AGENT Guidelines for dataflow

Authors *(in alphabetical order)*:
Anne-Françoise Adam-Blondon (INRAE, https://orcid.org/0000-0002-3412-9086)
Michael Alaux (INRAE, https://orcid.org/0000-0001-9356-4072)
Matthijs Brouwer (WR, https://orcid.org/0000-0001-8183-0484)
Paul Kersey (RBGK, https://orcid.org/0000-0002-7054-800X)
Matthias Lange (IPK, https://orcid.org/0000-0002-4316-078X)
Erwan Le Floch (INRAE, https://orcid.org/0000-0002-1010-6859)
Cyril Pommier (INRAE, https://orcid.org/0000-0002-9040-8733)
Danuta Schüler (IPK, https://orcid.org/0000-0003-4277-9879)
Nils Stein (IPK, https://orcid.org/0000-0003-3011-8731)
Stephan Weise (IPK, https://orcid.org/0000-0003-4031-9131)
The AGENT consortium

Activated GEnebank NeTwork (AGENT) project website.

This document is complementary to the following Research Data Management toolkit (RDMkit) pages:

- Plant sciences domain page
- Plant Genomics tool assembly
- Plant Phenomics tool assembly

The AGENT project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 862613.

1 Background and aim

In AGENT, we aim to establish a global genebank network to sustainably unlock the genetic diversity of food crops for future generations and make them intuitively accessible for modern breeding programmes.

We will mention the Work Packages (WP) in this document, cf. Figure 1 for WPs details.

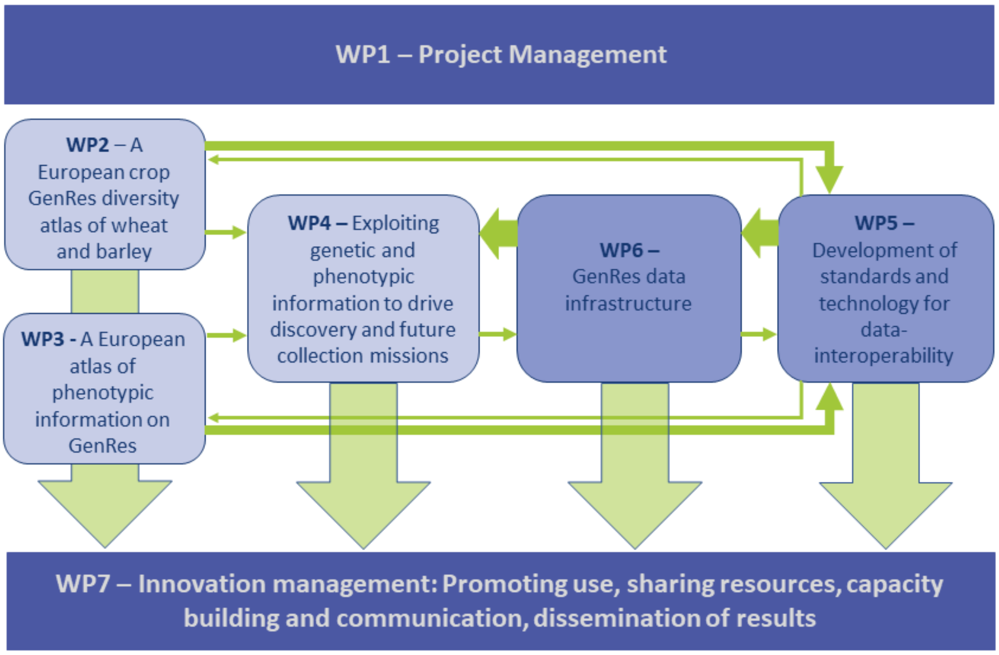

WP1 – Project Management
WP2 – A European crop GenRes diversity atlas of wheat and barley
WP3 - A European atlas of phenotypic information on GenRes
WP4 – Exploiting genetic and phenotypic information to drive discovery and future collection missions
WP6 – GenRes data infrastructure
WP5 – Development of standards and technology for data-interoperability
WP7 – Innovation management: Promoting use, sharing resources, capacity building and communication, dissemination of results

Figure 1. Pert chart for the interrelation and linkage between WPs of the AGENT workplan

To fulfil its goals, the AGENT project aims at integrating data from different sources, namely:

- Genebanks' and research institutes' information systems

The AGENT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613.

- Tools and APIs managing phenotypic data or genomic data using FAIRDOM and AGENT Portal/BrAPI endpoints for access
- European and International core data deposition repositories (EURISCO, Genesys, EMBL-ENA/EVA)

Within the framework of the project, different types of data are collected:

- Passport data
- Phenotypic data
- Genomic data

These data are collected by the project partners using the WP5/WP6 toolset. They are shared using the AGENT FAIRDOM platform, which serves as an internal staging area for files prior to validation and upload into the AGENT database. Where a suitable public repository exists, data are also submitted here, in order to ensure availability after the end of the project. The goal is to ensure that AGENT data are FAIR (Findable, Accessible, Interoperable and Reusable). Figure 2 shows an overview of the most important information systems in the AGENT project.

