# EXCELERATE Deliverable D7.3

| | | |
|---|---|---|
| **Project Title:** | ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences | |
| **Project Acronym:** | ELIXIR-EXCELERATE | |
| **Grant agreement no.:** | 676559 | |
| | H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1 | |
| **Deliverable title:** | Implementation of BRAPI (Plant Breeding API) at all participating nodes | |
| **WP No.** | 7 | |
| **Lead Beneficiary:** | 27 - INRA | |
| **WP Title** | Use Case B: Integrating Genomic and Phenotypic data for Crop and Forest Plants | |
| **Contractual delivery date:** | 31 July 2019 | |
| **Actual delivery date:** | 29 July 2019 | |
| **WP leader:** | Célia Miguel, Cyril Pommier | 18 - IBET, 27 - INRA |
| **Partner(s) contributing to this deliverable:** | 28 - CIRAD, EI, 1 - EMBL, 16 - FCG, 18 - IBET, 27 - INRA, 6 - NBIC, 33 - NIB, 37 - VIB | |


**Authors and Contributors:**

Cyril Pommier, Célia Miguel, Richard Finkers, Frederik Coppens, Dan Bolser, Manuel Ruiz, Anne-Françoise Adam-Blondon, Kristina Gruden, Živa Ramšak, Evangelia Papoutsoglou, Daniel Faria, Bruno Costa, Inês Chaves

**Reviewers:**

N/A

# Table of contents



# 1. Executive Summary

High-throughput “omics” technologies are widely used and increasingly important to support plant biology research and breeding of diverse plant species for production of food, feed, fibre and other biomaterials, and bio-energy. Significant advances in plant science can be obtained from the integration of available genomic and genotyping data with diverse types of phenotyping data, including field or greenhouse experimental data, molecular, -omics and image data. Although most -omics data, and especially phenomic data, are being generated in increasing scale, either from public or private research institutes, the dispersion of datasets and metadata among multiple repositories and their often poor description and annotation, make their use and exploitation still challenging or even unapproachable.

To help unlock the full potential of a multi-omics approach to plant science, the overarching goal of this work package is to make plant data reusable and interoperable in accordance with the FAIR principles (i.e. Findable, Accessible, Interoperable and Re-usable). Hence, several standards have been built these past years for the annotation of data sets, and the ELIXIR plant community is co-authoring most of them. Their use has been demonstrated by exemplary application to data from diverse species published in public repositories (Deliverable 7.1). They are also the foundation of the ELIXIR Plant Data Search Service, FAIDARE, that gives access to distributed and standardized datasets. This service uses the Breeding API (BrAPI[1]), an API for accessing data relevant for plant breeding developed by the international plant community (Deliverable 7.2). It is based on a federation of plant phenotyping/genotyping data repositories, accessible through a single data discovery webportal[2], that connects not only ELIXIR data repositories, but also any other Breeding API compatible database.

This work represents a major step towards plant FAIR data and a significant tool for the international plant community focused on plant breeding, either of crops or forest trees. Furthermore, with this deliverable we reach the main goal of WP7, to provide a distributed

[1] www.brapi.org
[2] https://urgi.versailles.inra.fr/faidare

infrastructure to allow plant genotype-phenotype analysis based on the widest available public datasets, and we give an important contribution to address goal one of the ELIXIR-EXCELERATE project: Deliver world-leading data services for academia and industry.

# 2. Impact

● ELIXIR Plants is providing a first version of an ELIXIR Plant Data Search service to a federation of data repositories through the data discovery portal for plants, FAIDARE. It currently connects ELIXIR data repositories, but has the potential to integrate in the searchable federation any Breeding API compatible database including some working prototypes with the EMPHASIS information system, PHIS[3]. North American PPN has also requested its addition to this federation[4] . In the near future, ELIXIR Plants is planning to extend the generic, non BrAPI, WheatIS search portal[5] by enabling it for all species and integrating it in FAIDARE. WheatIS lacks some of FAIDARE's functionalities for data visualisation and retrieval but enables the findability of any new type of data without previous development of web services specifications (see below).

