy for Plant Conservation[9]. Four key questions were addressed: (1) how are threatened plant species represented in *ex situ* collections; (2) to what extent are accessions of these species available for recovery and restoration programmes; (3) how are CWR species represented in *ex situ* collections; (4) how well is the infraspecific diversity of threatened and CWR species in *ex situ* collections documented in ENSCOBASE. To respond to the first three questions, the ENSCOBASE data were compared using time consuming procedures against external databases. This study did not consider the European seed banks not reporting to ENSCOBASE but to EURISCO. Moreover, the fourth question was not addressed — interoperating with external databases storing information on genetic or ecogeographical variables. These types of issues could have been addressed more accurately and efficiently if all databases (germplasm holders, international germplasm nodes, and external databases) were adhering to the FAIR Principles.

In 2021, the ECPGR decided to explore a new strategy trying to standardize information in all the 400 PGR-holders contributing to EURISCO. This strategy was described in the 'Plant Genetic Resources Strategy for Europe'[10]. By 2030 — just over five years from now — EURISCO would comprehensively apply the FAIR Principles, and the National Focal Points and PGR-holders would be trained to adopt these principles for managing local data, and in particular phenotypic data and data on *in situ* conserved populations of CWR and WFP.

Emergent germplasm initiatives are already beginning to adopt the FAIR Principles. The ongoing European Union funded projects AGENT ("Activated GEnebank NeTwork") (https://agent-project.eu/) and COUSIN ("Crop wild relatives utilisation and conservation for sustainable agriculture") (https://cousinproject.eu/) represent good examples. These projects aim to uncover the full potential for modern breeding programs that can be obtained via the plant genetic resources conserved in genebanks or in natural habitats. Both projects have adopted FAIR standards as a core objective. In 2022, the Conference of the Parties (COP15) to the Convention on Biological Diversity initiated a discussion on how to extend the ABS (access, benefit, sharing) principles of the Nagoya Protocol to digital sequence information (DSI) on germplasm. A document that establishes the basis for discussion was released[11]. The document acknowledges FAIR and CARE (Collective benefit, Authority to control, Responsibility, and Ethics) principles, as the framework for data governance.

Spain has a strong governmental focus on biodiversity. To organize information on plant genetic resources, there is a central node: *Centro Nacional de Recursos Fitogenéticos* (CRF). CRF is a public institution dependent on the *Ministerio de Ciencia e Innovación*. The associated catalogue — the *Inventario Nacional de Recursos Fitogenéticos* (NC) — includes around 90,000 accessions. The CRF node receives information from 38 germplasm banks throughout Spain belonging to the "*Red de Colecciones del Plan Nacional de Recursos Fitogenéticos*". This information is also sent to EURISCO. Unfortunately, neither the contributing banks databases, nor the NC have yet adopted FAIR Principles as established by the 'Plant Genetic Resources Strategy for Europe'[10]. It is also still under discussion if both central node and contributing PGR-holders will be incorporated to the GRIN Global Project.

With respect to the *ex situ* conservation initiatives for wild native species, currently, Spain does not connect to ENSCOBASE via a central node. Connection depends on the desire and capacity of each individual seed bank. The seed banks potentially reporting to ENSCOBASE are mostly members of RedBag (Spanish Network of Germplasm Banks of wild plants and autochthonous phyto-resources). RedBag is composed of about 30 public and private banks. However, the RedBag GBs that participate in reporting to ENSCOBASE are, in practice, fewer

than ten. Recently, the Spanish Ministry of Environment has created the *Banco de Germoplasma Forestal y de Flora Silvestre en Red* ("*Banco en Red*"). *Banco en Red* aims to facilitate access to information on *ex situ* conserved germplasm of forest and wild flora taxa and to promote synergy among its members and the link between its members and other actors. As with EURISCO and ENSCOBASE in Europe, NC and Banco en Red in Spain have not yet adopted the FAIR standards to synergically communicate with each other.

All of these observations circumscribe the interoperability problem — that is, the problem of discovering and utilizing the critical data held in these diverse GBs, and using that data to build contemporary, data-driven conservation strategies. We will now propose a lightweight, distributed, stakeholder-driven FAIR approach to bootstrapping a solution to the interoperability problem, particularly in Spain, but with generic technology that could be reused by other organizations or nations with similar barriers.