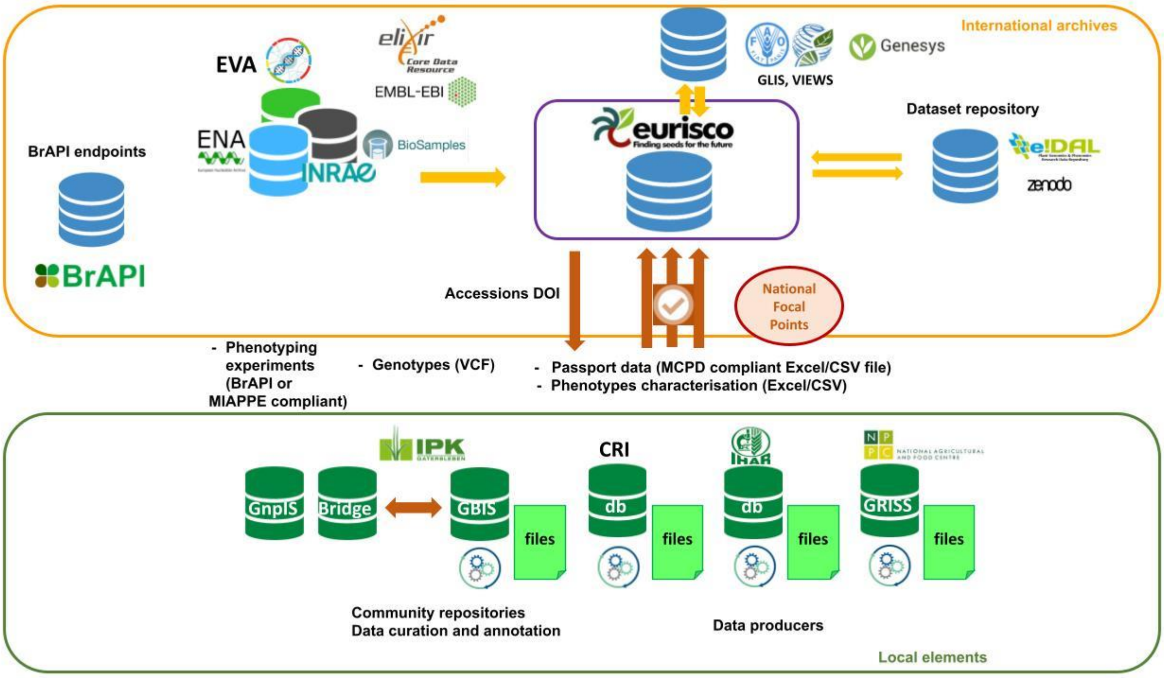


Figure 2: Overview of the existing important information systems related to the AGENT project before the start of the project.

Here, the focus is on the submission of passport and phenotypic data to the European Search Catalogue for Plant Genetic Resources (EURISCO), and genetic data to EMBL-EBI systems: the European Nucleotide Archive (EMBL-ENA) and the European Variation Archive (EMBL-EVA).

The objective is to develop guidelines and recommendations that support the collection, curation and validation of data, as well as the submission and publication of data in international repositories. In this context, minimum metadata standards need to be considered.

2 Current challenges

Against the background of supporting project data flows from production/collection to publication, there are a number of challenges that need to be considered.

2.1 Management of project data until its publication in international archives

EURISCO and EMBL-ENA/EVA make data publicly and openly available. They are not suitable for managing preliminary project data, especially data still under curation like newly collected and unchecked experimental data, raw historical data, or (in the case of EURISCO) data on SSD lines that may not be approved for distribution by the National Inventory Focal Points for Plant Genetic Resources. Separate infrastructures are required to make it possible to provide exclusive access for project participants, and support the imposition of embargo periods, metadata collection and data cleansing.

2.2 Unique identifiers of plant genetic resources material and samples

The general recommendations for the identification of plant material used in experiments is described in the ELIXIR Research Data Management Kit (RDMkit) [1].

In the specific context of plant material provided by genebanks, the unique identification of individual PGR material remains a major challenge. The primary identifier is a combination of the FAO-WIEWS code of the maintaining genebank (INSTCODE), the genus name (GENUS) and the accession number (ACCENUMB). This triplet is complemented in some cases by a DOI, which is a machine-readable identifier, stable over time. The use of DOIs as unique and persistent identifiers is recommended by the AGENT project since accession numbers might be subject to changes over time. The paper "Document or Lose It—On the Importance of Information Management for Genetic Resources Conservation in Genebanks" by S. Weise et al.[2] describes the clear identification of accessions.

Samples are another important type of biological entity to be unambiguously identified in AGENT. A sample is a limited quantity of in vivo material, e.g. organ, tissue or DNA, taken from a plant corresponding to a precise plant genetic resource. It is taken for analysis, testing, or investigation. To ease reuse of genomic databases such as EMBL-EBI ENA, samples can even include whole specimens from a larger population comprising many individual organisms. Sample ID management is also described in the Plant Sciences page of the RDMkit [1].

2.3 Standard exchange formats

To facilitate data exchange between different systems, it is essential, in addition to the agreement on appropriate metadata standards, that there is agreement on associated exchange formats. In the case of passport data, the Multi-Crop Passport Descriptors (MCPD[3], currently v2.1) standard, jointly developed by FAO and Bioversity International, has become widely accepted and adopted. Tabular

[1] https://rdmkit.elixir-europe.org/plant_sciences#plant-biological-materials-metadata-collection-and-sharing

[2] https://doi.org/10.3390/plants9081050

[3] https://hdl.handle.net/10568/69166

The AGENT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613.

MCPD compliant file formats[4] (straightforward .tsv or .xlsx) are used to exchange data between genebanks and international data aggregators such as EURISCO and Genesys.

This is more difficult in the case of phenotypic characterisation data that are currently still poorly standardised in the genebank networks. A format developed in the EURISCO network follows a minimum consensus approach in which only the exchange format is standardised, but not the data itself. It is also necessary to select the minimal metadata fields for considered use cases within the project, *i.e.* historic and experimental data.

The situation is simpler in the case of genetic data. Due to the long history of corresponding international repositories, there are widely accepted formats such as VCF or FASTA but more precise specifications for metadata formatting are needed to ensure interoperability with EURISCO and Genesys, as described in the FAIRcookbook recipe "Plant genomic and genetic variation data submission to EMBL-EBI databases"[5].

2.4 Interoperability and reusability of experimental data

The re-use of experimentally-determined phenotypic data poses particular challenges. First, the description of such experiments is often insufficient. Information such as experimental designs, soil parameters, cultural practices and experimental treatments may be missing or incomplete. Such information is often essential to allow the data to be reused. Second, there are still no generally accepted lists of environmental and phenotypic traits and methods for collecting measured and observed data. Proposals such as the IPGRI/Bioversity Descriptor Lists are an attempt towards achieving this goal, but they are not used in all genebanks or were adapted to specific needs, mostly because the methods and targets for phenotyping are constantly evolving.

More recently, the Minimum Information About Plant Phenotyping Experiments (MIAPPE[6]) metadata standard and the CropOntology trait dictionary have been developed as potential solutions to the challenge of proper annotation of experimental phenotyping data. These are already in use by international infrastructures and networks and are reused in the AGENT project.

3 Data flows in AGENT

3.1 Project data management

3.1.1 Curation of FAIR datasets

To support the collaborative development of FAIR data sets, two repositories restricted to AGENT partners have been set up:

[4] https://eurisco.ipk-gatersleben.de/apex/eurisco_ws/r/eurisco/eurisco-documents

[5] https://w3id.org/faircookbook/FCB061

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

The AGENT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613.

- *FAIRDOM*

An instance of FAIRDOM-SEEK[7] was established at INRAE to allow AGENT partners to share data files still under curation and guide the collection of metadata and its validation before submission to the AGENT database for the phenotyping data.

- *AGENT database*

The AGENT database is a project-specific extension of the EURISCO infrastructure whose development has started in the frame of the European Evaluation Network and is continued in the frame of the AGENT project. It is used in the AGENT project to manage the passport and phenotyping data as well as to connect with the genotyping information.