● ELIXIR Plants is a major contributor to the development of the Breeding API (BrAPI), an international standard for programmatic access to plant breeding data. We have not only contributed to the development of BrAPI through hackathons (and plan to continue this activity in the future), but also demonstrated its interest by making it a foundation of FAIDARE and collaborate to developments in EMPHASIS. We have hence contributed to its dissemination through talks and partnership, and we are co-authors of the BrAPI publication.

● ELIXIR has been instrumental for the coordination of the development of the MIAPPE and BrAPI standards, for keeping their alignment and their complementarity in collaboration with their respective communities of developers. Furthermore, we have collaborated with the Interoperability and Tools platforms to ensure their inclusion in the ELIXIR Standards Plan, in particular in the frame of Implementation Studies (Data validation, Plant community led). The work has also set the ground for the FAIR-ification of Plant Genotyping Data and its linking to Phenotyping using ELIXIR Platforms in frame of the ongoing Implementation Study FONDUE. ELIXIR Plant is now recognised as a major actor of the global plant community and is co-steering MIAPPE with EMPHASIS, the European Plant Phenotyping ESFRI and the CGIAR.

# 3. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

[3] http://www.phis.inra.fr
[4] https://github.com/elixir-europe/plant-brapi-etl-data-lookup-gnpis/pull/13
[5] www.wheatis.org/search



Horizon 2020 grant n. 676559

| No. | Objective | Yes | No |
|---|---|---|---|
| 1 | Deliver world-leading data services for academia and industry: Establish a distributed genotype–phenotype annotation that supports agriculture research and industrial development. | X | |
| 2 | Make data interoperable (in accordance with the 'FAIR' principles specified in WP5) through the development of controlled vocabularies and standardised APIs, proving the concept of a common phenotypic API through which any participant in an open network can advertise the availability of their data in a common domain. | X | |
| 3 | Annotate and submit key exemplar datasets to relevant public archives. | | X |

# 4. Delivery and schedule

The delivery is delayed: Yes No☑

# 5. Adjustments made

None

# 6. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

| | | | |
|---|---|---|---|
| Work package number | 7 | Start date or starting event: | month 1 |
| Work package title | **Use Case B: Integrating Genomic and Phenotypic data for Crop and Forest Plants** | | |
| Lead | Cyril Pommier, INRA; Célia Miguel, IBET | | |



Horizon 2020 grant n. 676559

**Participant number and person months per participant**

1- EMBL 12.00; 3 - TGAC 21.20; 6 - NBIC 0.00, DLO (LTP) 9.00; 16 - FCG 2.00; 18 - **IBET 53.00;** 26 - CNRS 4.00; **27 - INRA 24.70**; 28 - CIRAD 12.00; 33 - NIB 30.00; 37 - VIB 11.00

## Objectives

The main objective of WP7 is to design and test an infrastructure to allow genotype-phenotype analysis for crop plants based on the widest available public datasets. To achieve this, the specific objectives for WP7 are to:

• Make data interoperable (in accordance with the 'FAIR' principles specified in WP5) through the development of controlled vocabularies and standardised APIs, proving the concept of a common phenotypic API through which any participant in an open network can advertise the availability of their data in a common domain.

• Annotate and submit key exemplar datasets to relevant public archives.

• Engage industry in defining priorities in genotype/phenotype annotations, and collaborate with WP13 in showcasing the developed resources to the agroforestry industry.

• Collaborate with WP11 in delivering specific training for the use of developed resources.

Work Package Leads: Cyril Pommier, INRA (since 25/01/2018); Célia Miguel, IBET (since 01/03/2017)

Description of work and role of partners

**WP7 - Use Case B: Integrating Genomic and Phenotypic Data for Crop and Forest Plants [Months: 1-48]**

**INRA, EMBL, TGAC, NBIC, FCG, IBET, CNRS, CIRAD, NIB, VIB**

This work will facilitate the analysis of many of phenotypes against large panels of crop accessions through the aggregation of locally held data; and thereby, enable more powerful association analysis, opening the way to understanding of function, candidate gene prioritisation, and improved crop breeding. Working on exemplar species, we will establish a sustainable model for the interaction of distributed phenotypic repositories with defined genomic and sample reference data, in which organisations can expose data to the system through conformity with standards for annotation and interface, allowing the subsequent expansion of the approach to other species and domains. It will also provide resources (in the form of standards, ontologies and models for annotation and collaboration) for use within ongoing species-centric (e.g. the Wheat Initiative) and/or national endeavours.

