Innovative Management of Animal Genetic Resources

IMAGE

Innovative Management of Animal Genetic Resources

Grant Agreement Number: 677353
Horizon 2020 FRAMEWORK PROGRAMME

TOPIC: MANAGEMENT AND SUSTAINABLE USE OF GENETIC RESOURCES

Topic identifier: SFS-07b-2015

Type of Action: Research and Innovation Action (RIA)

DELIVERABLE **D5.1**

Report on the conceptual model for IMAGE web portal integrating data from gene bank with related genomic, geographical and phenotype data

Abstract: this deliverable describes the objectives of the IMAGE web portal and the conceptual model developed in order to integrate data sources and relevant archives.

Due date of deliverable: Month 24
Actual submission date: Month 28
Start date of the project: March 1st, 2016
Duration: 48 months
Organisation name of lead contractor: **PTP**
Contributors: EPFL, EMBL-EBI, WU (WP4 leader), INRA, DLO (WP2 Leader)
Dissemination level: PU
Revision N°: V1

Table of contents

Executive Summary

Background	The integration of information is a challenging task and it has become increasingly complicated due to the large amount of data currently generated by sequencing and genotyping technologies. Integration of phenotypic information is even more challenging, due to the high heterogeneity in the collection, definition and description of phenotypes. Thus, at present, data are scattered in different information systems and documentation of gene banks and genomic collections is not as rich and informative as it should be.
Objectives	Creation of a European web portal which integrates data from animal gene banks and collections, all over Europe, with genomics data, geographical data, and other relevant metadata. Objective of this deliverable is the construction of the conceptual model for a common database superstructure, allowing the leveraging of the access and use of genetic material and ensuring sustainability in time.
Methods	Steps taken to elaborate a consensus conceptual model include: Analyses of data sources; Networking with management of other relevant databases; Identification of sustainable archives; Definition of data flow metadata and relative standards.
Results & implications	A conceptual model was developed. According to the data flow plan, gene bank data from the main sources (Cryoweb, well organized National databases, and other smaller resources) are submitted to the EMBL-EBI's BioSamples database, through the IMAGE inject tool. Long term storage is ensured by the European BioSamples database. Each record is assigned a BioSamples ID that enables subsequent linkage to genomic data that is deposited in EMBL-EBI sequencing archives. This data structure, centred around this unique BioSamples ID, enables each sample to be integrated with molecular data held across public data archives and displayed in the central IMAGE portal, and allows direct mapping back to the records held in each gene bank. The portal enables direct cross referencing from each sample record to external resources, enabling integration of additional phenotypic, GIS and gene bank information held on external sites. The structure of the portal will also enable external sites to include direct links to specific samples and molecular assays by their unique ID. The common data pool (CDP), a local database, is routinely updated by downloads from the BioSamples database in order to provide the web portal with the latest direct link records. The CDP integrates data from different sources, including EUGENA and other EMBL-EBI databases, to allow more complex queries.

1. Objectives and approach

The main objective is to define the architecture and the data flow behind the IMAGE portal.
Providing a web portal with information on animal genetic resources addresses a long-term concern of the European Commission and particularly DG Agri Council Regulation since the council regulation No 870/2004. (see more on https://ec.europa.eu/agriculture/genetic-resources_en).

The IMAGE data portal will be closely linked and aligned on the one hand, with the EUGENA portal, managed by the European Regional Focal Point for Animal Genetic Resources, and on the other hand, with the European/global breed databases EFABIS/DAD-IS. Different user groups for the IMAGE portal include researchers, extension organisations, gene bank managers, policy makers, breeding associations, and industry/private breeders.
The approach taken aims to identify sustainable archives and facilitate the input of the data as well as easy querying of relevant collections.

2. Data sources

Four major classes of data were considered in the development of the conceptual model for the IMAGE common data portal (CDP):
1. genebank data
2. genomic data
3. geographical data
4. phenotypic data

Gene bank data
The gene bank data group includes data on samples stored in gene banks or sample collections. Typically, it includes metadata such as species, breed, animal ID, location *et cetera*. Depending on the data recording procedures, the information available may greatly differ, thus sources need to be grouped by their organizational structure and by their data recording setup. Two main possible sources for this class are gene banks (germplasm) and collections (genomic material).
Integration of gene bank data is of central importance in IMAGE. Such data are usually well organized. In the EFABIS-NET EU project, a data recording system was implemented in more than 12 national gene banks resulting in an identical data structure as collected in the CryoWEB system (Duchev *et al*. 2010). This data set is therefore well defined but its meta-data is country centered, recording information in their national language, and will require super national integration. A second group of gene banks has its own well-developed data collection which operates on a sizeable collection of samples, such as in the recently developed French national CRB-ANIM database.

Finally, gene banks exist with a less formalized data recording and storage schemes.
Genomic collection data are inventories of biological material often having been set up in the context of research projects (including from IMAGE) which differ in their objectives, but share a common concern regarding the safe preservation of samples, which can be of interest for further studies. These data sets show a broad range of organizational levels, from relational database with public access to spreadsheet.