3.1.2 Publication, long term availability and archiving

Long term availability of the passport and aggregated phenotypic data of the studied accessions will be provided, where possible, through EURISCO. Raw and aggregated phenotyping data will be made available in an open repository such as Zenodo or e!DAL-PGP (https://edal-pgp.ipk-gatersleben.de/). For this purpose, the participating genebanks must ensure that all accessions used in the project (and possibly derived SSDs) are documented in EURISCO, which requires the approval of the responsible National Inventory Focal Points[8]. Before this official clearance, passport and phenotype data will be loaded fully in the AGENT database and phenotyping raw and aggregated data are stored in MIAPPE compliant files in FAIRDOM. Data will be transferred to Zenodo/e!DAL-PGP and possibly to EURISCO after validation and clearance.

Genotyping data will be submitted to the EMBL European Variation Archive (EMBL-EVA) for long term archiving. Submission to EMBL-EVA will generate one to many BioSamples IDs for each accession used (depending on how many times the accession has been sampled), which can be used to make a link between EURISCO and EMBL-EVA data. A reverse link will be added by including in the EMBL-EVA submission the PGR DOI (if available), accession numbers, and other relevant MCPD metadata. Genotyping data will be submitted in VCF format.

3.2 Unique identifiers for the PGR material and samples

Genebanks traditionally assign accession numbers to collections of PGR material. It is also necessary to assign accession numbers to the SSD lines to be used in the AGENT project, as these are line selections of the original genebank accessions with intentionally reduced genetic variability. Samples differ from PGR material in that they are samples obtained from individual plants. These can comprise tissue samples, entire plants, or collections of these. For this purpose, we use the definition of the MIAPPE standard[9] that a sample "represents sub-plant material that was physically collected from an observation unit and was stored and processed before observations are made on it (e.g. in molecular studies)". This means that two levels of identifiers are considered in the data flow. On the one hand, the data flow to genebank materials (§3.2.1) and, on the other hand, the data flow to samples (§3.2.4) and genotyping data collected here.

[7] https://seek4science.org/

[8] A general discussion involving the National Focal Inventory Focal Points about the documentation of derived SSD lines in EURISCO must still be done.

[9] https://doi.org/10.1111/nph.16544

The AGENT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613.

3.2.1 Project internal identifiers: AGENT IDs

Project-internal identifiers (AGENT IDs) are assigned centrally by WP5/WP6 when PGR material is handed over for genotyping to facilitate internal data management. This is done centrally by WP5/WP6. The following scheme is used for this purpose: Prefix AW (AGENT wheat) or AB (AGENT barley) + underscore + five-digit number (see Table 1).

The following mandatory routine is defined for all project partners:

1. When material is sent to WP2 (IPK/IHAR) for genotyping, a list of the material provided must be sent to a dedicated email. A template for such a list is available from the EURICE project management platform[10].
2. As soon as this list is received, WP5/WP6 assigns the necessary AGENT IDs.
3. The sender of the material as well as the genotyping facilities (IHAR/IPK) then receive the list of the assigned AGENT IDs. These IDs must be used in all data exchanges in the further course of the project.

The use of AGENT IDs in the project is mandatory. Only material with assigned AGENT ID will enter into the genotyping routine (cf. Figure 3). The original accession numbers and/or DOIs remain unaffected of the AGENT IDs, as they remain under the authority of the individual project partners that are involved.

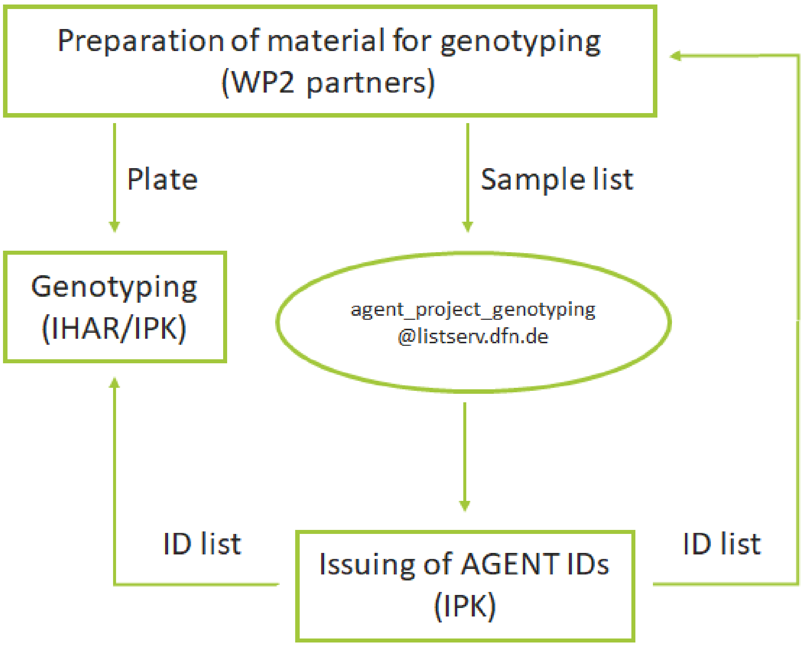

Preparation of material for genotyping (WP2 partners)
Plate
Sample list
Genotyping (IHAR/IPK)
agent_project_genotyping @listserv.dfn.de
ID list
Issuing of AGENT IDs (IPK)
ID list

Figure 3: Genotyping identifiers routine.

The list of material registered with an AGENT ID is continuously growing as the project progresses. It is available to project partners via the AGENT portal[11].

[10] https://agent.eurice.eu/projects/documents/dmsf?folder_id=76 (registered access to AGENT partners)

[11] https://agent.ipk-gatersleben.de/

The AGENT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613.

3.2.2 DOIs for genebank accessions and SSD lines

AGENT strongly recommends that all the characterised plant materials are associated with DOIs for permanent unique identification and resolvability, since accession numbers might be subject to changes over time. This is true for:

- The SSD lines used in the AGENT project, if integrated in the genebank catalogue.
- The original genebank accessions from which the SSDs were derived.
- Any new accession created through the transfer of material from one genebank to another.

For this purpose, the AGENT partners maintaining the accessions and the SSD lines are advised to use the DOI registration service of the International Treaty through the GLIS-DOI portal[12]. The International Treaty provides guidelines that describe the relevant process. For material already documented in EURISCO, the EURISCO coordination can support the DOI registration process on request.

In some cases, DOIs are assigned by the research institutions themselves through individual memberships in the DataCite Consortium. We recommend that genebanks register these DOIs with the GLIS-DOI portal of the International Treaty to ensure a consistent global data infrastructure on genetic resources.

Suggestions are made below on how SSD accession numbers could be formed (see ACCENUMB SSD in Table 1), although the final design is the decision of the genebanks under the sovereignty of the partner's institutes. In addition, project-internal identifiers (AGENT IDs) are assigned centrally by WP5/WP6 for all PGR materials used to facilitate internal data management. This is done centrally by WP5/WP6. The following scheme is used for this purpose: Prefix AW (AGENT wheat) or AB (AGENT barley) + underscore + five-digit number (see Table 1).