The Interoperability Problem, in Summary

The key features of germplasm data that create a challenge to interoperability are:

1. They are highly dispersed. Worldwide there are an estimated 871 seed banks, with more than 5.9 million accessions (unofficial statistics from a draft of the 2023 FAO report[12]). In Spain, alone, there are approximately 38 seed banks dedicated to crops and crop wild relatives and another 20 seed banks dedicated to wild species.
2. They are (geographically) sparse and biased. Collection expeditions travel to locations where they anticipate certain species will be post-bloom and setting seed. Expeditions are complex to organize, and require multiple levels of permissions, thus many locations remain unsampled or under-sampled.
3. The data are fragmented and of varying precision, due to most observations being generated by-hand in notebooks at time of collection, and most observations preceding the appearance of consumer GPS.
4. There are a surprisingly large number of non-identical plant taxonomies that may be used in seed banks.
5. The use of controlled vocabularies is limited, and many of the available controlled vocabularies are taxon-specific (for example, the ontologies hosted by Crop Ontology[13]), and focused on crop species versus wild species.
6. Many germplasm banks — those focused on wild species in particular— are severely under-resourced compared to other kinds of biological/biomedical databases, and thus have only modest Web interfaces (if any).
7. Germplasm access, and arguably data access also, is controlled at multiple governmental levels, from regional to international, through laws and treaties. The formalization of these constraints into machine-readable/processable formats lags far behind what is the case for other areas of data sharing (e.g. patient consent) and thus access requests must be handled manually.
8. Some germplasm data — particularly geolocation data for at-risk or endangered species — are extremely sensitive and are tightly controlled.

Perhaps surprisingly, many of these features are shared in another important domain - the domain of Human Rare Diseases. Rare disease registries are highly dispersed, and sparse in their individual holdings. They are often single-disease specific, or represent a narrow range of similar diseases, and the data is extremely sensitive. With respect to resourcing, however, funding for rare disease is comparatively rich. In 2019, the European Commission established the European Joint Programme on Rare Disease (EJP-RD). With a total budget exceeding €100 Million, one of the objectives of the EJP-RD project was to build a federated network from these highly dispersed, highly sensitive datasets, using the FAIR Principles as guidelines to improve machine-discoverability and processability. This culminated, in 2024, with the publication of the EJP-RD Virtual Platform and Portal (https://vp.ejprarediseases.org/), where rare disease registries can be discovered and accessed in a federated manner.

The FLAIR-GG Platform, Network, and Portal

In late 2022, the *Ministerio de Ciencia e Innovación* of Spain funded the FAIRification, Linking And Integrated Reuse of Global *ex situ* plant Germplasm resources (FLAIR-GG) project, jointly based at the *Banco de Germoplasma Vegetal "César Gómez-Campo"* of the *Universidad Politécnica de Madrid* (BGV-UPM), and the Center for Plant Biotechnology and Genomics (CBGP) of the *Universidad Politécnica de Madrid*. FLAIR-GG has the goal of building a federated network for the discovery and exploration of, primarily, wild germplasm resources, initially focused on Spain, but eventually reaching out more globally. This would be accomplished by following the model of EJP-RD, and as such, the FLAIR-GG infrastructure has four primary components:

1. FAIR metadata servers at every seed bank, adhering to the FAIR Data Point[14,15] (FDP) metadata specification. FAIR Data Points combine[16] two existing Web standards — the Data Catalog Vocabulary (DCAT: https://www.w3.org/TR/vocab-dcat-2/), and the Linked Data Platform (LDP: https://www.w3.org/TR/ldp/) — to create a fully machine-traversable, hierarchical metadata record, capable of leading a computational agent to discover catalog and dataset descriptors, downloadable data distributions, and data access and/or analytical services.
2. A dynamic index of all FDPs that have registered themselves to appear in the FLAIR-GG Network.
3. A Portal (https://vp.bgv.cbgp.upm.es) that explores all registered partners and distributes queries or analytical requests to each site that has advertised in their FDP metadata record that they are capable and willing to accept such a request.
4. An (optional) data transformation pipeline that depends on shared templates in the YARRRML[17] language to convert CSV "dumps" from the native germplasm bank data into FAIR formats. These adhere to pre-formulated and shared semantic models grounded in the SemanticScience Integrated Ontology[18] (SIO),

the FAO-IPGRI Multi-Passport Descriptor Ontology (https://w3id.org/fao-ipgr), and other widely used ontologies and taxonomies in the botanical space.