Massive sequencing and genotyping of crop and forest plants (and their pathogens and pests) is generating large quantities of genomic variation data. These efforts are likely to accelerate in the near future, with further expected reductions in the cost of sequencing and international efforts (such as the DivSeek Initiative) aiming to catalogue all genetic diversity present in global germplasm resources. Such data could serve as a powerful panel in association screens and facilitate precision breeding of increasingly complex traits. But structural variation in most crop plants is enormous (more so than in humans),

and phenotypic characterisation data is (i) often inaccessible (ii) diverse and non-standard (iii) lacks any route of unified access. Indeed, “phenotype” is a broad concept, covering many data types (descriptive data, molecular data, image data) at many scales (laboratory, phenotyping centre, field data) on many species; and moreover, a phenotype exists in the specific concept of an experiment (in contrast to a genotype, which is assumed to be constant in a given sample). Both technical and sociological progress in data definition and sharing are lagging experimental progress.

To address this problem, we will harness the domain-specific expertise and data held in a distributed fashion across many national Nodes with interests in agriculture. Seven ELIXIR Nodes will jointly establish a technical infrastructure and associated social practices to define an open model for the publication and sharing of plant genotype-phenotype data, working on a minimum of 3 exemplar species from different domains of plant life to establish a model applicable in all species. We will establish a scalable, distributed model, transparently integrated through the development and use of common vocabularies and search technologies, adhering to the FAIR principles60 (WP5), and using established repositories for genomic data and sample meta data. Domain-specific training will be coordinated with WP11 and will include training users and data curators. The expected impact is accelerated research and plant breeding through the exploitation of an interoperable commons of public data.

**Task 7.1: Development/adoption of appropriate controlled vocabularies for annotating plant phenotypic data (35.5PM)**

The use of controlled vocabularies, to define the material assayed, the form of the experiment and the observed phenotype are critical to enabling diverse datasets to be interrogated and compared. A number of initiatives have developed controlled vocabularies that can be used for the annotation of plant phenotypic data, including the Plant Ontology (http://www.plantontology.org), Crop Ontology http://www.cropontology.org),

Plant Trait Ontology (http://www.obofoundry.org/cgi-bin/detail.cgi?id=plant_trait), Environment Ontology (http:// environmentontology.org), XEML (http://xeml.codeplex.com). Different vocabularies apply in different species, with some specific and some overlapping features; in many countries, national lists of traits found in specific crop varieties are in use by breeders (distinct from the terms in use in academia). Slow-growing forest species have overlapping needs to annual food crops.

Together with representatives from the appropriate infrastructure resources, communities and ontology developers, and industrial/SME partners, we will work on establishing common guidelines for ontology usage when annotating crop and forest species. We will hold a workshop within the first 4 months of the project bringing experts together from all participating Nodes and key industrial participants, working on the target species to agree on a common set of vocabularies to be used in the project (by the end of the month 6). Existing ontologies will be extended where necessary, and cross-references established between corresponding high-level terms in the recommended vocabularies (e.g. between common anatomical concepts in different species-specific ontologies). Outputs will be regularly showcased to industry and list reconsidered at 6 monthly intervals.

Partners: EMBL-EBI, BE, FR, NL, PT, SI, UK

**Task 7.2: Annotation of key plant phenotypic datasets with agreed controlled vocabularies (54 PM)**

At least 3 exemplar species will be chosen, including one cereal species, one woody species, and one other crop species, each of which is of interest to at least 2 ELIXIR partners (maize, oak and potato have been identified as initial candidates). Phenotypic data is very varied and we will work on a variety of data types, including quantitative trait loci, association genetics (laboratory and field data), and biotic stress. Participating Nodes will collect and curate sample and experimental meta data and phenotypic description data to agreed standards using the vocabularies identified and extended in Task 7.1.