AGENT ID	INSTCODE	GENUS	ACCENUMB SSD	ACCENUMB parent
AB_00717	CZE122	Hordeum	AGENT_SSD_03C0601665	03C0601665
AW_02377	CZE122	Triticum	AGENT_SSD_01C0100222	01C0100222

Table 1: Examples of identifiers assigned to PGR materials.

Traceability and unique identification of sequencing samples is implemented by registering them in the EMBL BioSample Repository and assigning BioSample IDs. All genotyping analysis data, such as variant matrices, will be linked to the respective sample(s) and metadata for the sequencing runs by using the permanent BioSample IDs.

A description of the resulting data flow of DOI and identifiers of plant material used in AGENT is proposed in Figure 4.

[12] https://glis.fao.org/glis/

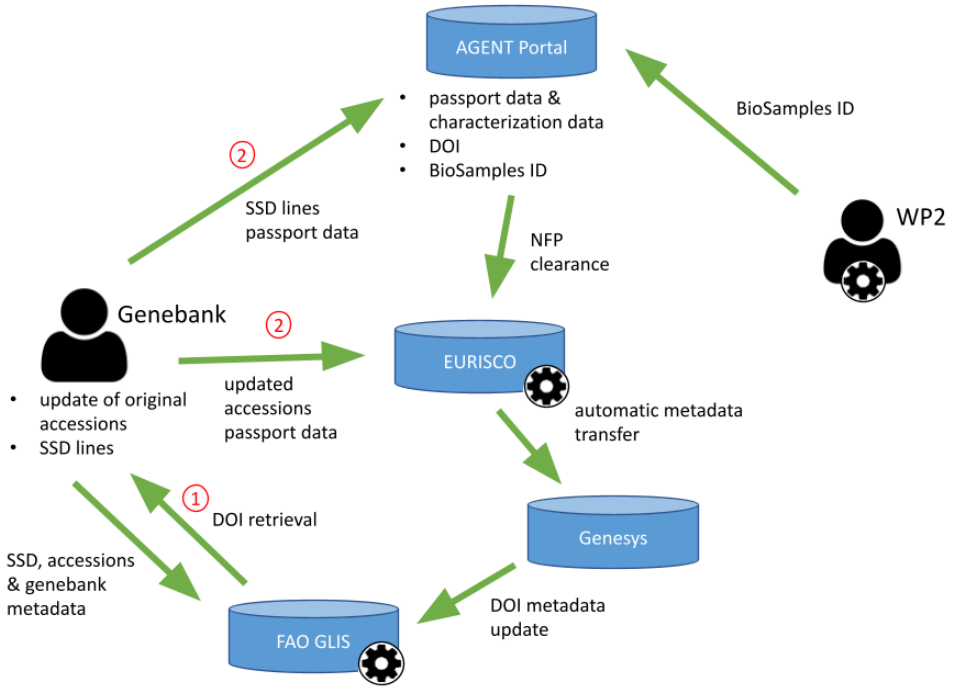

AGENT Portal
passport data & characterization data
DOI
BioSamples ID
BioSamples ID
WP2
2
SSD lines passport data
NFP clearance
Genebank
2
update of original accessions
SSD lines
updated accessions passport data
EURISCO
automatic metadata transfer
1
DOI retrieval
Genesys
SSD, accessions & genebank metadata
FAO GLIS
DOI metadata update

Figure 4: DOI and BioSamples identifiers dataflow. The genebanks are responsible to request the DOI for their accessions and SSD lines from the Treaty, while BioSamples ID are requested by WP2 partners in charge of genotyping operations.

3.2.3 Passport data of SSD lines

SSD lines are derived from genebank accessions, and AGENT recommends that genebanks should not copy-paste the full MCPD passport data from the original accession as their passport data. Instead, SSD lines are considered as new accessions selected from existing genebank accessions. Therefore, only a subset of the passport data should be transferred:

- Taxonomic determination
- Country of origin, collecting site, coordinates, if available
- The original accession number must not be used to allow a clear distinction between the SSD line and the genebank accession. Donor of the SSD line should be the genebank providing the material, the donor number should be the accession number of the original sample. See a suggestion for issuing SSD accession number in §3.2.1.
- The acquisition date should be the provision of the material or the selection of the SSD.
- If the original accession is flagged for AEGIS, this flag is not automatically transferred to the derived SSD line.
- In the case of an Annex 1 crop, it is recommended to make the SSD line available for distribution under the MLS (Multilateral System of Access and Benefit Sharing of the ITPGRFA) on publication in EURISCO.
- Regardless of the original value of the biological status, the derived SSD line should always be listed as “breeding/research material” (SAMPSTAT 400) or “genetic stock” (SAMPSTAT 420).

Other MCPD passport data of the original accession should not be included.

In addition, the link to the original accession must be documented as described above (DONORCODE/DONORNUMB). The DOI of the original accession, if available, can be provided by using the OTHERNUMB field of MCPD and specifying in the REMARKS field “OTHERNUMB contains the DOI of the donor accession”.

After the assignment of AGENT IDs for the SSD lines, WP5/WP6 performs a comparison with the original genebank accessions documented in EURISCO. In this process, passport data are assigned to the SSD lines according to the above proposal. If project partners do not agree with this, the SSD passport data can be changed or supplemented subsequently on request.

In the course of the project, genebanks should clarify with their responsible National Inventory Focal Points which of the SSD lines will be included in EURISCO after the end of the project. EURISCO will also make a recommendation on this.

3.2.4 BioSamples identifiers for DNA samples

The data flow for genotypic data is based on the IPK sequencing process that has been checked with regard to FAIR status and published as a reference process for FAIR aware data processes[13] within the framework of ELIXIR-DE or the German Network for Bioinformatics Infrastructure. The process starts with the creation of the sequencing order (genotyping order for AGENT). The AGENT IDs associated with the panel of SSDs to be genotyped are issued to allow this form to be created and the preparation of the necessary plant material (the cultivation of plants, tissue sampling and DNA extraction). These IDs comprise the AGENT ID of the source PGR, the AGENT ID of the SSD line, an internal unique sample ID and an external sample ID (BioSamples ID) to allow all relevant sequencing data to be traced in the Laboratory Information Management System (LIMS) of the IPK sequencing facility. All these IDs are included in the metadata of the sequence sample that is registered in the Biosamples database and linked to the EMBL-ENA sequence submission.

3.3 Phenotyping data flow

In order to achieve the most efficient management of phenotyping experimental data within the project as well as ensuring its future interoperability and reusability, it is essential to standardise: the collection of standard metadata.

The choice of consistent traits, methods and scales used for phenotyping are important aspects for data reuse. However, they are not in the scope of this document and are the responsibility of WP3.