Genomic data

One major thrust of IMAGE is the addition of genomic data to the gene bank material. Genomic data (SNPs and sequences) are increasingly becoming available from research projects or breeding programmes, while relatively little genomic data is available for gene bank material, as pointed out in D4.1 of IMAGE. For -omics related data, a number of public data resources have been funded within the EU, mainly hosted at EMBL-EBI, which have enhanced data collection and standardisation and have become an international reference for sustainable storage of genomic information. The volume and complexity of genomics data are such that the resulting outputs of a large-scale project cannot be stored in a single archive but must be organised and submitted to multiple public resources, each dedicated to a specific data type. Although EMBL-EBI is currently developing a unified submissions interface that will accept, validate and distribute data of different types to its various archives from a single submission.

Geographic data

The third data class is the geographic information system (GIS) defining the origin of the preserved material by connecting a sample with a georeference. All the sources of gene bank data are potential suppliers of georeferences. While for some material the GPS coordinates data are available, for others only addresses may be present. Again, a complete integration is required.

Phenotypic data

Finally, the fourth class refers to phenotype data, available for a subset of gene bank samples. While genomic data are relatively well defined, as are GPS coordinates, phenotypes can be anything from metric performance data on a host of traits to verbal description, with the resulting issue of managing this class across traits/populations/species. Some ontologies have been developed (ATOL) and the international consortium ICAR is developing standards for livestock, which mainly deal with ruminants. Furthermore, phenotypic data being the basis for breeding programmes, they are generally not publicly available and agreements from breeders are needed to grant access to them.

3. Identification of suitable archives

To ensure sustainability in time and to leverage the opportunity to reach the largest possible audience for animal genetic resources data, EMBL-EBI, also a partner in the project, will facilitate the use of its public archives and tools within the framework of IMAGE.

EMBL-EBI archives, Biosamples (https://www.ebi.ac.uk/biosamples/), EVA (European Variation Archive, https://www.ebi.ac.uk/eva/), ENA (European Nucleotide Archive, https://www.ebi.ac.uk/ena/) are identified as the best choice for the objective and are also in compliance with relevant international initiatives such as the FAANG (Functional Annotation of ANimal Genomes) project.

In particular, all samples will be registered in BioSamples at EMBL-EBI as this sample archive has the best support for 'child of' and 'derived from' sample relationships. The NCBI BioSample database is a peer of the EMBL-EBI BioSamples, and they exchange data regularly. IMAGE samples should be registered in the EMBL-EBI BioSamples prior to data submission.

4. Conceptual model

Definition of data flow, metadata and relative standards

A conceptual model has been developed and it is represented in Figure 1. According to the data flow plan, genebank data from the main sources (Cryoweb, well organized National DB such as CRB–Anim in France, and other smaller resources) are transferred to the EMBL-EBI's BioSamples database, through a bespoke IMAGE Inject tool. Sustainable and scalable storage is ensured by BioSamples. Each record is provided with a unique BioSamples ID which allows linkage to genomic data subsequently deposited in EMBL-EBI sequencing archives. This data structure enables all data to be integrated and displayed in the IMAGE portal, and allows mapping back to the records in each genebank.

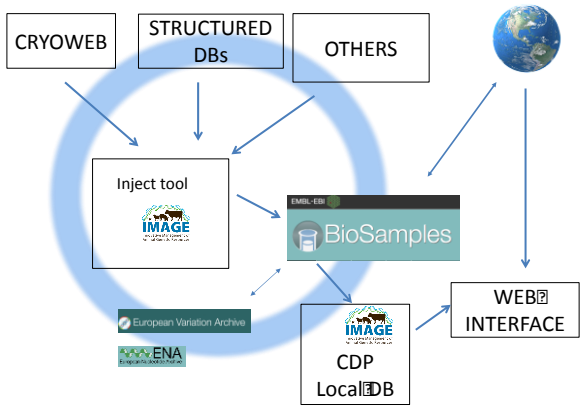

Figure 1. Data flow behind the IMAGE Common Data Portal (CDP).

The common data pool (CDP), i.e. the internal database backend that feeds the web and programmatic interfaces, is routinely updated from the information submitted to BioSamples. The CDP integrates data from the different archive sources (EVA, ENA, EUGENA) to allow for more complex queries across datasets.

We have developed metadata standards for the IMAGE project that incorporate a review of information available from genebanks and the standards adopted by other livestock projects such as FAANG. The standards include minimal requirements and enrichment through ontologies using EMBL-EBI's Zooma tool (http://www.ebi.ac.uk/spot/zooma/).

Coordination with the EUGENA portal will be ensured: information on the gene banks will be searchable at the IMAGE CDP. Each sample with a reference to a specific gene bank will provide basic information about the location and contacts for the bank but will also connect the user to the EUGENA portal for further information.

This integration of data will provide insight into the genetic make-up of the genebank material, adding to this the geographical dimension with its socio-economic, socio-demographic and climatic conditions, while connection to available phenotypic information will further add the production relevant aspect. The portal will achieve this integration by both importing and enriching data directly imported into its own internal database (the common data pool), but also through cross referencing each sample with the extensive phenotypic, GIS and gene bank data held in external resources. Providing access to genomics, GIS and phenotype data to genebank collections through a single unified access point will constitute a major achievement and great value to the collections from Europe.