Partners: BE, FR, NL, PT, SI, UK

**Task 7.3: Submission of exemplar genomic and phenotypic datasets to appropriate public repositories (35.4PM)**

Annotated datasets will be submitted to appropriate repositories wherever possible, using existing platforms designed for such scope; for example, the European Nucleotide Archive (developed by EMBL-EBI) (for genomic and transcriptomic data), or phenotypic archives e.g. PIPPA (the PSB Interface for Plant Phenotypic Analysis, in development at VIB), BreeDB (in development at NBIC), GoMapMan (in development at NIB), and Ephesis (in development at INRA). Sample identification will be handled through the BioSample DB at EMBL-EBI, or, where the sample is an accession from a public gene bank, by cross-references to EURISCO, the European catalogue of plant collection data (http://eurisco.ipk-gatersleben.de). We will work closely with EURISCO and the gene banks to ensure that appropriate meta data is reliably, non-redundantly associated with samples, and that genomic and phenotypic data held in different resources but derived from the same biological material can be reliably identified.

Partners: EMBL-EBI, BE, FR, NL, PT, SI, UK

**Task 7.4: Development and implementation of agreed public APIs for access to data in participating repositories and exposure via public computational infrastructures (54PM)**

We will develop a common API for data query and retrieval, in close collaboration with WP5. We will build from the experiences already gained by partners in developing repositories and interfaces and will utilise established standards for programmatic data access (e.g. REST, RDF, etc.). The API will be implemented at each partner holding a genotypic, phenotypic or sample repository, allowing users to query a single end point that will return results meeting a common standard from dispersed resources. The API will be exposed to users via the ELIXIR computational infrastructure and other computational infrastructures in the plant sciences (for example, on the iPLANT infrastructure in the United States).

The first version of the API will support identification and query of datasets held in distributed repositories matching specified experimental and sample meta data. The API will be later be extended to encompass the querying of resources based on the phenotypic descriptions.

We would also like to enhance the interpretability of results and lower the barrier of computer competency required by users to access data queries from this platform. We will use the data served by the API to develop intuitive visualisation components to allow researchers to mine phenotypic data accessible through the API. These components will be developed within existing frameworks e.g., PIPPA or BioJS (a JavaScript library of open source components for biological visualisation), allowing their re-use in other contexts.

As part of this task, we will look into the re-utilisation of beacons for plants progression on the application of FAIR principles to services.

Partners: EMBL-EBI, BE, FR, NL, PT, SI, UK

Relation to other WPs

Propose phenotype resources to service registry (WP3)

Exposure on the ELIXIR cloud resources (WP4)

Adhering to and develop vocabularies and API compatible to the FAIR principles (WP5)
Training in the development of plant ontology development and in the use of resources (WP11)

# 7. Appendix 1: Implementation of BRAPI (Plant Breeding API) at all participating nodes

## Introduction

The improvement of plant characteristics and traits for the benefit of mankind has been ongoing for thousands of years. However, the current capacity to speed up plant breeding is higher than ever before due to the power of current plant -omics technologies. Improvement of traits such as disease resistance or increased yield under specific environmental conditions can be tackled by making use of available genomic/ genotyping and phenotyping data for a growing number of crops. The analysis of genotype-phenotype associations can be very complex, since desired phenotypic outcomes can result from single or multiple genes, and from their interaction with each other and with the environment. The systematic study of phenotypes on a genome-wide scale, and its association with genomic/genotypic information under a range of environmental conditions is being increasingly adopted within breeding programs in diverse plant species.

Whether phenotypic data are obtained from field or greenhouse experimental settings, or from phenotyping platforms allowing to generate volumes of data that are several orders of magnitude higher, a number of gaps often make genotype-phenotype associations still very difficult to approach. Unlike the publicly accessible and curated repositories for DNA and protein data, there is no equivalent public repository for the deposition of large amounts of phenotypic data, which are often stored on multiple systems. In cases where data generated from heterogeneous sources can be found and accessed, they are often of poor value because they are not standardized and lack adequate annotation, preventing their re-use.