3.3.1 Metadata standard for phenotyping data

There are two different types of phenotypic data associated with genetic resources:

- Descriptors that represent a general characteristic of the accession that are used to identify the accession or to classify it in relation to its use in agriculture (e.g. morphologic characteristics). These descriptors are present as general knowledge or as an observation (characterization). There have been approaches to standardise descriptors in descriptor catalogues (e.g. IPGRI, UPOV, Comecon-Russian, internal lists), which define the names of traits and the expressions of trait values.

[13] https://doi.org/10.1093/bib/bbab010

- Phenotypes measured in experiments with repetitions and possibly associated experimental factors and environmental data (Evaluation). The Minimum Information About Plant Phenotyping Experiment (MIAPPE) standard addresses this type of data, as described on the RDMkit plant sciences page[14].

The CropOntology[15] provides a standard framework for the capture of the metadata of any phenotypic measurement, which should be associated with a trait, a method of measurement and a scale. A constantly updated dictionary of phenotyping variables exists for wheat and barley: CO_321[16] together with a dictionary of anatomy and development stages: CO_121[17].

3.3.2 Exchange formats for phenotyping data

An exchange format for phenotyping data has been used in the EURISCO network for several years, based on a minimum consensus of what is relevant metadata. Due to its abstract form (Entity-Attribute-Value-like approach), it has mainly been adopted by institutions that already manage their phenotypic data in database management systems, and less suitable for the manual compilation of phenotypic data, e.g. in .xlsx files. To take this into account, the format has been adapted within AGENT.

Two different formats are proposed to AGENT partners:

- A simplified exchange format[18] for historical phenotypic data and aggregated data (e.g. BLUEs, BLUPs, …) produced by WP3 partners for which relatively few descriptive metadata exist.
- A MIAPPE compliant exchange format[19] for phenotyping experiments data and the resulting aggregated data (e.g. BLUEs, BLUPs, …).

[14] https://rdmkit.elixir-europe.org/plant_sciences#phenotyping-metadata-collection-and-publication

[15] http://www.cropontology.org

[16] https://cropontology.org/term/CO_321

[17] https://agroportal.lirmm.fr/ontologies/CO_121

[18] https://urgi.versailles.inrae.fr/fairdom/documents/2 (registered access to AGENT partners)

[19] https://urgi.versailles.inrae.fr/fairdom/documents/9 (registered access to AGENT partners)

3.3.3 Data flow for phenotyping data

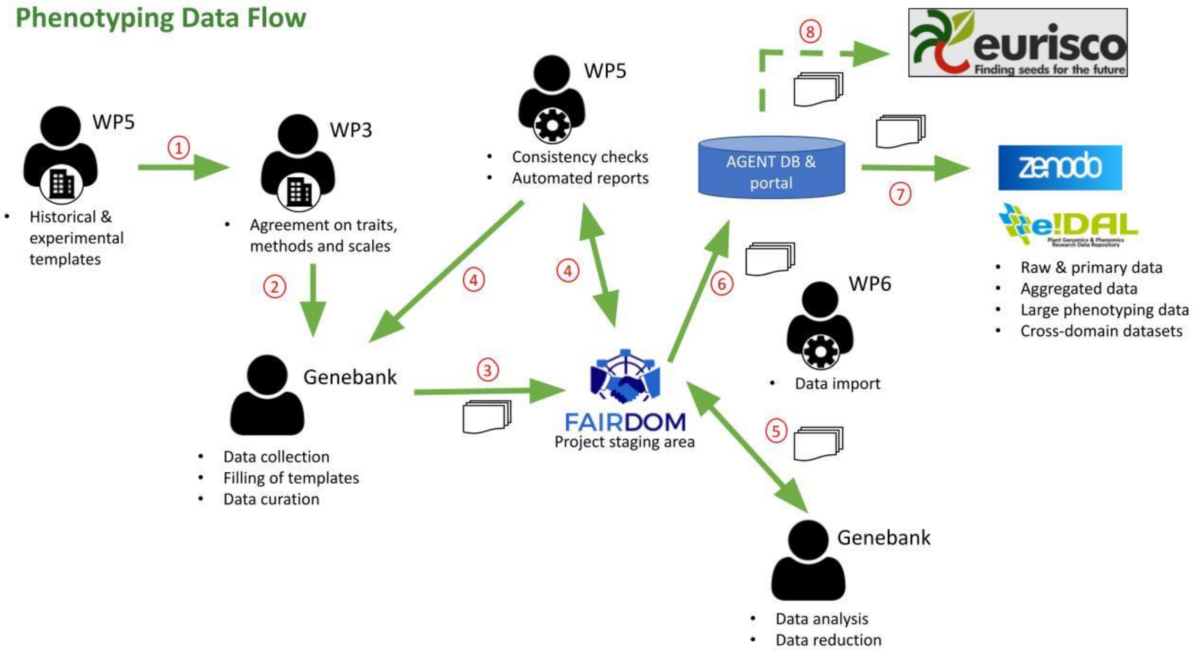


Figure 5: Data flow for AGENT phenotyping data[20]

The following phenotyping data flow is proposed for AGENT:

- Historical phenotypic data[21] and historical aggregated data[22] is provided on the FAIRDOM platform using the simplified format template[23].
    - For aggregated data: the analysis method and parameters are available in the description of the experiment, including a link to the analysis software and relevant resources when available. A dedicated data quality sheet is added with the heritability results.
- Any experimental data generated during the AGENT project (precision, bridging and checks collection)[24] and related aggregated data[25] are compiled using the extended format in the FAIRDOM platform.
    - For aggregated data: the analysis method and parameters must be added in the description of the study and in the method description of each variable, including a link to the analysis software and relevant resources when available. The heritability results will be stored as key-value pairs in the Study sheet.

[20] For automatic data exchanges, MIAPPE has been implemented into the specifications of the BrAPI web services. Key BrAPI end points will be implemented in the infrastructure.

[21] https://urgi.versailles.inrae.fr/fairdom/investigations/2 (registered access to AGENT partners)

[22] https://urgi.versailles.inrae.fr/fairdom/investigations/14 (registered access to AGENT partners)

[23] https://urgi.versailles.inrae.fr/fairdom/documents/2 (registered access to AGENT partners)

[24] https://urgi.versailles.inrae.fr/fairdom/investigations/3 (registered access to AGENT partners)

[25] https://urgi.versailles.inrae.fr/fairdom/investigations/35 (registered access to AGENT partners)

The AGENT project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 862613.