Component 3 is diagrammed in Fig. 1, while components 1 and 4 are shown in Fig. 2.

We use the word "Platform" to describe the set of standards and expectations circumscribed by the FLAIR-GG project that allows participants to work together without centralization of their data. We use "Network" to describe the loosely coupled participant sites, and we use "Portal" to describe a prototype Website that is capable of exploring each Network member, highlighting their holdings and the data services that they support.

The FLAIR-GG Network Onboarding Process

The process for onboarding a new resource into the network is distributed and only loosely controlled at this time to reduce as much as possible barriers-to-entry for early adopters (complete instructions are available on the project Website: https://wilkinsonlab.github.io/FLAIR-GG/). The minimum requirements to join the network are:

- To have a FDP that contains at least the core metadata about your germplasm bank, such as its name, location, contact information, scope, etc. The Platform supports a range of different options for the method of FDP publication ranging from a dedicated FDP server to simple text files.
- Within that metadata record, "flag" the portions you wish to appear in the Portal.
- Notify the central FLAIR-GG network index of the URL of your FDP.

The content of the metadata is only minimally governed by the Platform and attempts to align with the minimal requirements defined by the European DCAT Application Profile Version 3 (https://semiceu.github.io/DCAT-AP/releases/3.0.0/); however, providers are strongly encouraged to provide richer metadata records that will assist Platform participants and users to discover them. In addition, there is a project specific metadata element that allows members to opt-out of announcing any portion of their data holdings on the Portal, thus giving an additional layer of control to the data provider. The provider sends the URL of their FDP record to the Network indexer, which gathers core information such as the URL, publisher, title, description, the nature of the resources exposed (Catalog, Dataset, Service, etc.) and keywords or ontology terms that have been provided for the purpose of discovery. This metadata record is the primary source of information used by the Portal and is intended to support search across all participating resources, as shown in Fig. 1A,B, as a first step toward building better conservation strategies through the ability to discover general information about the content and policies of each participant.

The Portal also intends to support more intensive data science approaches to building conservation strategies. The data transformation and publication workflow shown in the upper portion of Fig. 2 is optional, and entirely under the control of the Platform participant. All or selected parts of a participant's data may be transformed in a manner compatible with the FLAIR-GG Platform's shared ontological models. After transformation, the FAIR data may be kept entirely private, may be fully or partially exposed for direct query, or may be made available via one or more services that query or analyse the data. These services are not intended to be platform-specific — the Platform supports all (open) Web APIs such as BrAPI or OpenStreetMaps — however, the motivation to build Platform-specific services will potentially be high, due to the availability of harmonized FAIR germplasm records from all Platform participants. The function of the Portal in this case is to identify when similar services are available from multiple Network participants and aggregate them such that they can be simultaneously invoked, thus federating the results over all participants (Fig. 1C). The Portal also offers pre-configured analytics environments (not shown), where data scientists can do additional integrations on the federated output using their preferred language (Python Pandas is the current default).

Examples of Anticipated Benefits

Examples of the benefits of joining the FLAIR-GG network will include a facilitation of both inter-bank collaborations ("horizonal" data sharing) as well as simplifying the aggregation of individual site metadata into national or international catalogs ("vertical" data sharing) either through simplified voluntary uploads, or via harvesting by these higher-level initiatives.

As the network grows in number and diversity of resources, it will increasingly support to the (more) accurate assessment of the geographical areas that are underrepresented in national and international germplasm banks due to the biases mentioned previously. Improvement in interoperability among FLAIR-GG network members means that not only is it possible to check for areas that have not yet been sampled by other germplasm banks, but also enhance the information about those places by using data fromsoil, geographical, climate databases and other relevant environmental data sources. Another example could be to monitor the relative abundance of a species between regions — certain species are abundant in some regions but scarce in others. With this information, germplasm banks can better focus their collection strategies to prioritize species that are universally under-represented. The obstacles to "horizontal" germplasm data sharing and integration makes these kinds of comparisons complicated and frustrating using current approaches. The more participants who join the FLAIR-GG network, the more accurately these kinds of important integrative comparisons can be made, leading to better conservation strategies.