This Work Package has already produced an extended and revised version of the Minimal Information about a Plant Phenotyping experiment (MIAPPE[6]) standard, which was described in Deliverable 7.1. In Deliverable 7.2. the implementation of 7 BrAPI endpoints over datasets held at the several ELIXIR nodes participating in the Plants Use Case was achieved, thereby publicly exposing these datasets and making them accessible and

[6] http://www.miappe.org

contributing for their interoperability and re-use. In parallel, ELIXIR Plant partners are major contributors to the BrAPI specification and ensured the alignment with the MIAPPE standard for plant phenotypic data, based on our expertise in phenotypic data. The final step of the ELIXIR work plan is to elaborate on the two first deliverables an ELIXIR Plant Data Search Service, named FAIDARE, that allows FAIR phenotyping and genotyping data publication and access through a web interface and a web services API. This service brings sustainability beyond EXCELERATE to all WP7 results, including the presented D7.3 deliverable, and will be available both to academia and industry. It is being deployed as a beta[7] and the final version[8] will be available at the end of the project.

## Report

The work developed to achieve the present deliverable is included in the fourth and last task of the work package, Task 7.4: Development and implementation of agreed public APIs for access to data in participating repositories and exposure via public computational infrastructures. As part of this task, we had previously reached milestone M7.4 (specification of an API for data access) by choosing BrAPI as an interoperability standard for the exposure of plant phenotypic data by participant Nodes within the plant use-case at this stage.

BrAPI specifies a standardized web service interface for plant phenotype/genotype data access to enable interoperability among plant databases, allowing breeders and researchers to exchange, compare and combine data across databases. It is an open API based on the principles of open linked data and a web-based architecture that has been driven by a collaborative community-based global effort involving major stakeholders including the CGIAR, ELIXIR and EMPHASIS. Several members of the ELIXIR Plant work package are key members of this BrAPI community and major contributors to its development. The specification includes the protocols, data structures and services available for automated access and is available at www.brapi.org. BrAPI calls are organized into categories in alignment with the major domains needed for exchanging information between plant breeding information systems and client applications. In the current version, BrAPI calls cover information about germplasm *i.e.* plant material identification, phenotypes, experiments, studies, geographic locations, samples, and genetic markers.

The main content of this deliverable is (1) the improvement of BrAPI specifications, (2) their implementation by all participants to enable exposure of MIAPPE and genotyping datasets across ELIXIR Nodes and (3) the public release of the FAIDARE ELIXIR Plant Data Search service for easy access to the BrAPI data federation.

The following endpoints allowing access to multiple datasets from the several participating ELIXIR Nodes have been implemented, and are described below:

**Table 1** API endpoints implemented by ELIXIR Nodes

| | Interface | API endpoint | Datasets |
|---|---|---|---|

[7] https://urgi.versailles.inra.fr/gpds
[8] https://urgi.versailles.inra.fr/faidare



Horizon 2020 grant n. 676559

| | | | |
|---|---|---|---|
| BE | PIPPA[9] | https://pippa.psb.ugent.be/pippa_experiments/brapi/v1/ | maize |
| FR | GnpIS INRA/GnpIS[10] | https://urgi.versailles.inra.fr/gnpis-core-srv/swagger-ui.html | several crop and forest sp. |
| | TropgeneDB[11] | http://tropgenedb.cirad.fr/api/ | rice |
| NL | EU-SOL BreedDB[12] | http://www.eu-sol.wur.nl/brapi/v1/ | several Solanaceae sp. |
| PT | PHENO[13] | https://brapi.biodata.pt/brapi/v1 | cork oak, rice, *Jatropha curcas* |
| SI | PISA[14] | http://pisa.nib.si/brapi/v1 | potato |
| UK | Brassica Information Portal (BIP)[15] | https://bip.earlham.ac.uk/api_documentation | Brassica sp. |
| EBI | ENA and EVA | ftp://ftp.ensemblgenomes.org/pub/misc_data/plant_index | Genotyping data for all plantae |

The federation built on top of those endpoints capitalizes on the experience gained through the WheatIS and Transplant data discovery portal. Therefore, we decided to build an architecture based on a centralized index to cope with deep pagination and result sorting. Furthermore, this avoids overload and risks of denial of service on the BrAPI endpoints. The Elasticsearch technology has been chosen because it deals well with the BrAPI JSON format and provides powerful, extensible and easy querying as well as scalability, hence being sustainable and adapted to the future development of FAIDARE. Hence, an indexing software has been developed that allows harvesting of metadata and is publicly available[16]. The harvester is developed using the Python language and feeds the Elasticsearch central index (Figure 1). For genotyping data, we have decided to