- A validation process takes place: a validator[26] tool automatically checks the metadata and data formats and returns a report. A helpdesk[27] can be mobilised to support error corrections when needed. After successful curation and validation of the phenotypic data, the corresponding data files are marked as final in FAIRDOM.
- Data publication will be done in parallel in two complementary repositories (see also chapter 4):
  - The validated files will be imported into the AGENT database.
  - Consistent data sets consisting in raw and aggregated data and their descriptive metadata will be published in the Zenodo or e!DAL-PGP repositories[28].

3.4 Genotyping Data flow

3.4.1 DNA samples identifiers

In order to make internal DNA sequence identifiers, as described in §3.2.4, globally resolvable and the accompanying metadata interoperable, AGENT adopted the standards for PGR sample tracing developed in the ELIXIR FONDUE project[29]. Here, the EBI BioSamples infrastructure is used to create PUIDs for samples and link to machine actionable metadata. In an MIAPPE extension to the BioSamples metadata schema[30], a core set of relevant MIAPPE attributes is added to the JSON specification. This comprises in general MCPD defined metadata as well as linked material or sample identifiers, tissue related information, developmental stage of the plant and information to the environment, ploidy and environmental data. A subset of this is provided by the partner as described in §3.2.1. BioSamples IDs have been issued for AGENT[31] DNA samples prior to DNA sequence publication. They will be publicly available alongside publication of sequence data at EMBL-EBI archives.

IPK registers a BioSamples identifier (BioSamples ID) for all DNA material and uploads the corresponding JSON metadata record. Each sequence that results from this DNA sample is assigned to the particular BioSamples ID. Within the submission process of DNA sequences to EMBL-ENA this BioSamples ID will be assigned, which enables the traceability of the DNA sample and its relationship to the PGR material within the downstream GBS analysis steps and even its use in literature.

In particular, the computational workflow of variation analysis uses the DNA sequences and therefore the BioSamples ID to identify a genotype uniquely. As a consequence, each genotype in the resulting VCF encoded SNP matrix is identified by a BioSamples ID. In addition, each BioSamples ID is accompanied by the genotype name and source PGR material. This is redundant to the deposited BioSamples metadata but enables ad-hoc human readability of a VCF file.

3.4.2 Exchange format

WP2 takes responsibility to convert vendor formats and SNP/marker mapping to chromosome/position of the reference genome.

[26] https://github.com/AGENTproject/agent_validation

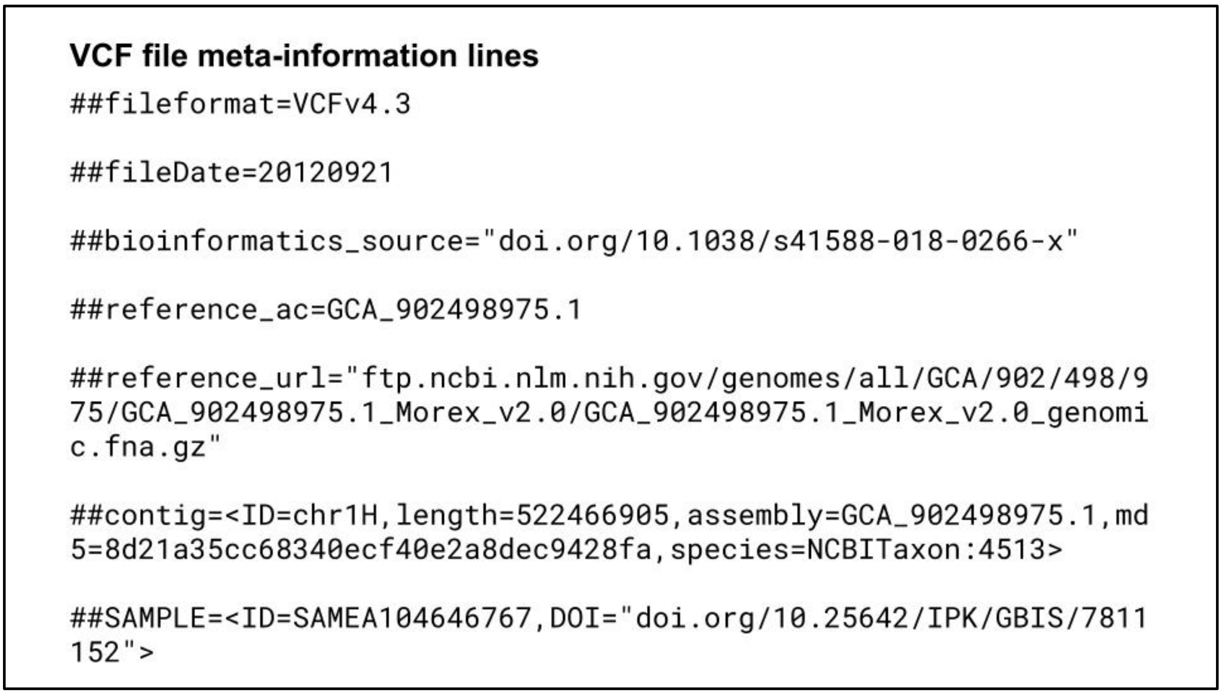
[27] Helpdesk email (internal use), FAQ: https://urgi.versailles.inrae.fr/fairdom/documents/1 (registered access to AGENT partners)

[28] https://edal-pgp.ipk-gatersleben.de

[29] https://elixir-europe.org/about-us/commissioned-services/fondue

[30] https://www.ebi.ac.uk/biosamples/schemas/certification/plant-miappe.json

[31] https://wwwdev.ebi.ac.uk/biosamples/samples/SAMEA7998638



VCF file meta-information lines

```
##fileformat=VCFv4.3

##fileDate=20120921

##bioinformatics_source="doi.org/10.1038/s41588-018-0266-x"

##reference_ac=GCA_902498975.1

##reference_url="ftp.ncbi.nlm.nih.gov/genomes/all/GCA/902/498/9
75/GCA_902498975.1_Morex_v2.0/GCA_902498975.1_Morex_v2.0_genomi
c.fna.gz"

##contig=<ID=chr1H,length=522466905,assembly=GCA_902498975.1,md
5=8d21a35cc68340ecf40e2a8dec9428fa,species=NCBITaxon:4513>

##SAMPLE=<ID=SAMEA104646767,DOI="doi.org/10.25642/IPK/GBIS/7811
152">
```

Figure 6: Example of a VCF format with minimal agreed info lines and encoded data fields.

The exchange format is VCF. The recommendation of encoding SNP data in VCF format was developed in strong collaboration among WP5, WP6 and the ELIXIR FONDUE project. In the course of a 3 days thinkathon[32] in September 2021, a minimal set of header information and encoding of references to the used genome assembly, SNP calling pipeline and encoding of genotype identifier was agreed as shown in Figure 6. Furthermore, WP5, WP6 and FONDUE partners wrote a white paper in F1000Research for detailed format specification: “Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR”[33] by S. Beier et al., 2022. The FAIRcookbook recipe “Plant genomic and genetic variation data submission to EMBL-EBI databases”[34] by S. Beier et al. describes all the steps and necessary metadata fields. A validator checking VCF-files for compliant metadata header has been developed and is distributed as vcfvalidator[35] on the Python Package Index.