Beyond improving support for scholarly exploration and strategy-building, we see FLAIR-GG simplifying participation of the numerous under-resourced wild germplasm banks in national and international germplasm infrastructure initiatives such as ENSCOBASE. As noted above, in practice, only a handful of germplasm banks are reporting or depositing their information in these larger collections, and this is at least partially due to

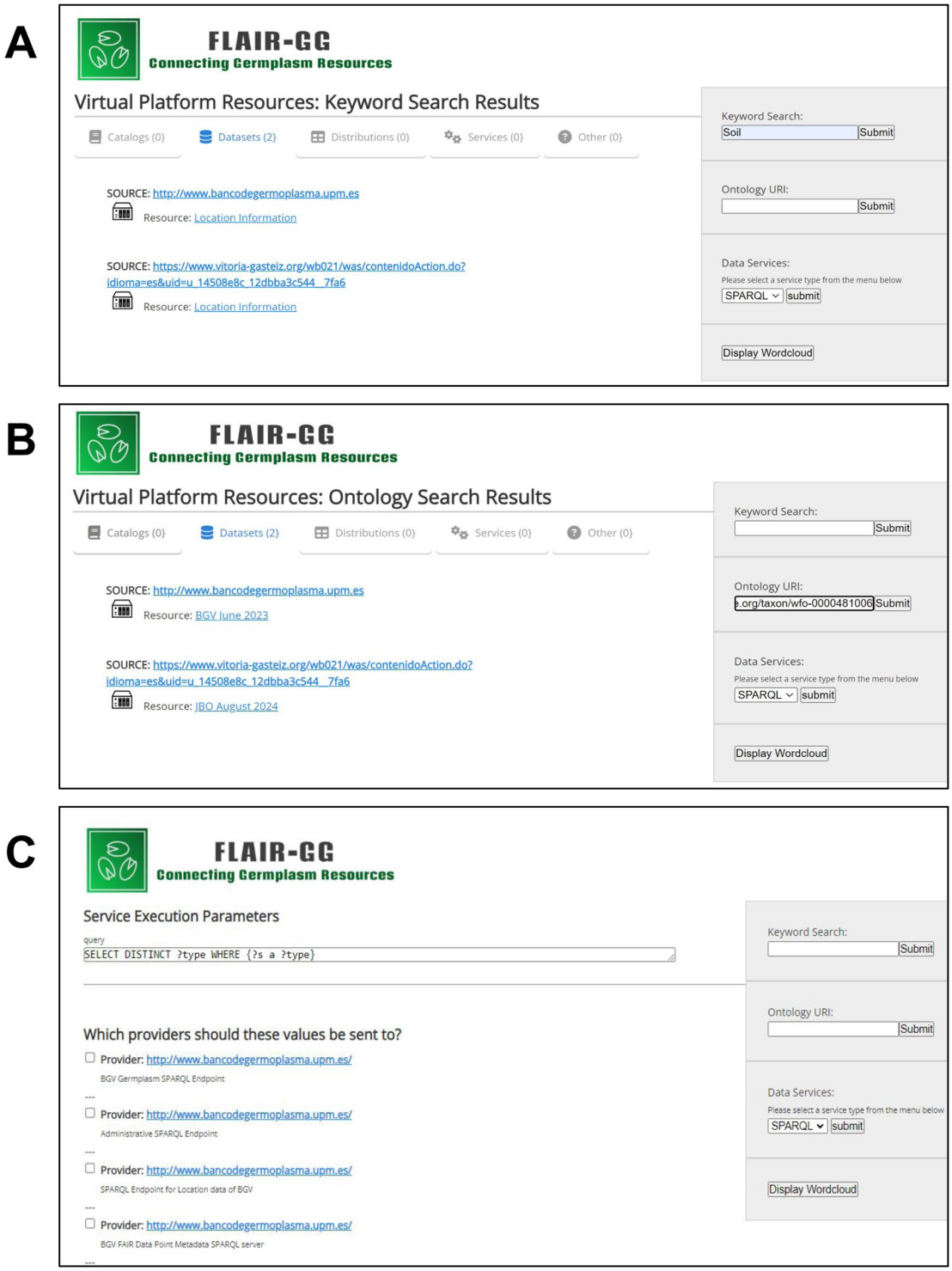

Fig. 1 The FLAIR-GG Portal (https://vp.bgv.cbgp.upm.es/flair-gg-vp-server/resources) provides an overview of all of the catalogs, datasets, and other kinds of resources that appear on the FLAIR-GG network, and some simple federated search functionalities by keyword or ontology term. For example, **(A)** shows a keyword search for the word "Soil", where two datasets have been discovered that are annotated with that keyword, from the BGV Madrid, and the Vitoria Gasteiz botanical garden respectively. **(B)** shows a search using the World Flora Online taxonomy ID for *Papaver rhoeas* L., showing that this germplasm is part of both of these same two collections. **(C)** is an interface aimed at data scientists, providing more detailed exploration of the federated network via Web Services such as SPARQL or other APIs; the