IMAGE

IMAGE
Innovative Management of Animal Genetic Resources

European Project n° 677353

IMAGE data portal and tools for efficient access and use of information

Paolo Cozzi and Alessandra Stella

GenResBridge meeting 28/04/2021

The IMAGE WEB PORTAL

One of the main goals of IMAGE was the creation of a European web portal that integrates data from gene banks and collections with genomics data, geographical information systems data, and other information generated by IMAGE

NEED

- Integration of data from various sources
- Standardization
  - Consistency of ontologies
- Sustainability
- Visibility
- Ease of uploading
- Integration with existing DB

Exploit collections data

Gene banks collections are a backup of animal genetic diversity that are not fully exploited. Data are complex by nature

Standard definition

We need to unify access to gene bank data and create standards and protocols to exchange data

Information Integrations

Challenging task due to heterogeneity and large amount of data generated

IMAGE

IMAGE Conceptual Model

Ruleset to ensure comparable data

Collect, standardize and submit data

Data sustainability and persistence

Data integration

- Support for GIS
- Querying accross data

IMAGE PORTAL

- Step taken towards the development of such infrastructure were:
  - Selection of sustainable archives
  - Definition (and refinement) of rules and implementation of metadata standardization;
    - use of the ontologies to improve data quality and comparability
  - Construction of a custom Inject Tool that applies the ruleset to import data from gene banks, unifying units, terms and languages, and submitting the enhanced data to public EMBL-EBI BioSamples archive

IMAGE PORTAL

- Creation of the Common Data Pool to integrate:
  - Data from participating gene banks (archived in BioSamples)
  - IMAGE ‘omics datasets
  - External data (DAD-IS, EuGeNa)
- Tools to guide partners in the input of gene bank /collection data
- Creation and automation of tools to enhance use of data
  - Diversity browser
  - Breeder Interface

IMAGE Data model

- Simplified schematic of IMAGE data model.

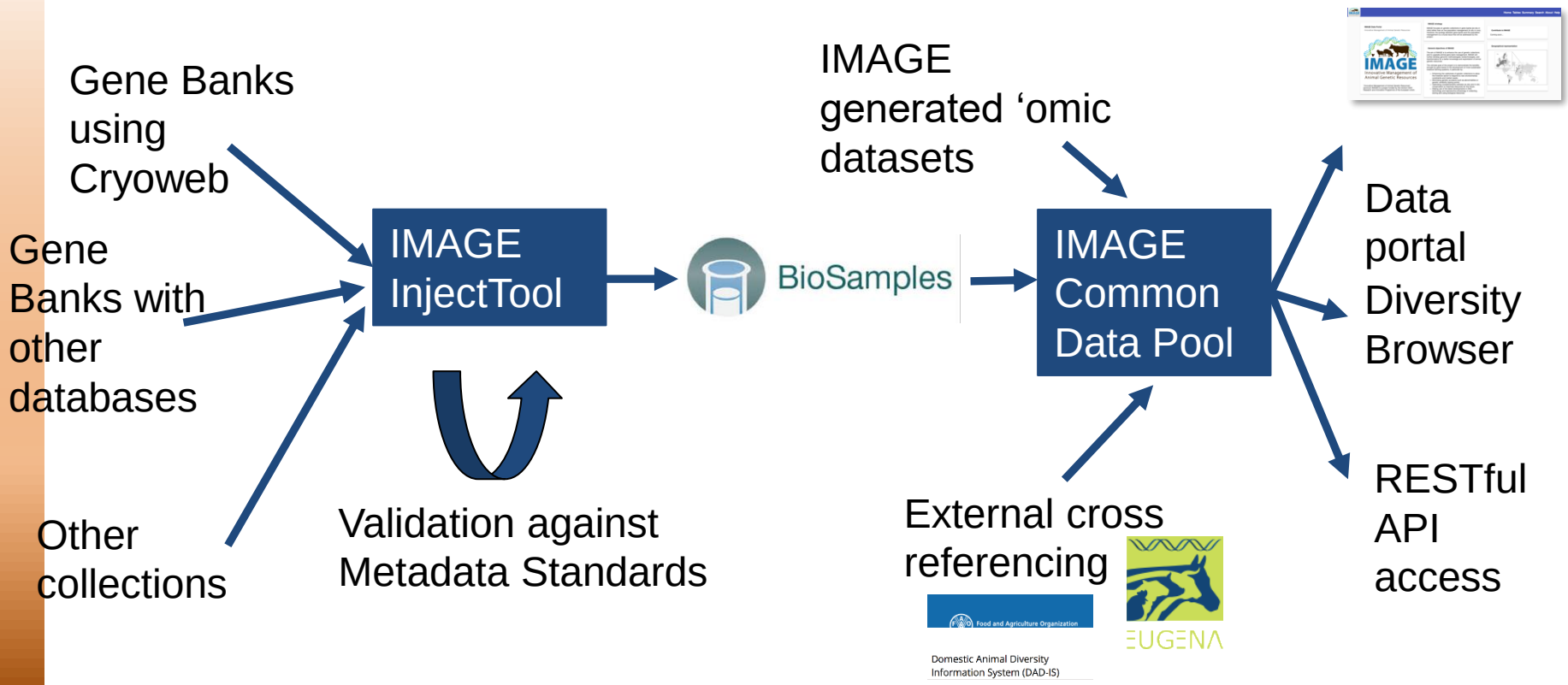
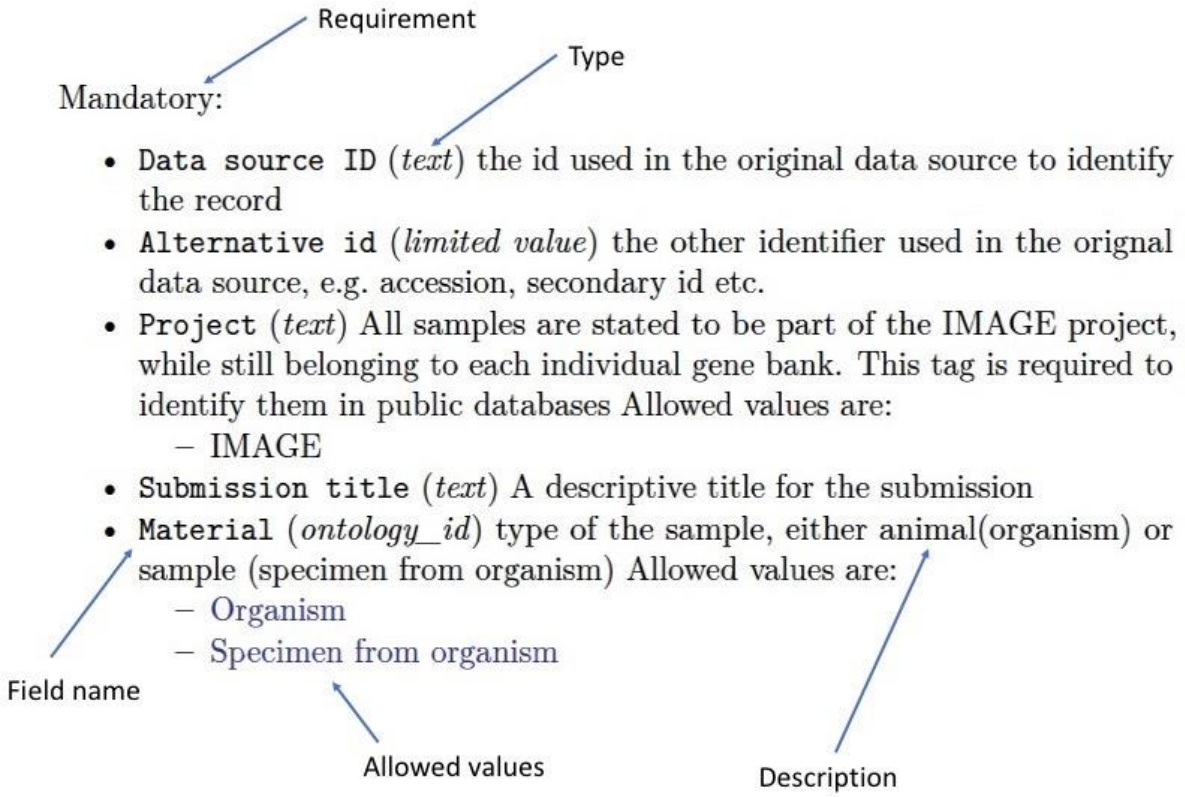


