SEED BANKS INFORMATION UNDER THE FAIR PRINCIPLES

Verykaki, Evrykleia Sofia[1]; Cámara Ballesteros, Alberto[2], Aguayo, Elena[3]
Tutores: Moreno Vázquez, Santiago[1]; Wilkinson, Mark D.[2]
[1] *Departamento de Biotecnologia y Biologia Vegetal E.T.S.I.A.A.B. Universidad Politécnica de Madrid*
[2] *Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas. Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC). Campus Montegancedo 28223 Pozuelo de Alarcón (Madrid), Spain.*
Correo electrónico (AUTOR/ES): sofia.verykaki@alumnos.upm.es; alberto.camara.ballesteros@alumnos.upm.es; elena.aguayo@estudiante.uam.es

ABSTRACT

Seed banks are a pivotal service for scientific research and a key tool for conservation of biodiversity. More specifically, in agricultural science, seed banks operate under a wide range of regional, national, and international rules aimed at maximizing the preservation of plant biodiversity. Unfortunately, automated, and global query of seed banks by researchers and other external users is not yet possible. Conservation actions requiring integrated information spanning all seed banks cannot yet be done, and particularly not in an efficient and automated manner. In order to overcome these limitations, we are attempting to implement the FAIR principles over seed bank data, to improve their accessibility for machines.

Keywords: germplasm banks, data sharing, FAIR principles

BACKGROUND INFORMATION

SEED BANKS: GOALS

Seed banks, also called germplasm banks, are a form of *ex situ* conservation that complement conservation actions *in sit*u (Peres, 2016). Seed banks were created not only to preserve genetic material, but also to speed up research facilitating scientists' access to germplasm. Some banks focus on conservation of wild species, in most instances giving priority to endangered taxons (Engels et. al, 2003). On the other hand, there are seed banks dedicated to crops and their wild ancestors. They aim to ensure the conservation of underused and neglected crop species or landraces within common crops (Bacchetta et al., 2008).

SEED BANKS IN SPAIN

Worldwide, there are approximately 1800 seed banks with more than 7.4 million seed accessions, one third of them in Europe (FAO, 2010). In Spain, there are approximately 30 seed banks dedicated to crops and crop wild relatives and they are grouped in the *Red de Colecciones del Plan Nacional de Recursos Fitogenéticos*. In addition, Spain has around 20 seed banks dedicated to wild species, most of them are grouped in *Red Española de Bancos de Germoplasma de Plantas Silvestres y Fitorecursos Autóctonos (REDBAG) (Rubio et.al, 2018)*. In the latter group is the *-Banco de Germoplasma Vegetal "César Gómez Campo" (BGV-UPM)*, the first seed bank in the world for wild species (Bacchetta et al., 2008; Gómez-Campo C., 2006).

ACCESS TO SEEDS: ABS PRINCIPLES

There are two international agreements regulating access to the germplasm conserved by germplasm banks and also to the associated information. Crop species and crop wild relatives are accessed under the International Treaty on Plant Genetics Resources for Food and

Agriculture (ITPGRFA), other wild species are accessed under the Nagoya protocol. Both international agreements operate under the ABS (Access Benefit Sharing) principle: fair and equitable sharing of benefits arising from the utilization of the accessed germplasm and associated information. Most European countries, including Spain, have signed both treaties.

DATA WITHIN SEED BANKS

Inseparable from the seed bank's goals (conservation and research support) are data or documentation on the seed stocks. A germplasm bank is composed of seeds grouped in accessions(Pita and Martínez-Laborde, 2001). An accession includes the seeds of a particular species, collected in the same place at a particular time. An accession can be described through different types of data. Passport data includes the scientific name of the accession and the data gathered at the collecting site, including location of the collecting site, size of the sampled population, size of the sample, etc. Some authors include also as passport data environmental data obtained *in situ*. The environment shapes the genome; information about the environment surrounding an accession, indirectly informs about the genome(Pita Villamil, 2001). Environmental data includes abiotic (topography, soil chemistry, geology, etc.) and biotic (parasites, competitor plant species around, etc.) factors. Characterization data refers to the information that results from subsequent research and includes genetic, physiological, ecological data, etc.

THE OBJECTIVE: COMPUTATIONAL ACCESS TO SEED BANK DATA

Machine-understandable seed bank data will nourish better scientific research and conservation strategies. Germplasm banks face multiple barriers that impede their interoperability, especially by computers. Some of the main hindrances include outdated or unmaintained databases; disparity of data formats and underlying data storage mechanisms/software; lack of community standards for naming of descriptors; and lack of documentation on how to interact with the data.

THE SOLUTION: FAIR GERMPLASM DATA THROUGH THE FLAIR GG PROJECT

The FLAIR GG PROJECT

To address this challenge a BGV-UPM and CBGP-UPM-INIA-CSIC collaborative project 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 proposes to create a FAIR representation of native seed bank data and to improve the public metadata that describes all participating seed banks. Goal: to enhance their discovery and use.