[32] https://agent.eurice.eu/issues/581 (registered access to AGENT partners)
[33] https://doi.org/10.12688/f1000research.109080.2
[34] https://w3id.org/faircookbook/FCB061
[35] https://pypi.org/project/vcfvalidator/

The AGENT project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 862613.

3.4.3 Data flow for genotyping data

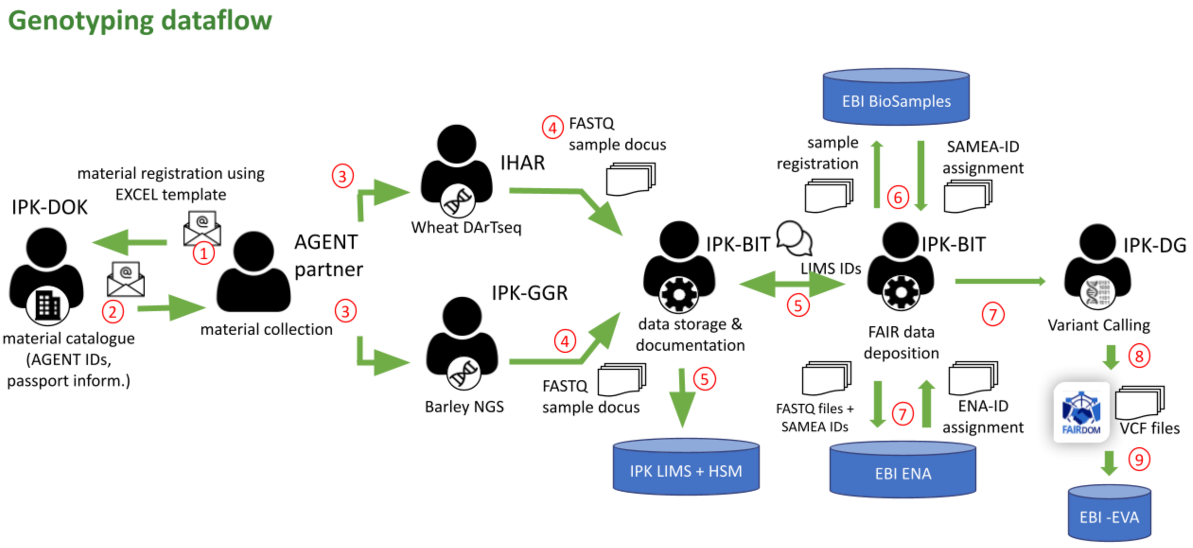


Figure 7: Data flow from data acquisition to EMBL-EVA

The first step is the deposition of DNA sequences by the partner IPK to the EMBL-ENA database. A collaborative AGENT account is already used for this purpose.

In the following steps, the sequences are made available to the WP2 partners responsible for implementing the GBS SNP calling pipeline. The resulting SNP matrices will be provided in accordance with the format specified in 3.4.2. These are first made available to the partners internally via the FAIRDOM platform at INRAE and transferred to EMBL-EVA after quality control and validation.

The data transfer of variation data resulting from SNP chip experiments, and the process of marker and reference genome mapping, which is required to produce a valid VCF file are not yet specified.

Finally, the VCF files are made available to WP4 to run GWAS analysis pipelines and WP6 to pack containers of optimised SNP matrix representation and integrate them with phenotyping data and pre-computed PCA results to serve interactive visual data exploration.

To add external sequence data (generated outside the AGENT project) to the genotyping analysis, some checks, like metadata completeness check, sample and sequence data registrations have to be performed. Furthermore, the license of the external data has to be evaluated for permission to include the data in AGENT data deposition, analysis and publication processes. The steps are detailed in Figure 8.

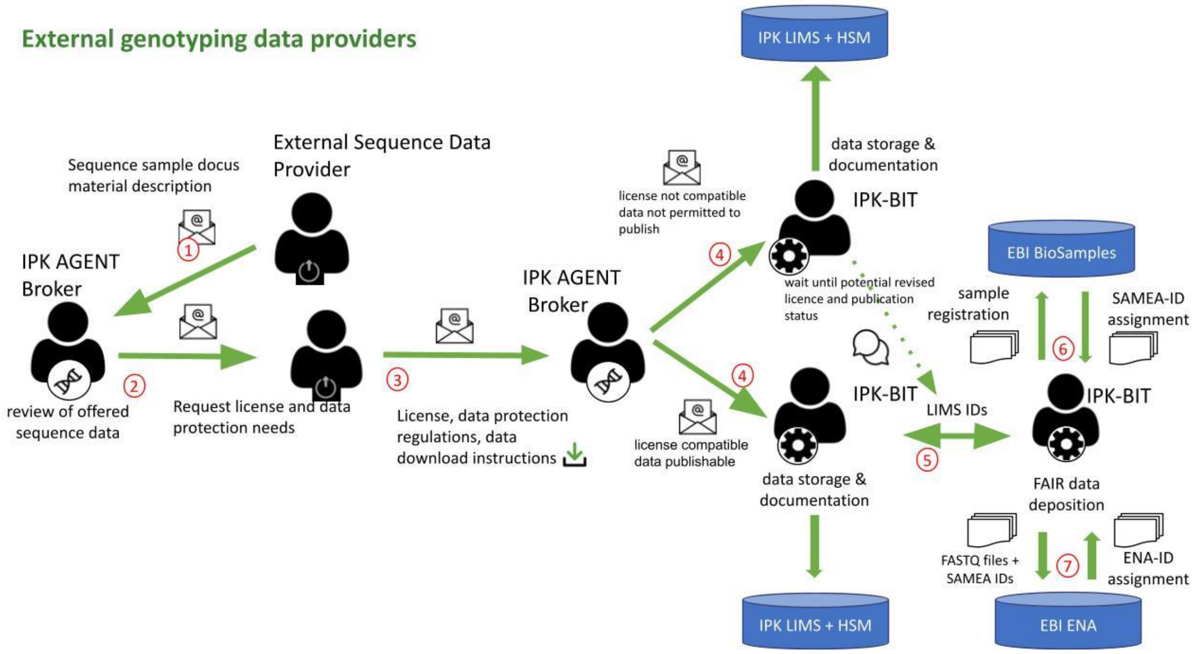

External genotyping data providers
External Sequence Data Provider
Sequence sample docus material description
IPK AGENT Broker
review of offered sequence data
1
2
Request license and data protection needs
3
License, data protection regulations, data download instructions
IPK AGENT Broker
license not compatible data not permitted to publish
4
4
license compatible data publishable
IPK LIMS + HSM
data storage & documentation
IPK-BIT
wait until potential revised licence and publication status
IPK-BIT
data storage & documentation
IPK LIMS + HSM
LIMS IDs
5
EBI BioSamples
sample registration
SAMEA-ID assignment
6
IPK-BIT
FAIR data deposition
FASTQ files + SAMEA IDs
7
ENA-ID assignment
EBI ENA