[9]https://pippa.psb.ugent.be/
[10]https://urgi.versailles.inra.fr/gnpis/
[11]http://tropgenedb.cirad.fr/
[12]http://www.eu-sol.wur.nl/
[13]https://brapi.biodata.pt/
[14]http://pisa.nib.si/
[15]https://bip.earlham.ac.uk/
[16]https://github.com/ELIXIR-europe/plant-brapi-etl-data-lookup-gnpis


Horizon 2020 grant n. 676559

directly harvest JSON generated from Biosamples, ENA and EVA APIs rather than developing a dedicated BrAPI endpoint at EBI. Considering the data volume and the workflow, this is the most efficient choice.

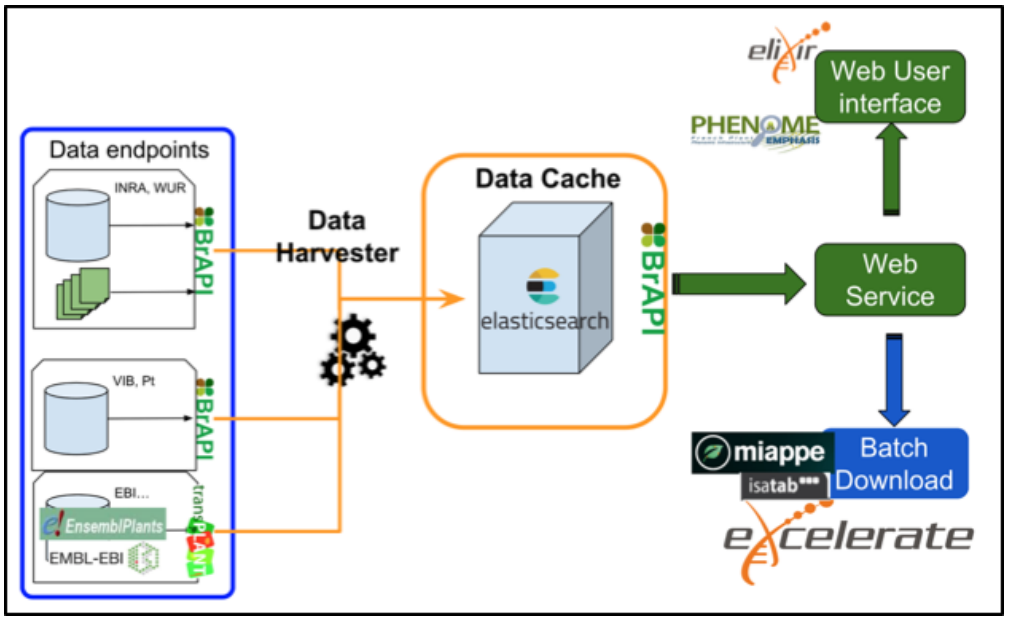


**Figure 1.** FAIDARE (ELIXIR Plant Data Search) Architecture

FAIDARE provides both web service[17] and web user interface access[18] (Figure 3). It will be fully available by the end of the project, while the beta version is already available[19] (Figure 2). The web interface allows searching of data sets based on plant material description and traits. A faceting mechanism allows filtering by data types or data sources. Each result allows the display of a description of the dataset and to link back to the original database.

[17]https://urgi.versailles.inra.fr/faidare/swagger-ui.html
[18]https://urgi.versailles.inra.fr/faidare
[19]https://urgi.versailles.inra.fr/gpds