Portal provides unified access to all data services that share a common type, among all network participants. In this case, the user has selected the "SPARQL" service type, and is presented with a place to type their query, and an offer to select the providers to whom they want to send this query. Every provider responds independently, and the overall results are displayed in a federated manner (not shown).

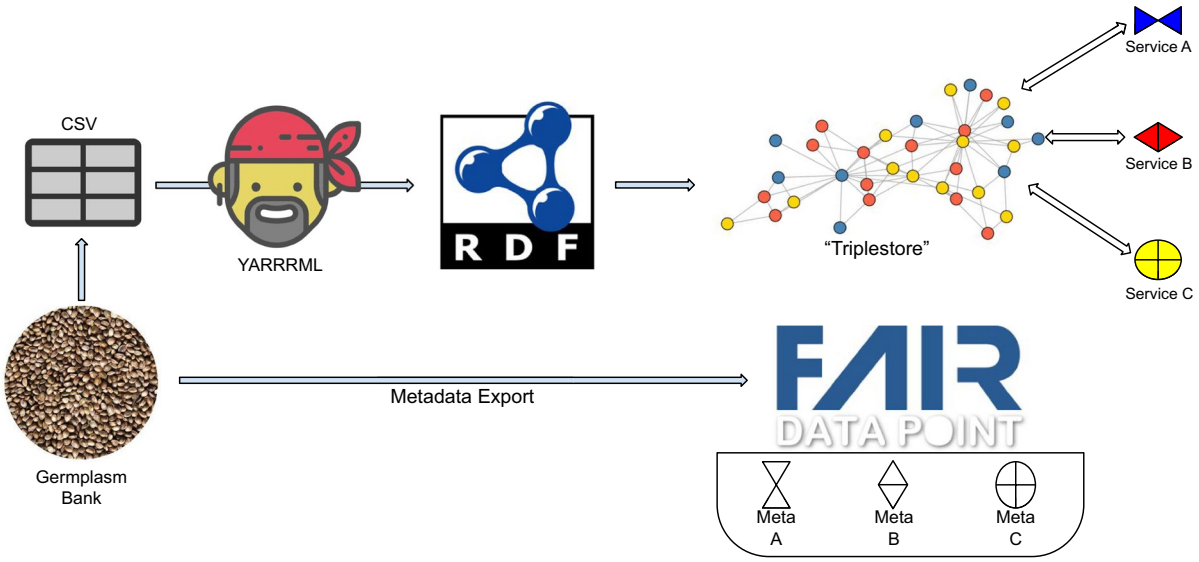

Fig. 2 The lower horizontal workflow shows the export of metadata describing the germplasm bank, and the nature of the data it holds, into a FAIR Data Point. This includes metadata describing any interfaces that might exist that allow exploration of the data itself (Meta A/B/C in the diagram). The upper workflow shows the optional data transformation pipeline that uses CSV as an intermediate representation between the native germplasm database, and the final FAIR data that appears in the Triplestore. Pre-configured and shared YARRRML templates are used to direct the transformation of the CSV into RDF, which is then published in the triplestore. Any interfaces (Service A/B/C) into the data are then pointed at this FAIR representation, rather than the database itself, enabling interoperability between non-coordinating germplasm banks.