Figure 8: Data flow from external sequence data providers

4 Data publication

4.1 Open repositories

Data formatted using the guidelines provided in section 1, 2 and 3 should be well described by rich metadata and well formatted facilitating their reuse beyond the project. They will be published in two complementary ways:

- Phenotypic data in the AGENT database and potentially EURISCO; and genotyping data in EMBL-EVA
- Consistent data sets associated with scientific papers described must be published in an open repository such as Zenodo or e!DAL-PGP. They will be used to store part of the data (e.g. phenotyping raw and aggregated data sets; GWAS associations, …) and link to EMBL-EVA data sets when necessary. Each Zenodo or e!DAL-PGP data set is associated with a DOI that can be cited.

  Phenotyping data can be published using the AGENT template, plus optionally an MIAPPE-compliant ISA TAB archive in the same data set.

  A Zenodo community has been created for the AGENT project. To upload a new dataset in this collection:

  - go to https://zenodo.org/communities/agentproject/
  - authenticate
  - proceed to the upload of your data set, and mention “AGENT has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No. 862613”.

WP6 will extract the consistent data sets from the AGENT database and will coordinate their submissions to the chosen open repository.

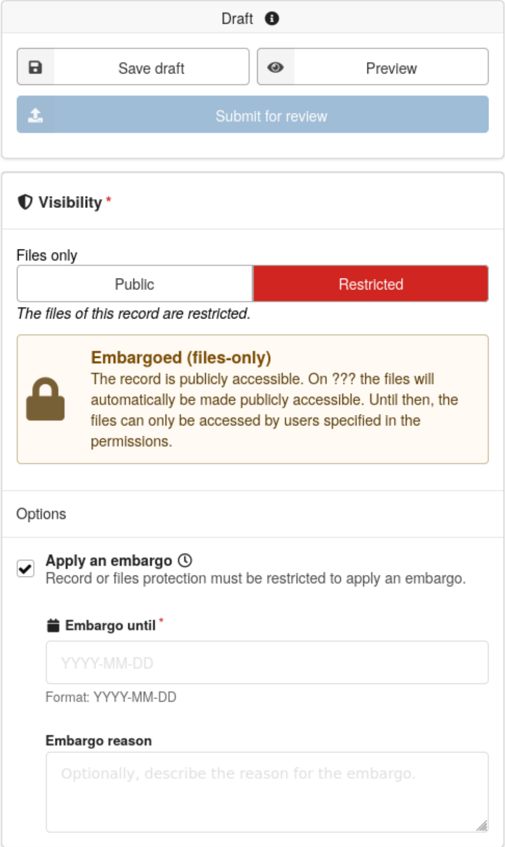
Pre-publication access to data is provided under the terms of the Toronto Agreement. The AGENT portal will display a Toronto agreement disclaimer on its front page and the BrAPI endpoint will specify the Toronto agreement in its metadata.

Example of Toronto agreement disclaimer:
All of the data listed here is available under the prepublication data sharing principle of the Toronto agreement (1). By using this data, you agree to:

- respect the rights of the data producers and contributors to analyze and publish the first global analyses and certain other reserved analyses of this data set in a peer-reviewed publication.
- not redistribute, release, or otherwise provide access to the data to anyone outside of the group, until the data has been published & submitted to the public data repositories.
- contact the authors to discuss any plans to publish data or analyses that utilize this data to avoid the overlap of any planned analyses.
- fully cite the prepublication data along with any applicable versioning details.
- understand that this data as accessed is precompetitive and is not patentable in its present state.

This agreement does not expire by time but only upon publication of the first global analysis by the data producers and contributors.
(1) Toronto International Data Release Workshop Authors. Prepublication data sharing. Nature 461, 168–170 (2009). https://doi.org/10.1038/461168a

In the open repositories, it is possible to apply an embargo during the submission. Figure 9 shows an example in Zenodo.



Draft

Save draft | Preview

Submit for review

Visibility *

Files only

Public | Restricted

The files of this record are restricted.

Embargoed (files-only)
The record is publicly accessible. On ??? the files will automatically be made publicly accessible. Until then, the files can only be accessed by users specified in the permissions.

Options

☑ **Apply an embargo**
Record or files protection must be restricted to apply an embargo.

Embargo until *
YYYY-MM-DD
Format: YYYY-MM-DD

Embargo reason
Optionally, describe the reason for the embargo.

Figure 9: How to apply an embargo in Zenodo

4.2 Data FAIRness

The FAIRness of the data published is guaranteed by the use of the following tools and API:

- The Breeding API (BrAPI) project is an effort to enable interoperability among plant breeding databases and tools. It is a standardised RESTful web service specification for exchanging plant breeding data. This community driven standard is free to be used by anyone interested in plant breeding data management. BrAPI compatible endpoints are currently being implemented on the AGENT database to improve the data accessibility and the interoperability with other information systems (https://github.com/AGENTproject/BrAPI).
- e!DAL-PGP is a comprehensive research data repository, which is hosted at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben and is mainly focused on sharing high valuable and large genomics and phenomics datasets. It is the first productive instance, which is based on the open source e!DAL infrastructure software and is furthermore a part of the de.NBI/ELIXIR Germany services. All provided datasets are FAIR compliant and citable via a persistent DOI. By using the widely established LifeScience AAI. The key feature of e!DAL-PGP is its user-friendly, simple and FAIR-compliant data submission and review procedure. The repository has no general limit to any type of size of datasets.
- Zenodo is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artefacts. For each submission, a persistent digital object identifier (DOI) is minted, which makes the stored items easily citeable. Zenodo self-assesses its compliance with the FAIR principles (https://about.zenodo.org/principles/).
- The FAIDARE[36] portal facilitates the discoverability of public data on plant biology from a federation of established data repositories. It is based on the Breeding API (BrAPI) specifications and facilitates the access to genotype and phenotype datasets for crop and forest plants through an easy to use web interface. It also provides a standard interface that can be accessed programmatically through web services. FAIDARE will index the AGENT BrAPI endpoint as well as AGENT datasets from the open repositories. EMBL-EVA and e!DAL-PGP are already indexed, and the Zenodo AGENT community will be indexed. FAIDARE will ensure the findability of all the AGENT data for re-use by the community.

[36] https://urgi.versailles.inrae.fr/faidare

The AGENT project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613.