The FAIR PRINCIPLES

The FAIR principles emphasize the importance of making data Findable by employing unique identifiers and standardized metadata, as well as ensuring that data can be easily located and accessed. Accessibility involves removing barriers to data access through **standardized and non-proprietary communications protocol**, and ensuring maintenance of metadata even if the data is no longer available. Interoperability is crucial for accurate integration and data exchange across different platforms. By adhering to **common data standards and formats**, seed banks can enhance the interoperability of their datasets, facilitating important efforts such as coordination of conservation actions. Finally, the Reusability aspect of FAIR focuses on comprehensive documentation on data usage agreements and licensing to simplify, clarify, and promote the reuse of germplasm data, and provide adequate provenance information to ensure that data creators are accurately credited for their efforts (Wilkinson et al., 2016).

Originally proposed in 2016, The FAIR principles have since gained widespread recognition and adoption across diverse domains, with the life sciences being one of the earliest and biggest adopters, especially in Europe. One of the biggest evolutions in international data management has recently taken place in the human rare diseases field, where data usage and interoperability faced colossal challenges including data scarcity, and widely dispersed, non-coordinating data repositories, making it hard for researchers to find enough data to perform statistically relevant analyses. To help alleviate these issues, the European Joint Programme on Rare Diseases (EJP-RD), more specifically a >100 million Euro international initiative, was tasked with helping rare disease data repositories comply with the FAIR principles, while still maintaining total control over their data and its governance. With the interoperability infrastructure in-place, a Virtual Platform was created, allowing users to discover problem-relevant datasets, and guiding them to these participating data resources in a standardized and well-documented manner, all while avoiding the centralization or exposure of the data itself. This Virtual Platform has been deployed.

The FAIR PRINCIPLES AND THE SEED BANK DATA

The germplasm conservation domain faces similar problems to that found in human rare diseases, albeit with different root causes. Seed collections and the corresponding databases are in general small and scattered, with the information stored under different systems and using non-standardized descriptors. Data is also highly private in many cases; for example, data about endangered species, or the personal information of the people that collected the accession. Information on the traditional use of germplasm is regulated by international treaties (for example, the Nagoya protocol) and could also require conditioned access.
Adherence to the FAIR principles would greatly benefit the seed banks goals and management, by facilitating: 1) integrated conservation strategies, 2) fast and extensive germplasm searches across seed collections based on several variables: taxonomy, geography, climate ranges at the collection site (of paramount importance in a world where climate change is only worsening), type of soil, etc., 3) the workload that germplasm curators face when asked for information about their conserved material, not only by the users, but also by genebank networks, local governments, European administration, and so on. FAIR data makes the germplasm collection more discoverable, increasing the usage of the seeds they store. It also makes the database more interoperable with public databases such as weather and climate agencies, geographical information agencies, annotated collections of publicly available nucleotide sequences and their protein translations, and geological and soil agencies.

CONCLUSIONS

A model for including FAIR principles in Germplasm Bank data has been successfully designed. This model is ready to spread in the different networks of Germplasm Banks in Spain.

BIBLIOGRAPHIC REFERENCES

Bacchetta G., et al. (2008). Conservación ex situ de plantas silvestres. Principado de Asturias / La Caixa. 378 pp.
Engels, J.M.M & Visser, L (eds). (2003). A guide to effective management of germplasm collections. IPGRI Handbooks for Genebanks No. 6. IPGRI, Rome, Italy.
European Cooperative Programme for Plant Genetic Resources. (2021). Plant Genetic Resources Strategy for Europe.
European Commission. Rare diseases. Research-And-Innovation.ec.europa.eu. Retrieved April 5, 2024. https://research-and-innovation.ec.europa.eu/research-area/health/rare-diseases_en
FAO. (2010). Second report on the state of the world's plant genetic resources for food and agriculture. Food and Agriculture Organization of the United Nations (FAO). http://www.fao.org/docrep/013/i1500e/i1500e.pdf.
Gómez-Campo C. (2006). Erosion of genetic resources within seed genebanks: the role of seed containers. Seed Science Research 16, 291– 294

Pita Villamil, (2001a). Documentación de recursos fitogenéticos. In Conservación y Caracterización de Recursos Fitogenéticos. González-Andrés,F.& Pita Villamil, J.M. Ed. Publicaciones I.N.E.A, Valladolid, España
Pita Villamil, J.M. & Martínez Laborde J.B. (2001b). Bancos de semillas. In Conservación y Caracterización de Recursos Fitogenéticos. González-Andrés,F.& Pita Villamil, J.M. Ed. Publicaciones I.N.E.A, Valladolid, España
Papoutsoglou, E. A., Athanasiadis, I. N., Visser, R. G. F., & Finkers, R. (2023). The benefits and struggles of FAIR data: the case of reusing plant phenotyping data. Scientific Data, 10(1), 457. https://doi.org/10.1038/s41597-023-02364-z
Rubio Teso, M. L., Torres Lamas, E., Parra-Quijano, M., De la Rosa, M., Fajardo J., Iriondo, J.M. (2018). National inventory and prioritization of crop wild relatives in Spain. Genetic Resources and Crop Evolution 65:1237–1253
S. Peres (2016). Saving the gene pool for the future: Seed banks as archives. Studies in History and Philosophy of Biological and Biomedical Sciences 55, 96-104
Torres, E., & Iriondo, J. M. (2022). La conservación de los parientes silvestres de los cultivos y la necesidad de publicar datos según los principios FAIR. Conservación Vegetal, 26, 3–6.
Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1). https://doi.org/10.1038/sdata.2016.18