their limited ability to create the correctly formatted submission forms. The FLAIR-GG FAIR transformation pipeline is modular, can be applied incrementally, and is based on CSV which is easily generated by any source. Over time, individual sites could build increasingly comprehensive records that not only meet the participation requirements of the large national and international catalogues —facilitating a broad participation in these important initiatives — but also include rich data elements that can only be obtained through direct interaction with the source germplasm bank, ensuring that the bank itself does not become less valuable as a result of "vertical" data sharing. FLAIR-GG will also draw these critical germplasm banks into FAIR compliance in time to meet EURISCO's 2030 ambitions, and moreover, the individual sites will be able to demonstrate compliance with the FAIR Principles to their governments and funders — which is increasingly a requirement — and thereby may improve their ability to obtain funding.

In parallel with adoption of FAIRness by the source germplasm banks, we would similarly encourage the higher-level cataloguing and aggregating initiatives, both national and international, to similarly improve their reusability through coordinated adoption of the FAIR Principles. Within Europe we see the adoption of DCAT (the metadata standard underlying FLAIR-GG) being selected as one of the core technologies for organizing the emergent European Health Data Space[19]. There is little doubt that the same standard will be selected for the Agricultural Data Space initiatives due to the existing commitment of the EU to supporting and tooling-to this standard. While FAIRness alone improves information sharing, coordinated FAIRness — sharing FAIR metadata and data models throughout the community — will dramatically improve interoperability, both horizontally and vertically.

Current Scope and Limitations

The FLAIR-GG Platform is primarily focused on "bootstrapping" FAIRness and interoperability for resource-strapped wild and CWR GBs, with an initial focus on Spain, and ensuring that the critical information they contain can be made discoverable to enhance conservation strategies. This does not, however, exclude any GB in any nation from participating, since in general the larger national and global initiatives lack a shared metadata layer that would allow for simple federation and comparison between them.

FLAIR-GG does not, at this time, include any support for data that is not fully open for exploration (e.g. if the data or API is password protected, it cannot be used on the FLAIR-GG network). Moreover, it offers no differential protection for any data that is exposed by the host; if the host has both accessible and sensitive data, they should not put the sensitive data on the Platform. Similarly, while the Platform will expose the data access and usage conditions defined by the host in their FDP record, it has no way to enforce these; as (primarily) a browser of metadata, it can only simplify discovery of those documents. Future iterations of the Platform will begin to provide mechanisms for onboarding data sources that are not fully open, and technologies that support machine-readable licensing and data access policies are being explored.

At this time, the focus of FLAIR-GG is on germplasm and collection metadata data and does not expand into other domains such as phenotype or sequence data. These other data types are of-interest but given the current

focus on onboarding wild and CWR GBs, they are currently at a lower priority. Moreover, these data may have additional access constraints and international treaty considerations that are beyond the current scope of the FLAIR-GG project. Nevertheless, such data could be brought into the network through Web Services, that can be registered via their DCAT records. These would use whatever data models and standards they natively support. It would be the responsibility of the data host to ensure that they are not providing data to the Platform that is sensitive or in contravention of any treaty.

An Open Invitation

While this is only a superficial descriptor of the aims and functionality of the FLAIR-GG network and portal, more details are available on the project website linked above. The core FLAIR-GG infrastructure is largely complete, and is open for third-party submissions, following the instructions on the project homepage (https://wilkinsonlab.github.io/FLAIR-GG/). We hope that FLAIR-GG provides a pathway towards more optimal usage of our valuable national and international germplasm resources, and a straightforward path to bootstrapping the EURISCO 2030 objective of bringing all germplasm banks into FAIRness.