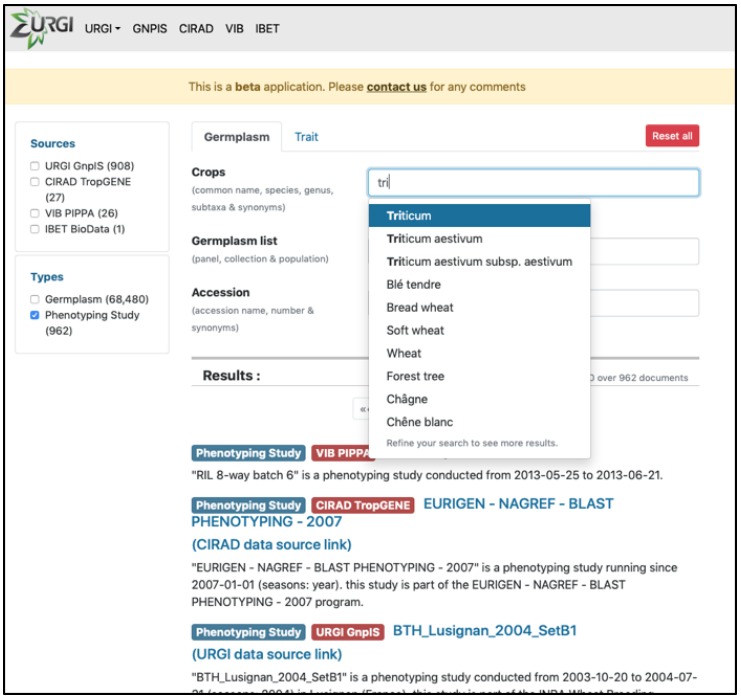


**Figure 2.** FAIDARE beta web interface

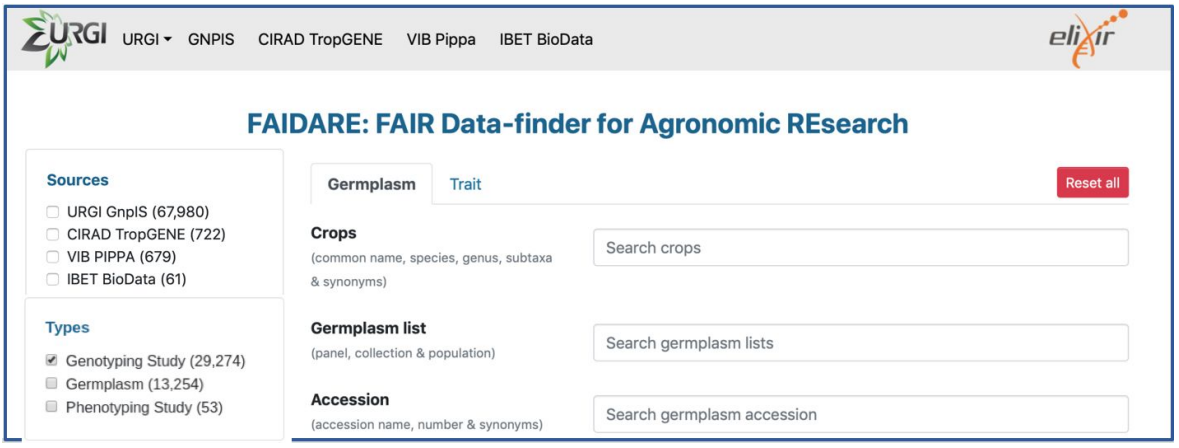


**Figure 3.** FAIDARE web interface with phenotyping and genotyping data access.

## Conclusion

The main objective of Work Package 7 is the implementation of an infrastructure to allow plant genotype-phenotype analysis based on the widest available public datasets. This deliverable (D7.3) ensures the findability and access to related phenotyping and genotyping datasets. In the previously reported deliverable (D7.1), we have worked on the development of controlled vocabularies and submitted exemplar datasets to public repositories, annotated in accordance with the adopted standards to ensure reusability and interoperability of the datasets. The second deliverable (D7.2) began the construction of the repository federation that gives access to distributed plant phenotypic datasets using an API developed by the international community.



Horizon 2020 grant n. 676559

D7.3 is the final step of the work plan conceived at the beginning of ELIXIR-EXCELERATE consisting on the exposure of plant phenotypic and genotypic datasets through the several BrAPI endpoints implemented by the participants. The FAIDARE ELIXIR Plant Data Search service will be sustainable beyond EXCELERATE and is already included in the work plan of several ELIXIR Implementation Study projects, in some EOSC Life demonstrators as well as in future projects.