Received: 19 June 2024; Accepted: 4 December 2024;
Published online: 18 December 2024

References

1. van Hintum, T., Engels, J. M. M. & Maggioni, L. AEGIS, the Virtual European Genebank: Why it is such a good idea, why it is not working and how it could be improved. *Plants* **10**, 2165, https://doi.org/10.3390/plants10102165 (2021).
2. Rivière, S., Breman, E., Kiehn, M., Carta, A. & Müller, J. V. How to meet the 2020 GSPC target 8 in Europe: priority-setting for seed banking of native threatened plants. *Biodivers Conserv* **27**, 1873–1890, https://doi.org/10.1007/s10531-018-1513-2 (2018).
3. Pragna, K., van Hintum, T., Maggioni, L., Oppermann, M. & Weise, S. EURISCO update 2023: the European Search Catalogue for Plant Genetic Resources, a pillar for documentation of genebank material, *Nucleic Acids Research*. **51**, D1465–D1469, https://doi.org/10.1093/nar/gkac852 (2023).
4. van Hintum, T. & Iriondo, J. Principles for the inclusion of CWR data in EURISCO. (2022)
5. O'Donnell, K. & Sharrock, S. Botanic gardens complement agricultural gene bank in collecting and conserving plant genetic diversity. *Biopreser. Biobank.* **16**, 384–390, https://doi.org/10.1089/bio.2018.0028 (2018).
6. Müller, J., Eastwood, R. & Linington, S. ENSCONET – A milestone for european seed conservation. *Studi Tren. Sci. Nat.* **90**, 209–210 (2012).
7. Wilkinson, M. D. *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
8. Rivière, S. & Müller, J. V. Contribution of seed banks across Europe towards the 2020 Global Strategy for Plant Conservation targets, assessed through the ENSCONET database. *Oryx* **52**, 464–470, https://doi.org/10.1017/S0030605316001496 (2018).
9. Convention on Biological Diversity. *Global Strategy for Plant Conservation: 2011–2020* (Botanic Gardens Conservation International, 2012).
10. ECPGR. Plant Genetic Resources Strategy for Europe (European Cooperative Programme for Plant Genetic Resources, 2021).
11. Conference of the Parties (COP15). *Digital Sequence Information on Genetic Resources*. (2022).
12. Commission on genetic resources for food and agriculture. *Revised Draft Third Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. (2023).
13. Shrestha, R. *et al*. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front. Physio.* **3**, 326, https://doi.org/10.3389/fphys.2012.00326 (2012).
14. Bonino da Silva Santos, L. O. *et al*. in *Enterprise Interoperability in the Digitized and Networked Factory of the Future* (eds. Zelm, M., Doumeingts, G., Mendonça, J. P.) (iSTE Press, 2016).
15. Da Silva Santos, L. O. B., Burger, K., Kaliyaperumal, R. & Wilkinson, M. D. FAIR data point: A FAIR-oriented approach for metadata publication. *Data Intelligence* **5**, 163–183, https://doi.org/10.1162/dint_a_00160 (2023).
16. Wilkinson, M. D. *et al*. Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Comput. Sci.* **3**, e110, https://doi.org/10.7717/peerj-cs.110 (2017).
17. Heyvaert, P., De Meester, B., Dimou, A. & Verborgh, R. in *The Semantic Web: ESWC 2018 Satellite Events* (eds. Gangemi, A. *et al*.) 213–217 (Springer International Publishing, 2018).
18. Dumontier, M. *et al*. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomed. Semant.* **5**, 14, https://doi.org/10.1186/2041-1480-5-14 (2014).
19. Hussein, R. *et al*. Getting ready for the European Health Data Space (EHDS): IDERHA's plan to align with the latest EHDS requirements for the secondary use of health data [version 1; peer review: 3 approved, 1 approved with reservations]. *Open Research Europe* **4**, https://doi.org/10.12688/openreseurope.18179.1 (2024).

Acknowledgements

FLAIR-GG: Proyecto TED2021-130788B-I00 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). Several of the technologies described in this work are based on discussions in the European Joint Programme on Rare Disease (EJP-RD) funded by European Union's Horizon 2020 research and innovation programme under grant agreement N°825575.

Competing interests

The corresponding author is co-founder of a consulting company that specializes in FAIR data infrastructures.

Additional information

Correspondence and requests for materials should be addressed to M.D.W.