IMAGE gene bank metadata standardisation

- IMAGE metadata standards and data model
- IMAGE ruleset standardises metadata from diverse gene bank records
- Utilises ontologies (controlled dictionaries for biology) to ensure consistency in supplied terms and synonyms, and provide structure
- Metadata standards are not fixed and evolve with requirements of users
- https://github.com/cnr-ibba/IMAGE-metadata/tree/master/docs

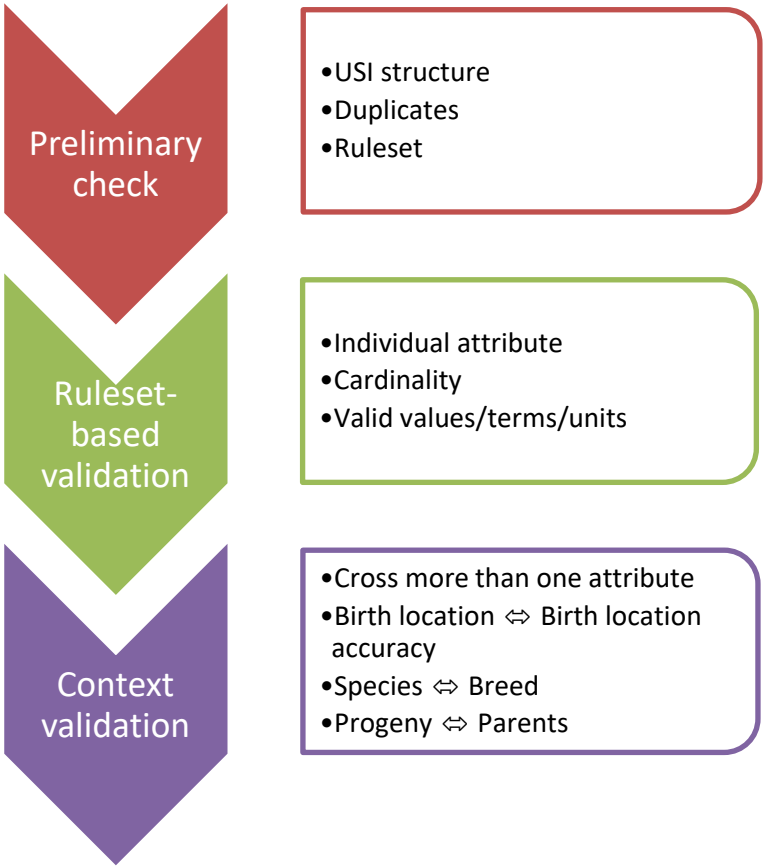
Metadata rules

- Mandatory
- Optional
- Suggested
- Enforce data types, requirements, allowed values
- To ensure high quality and comparability

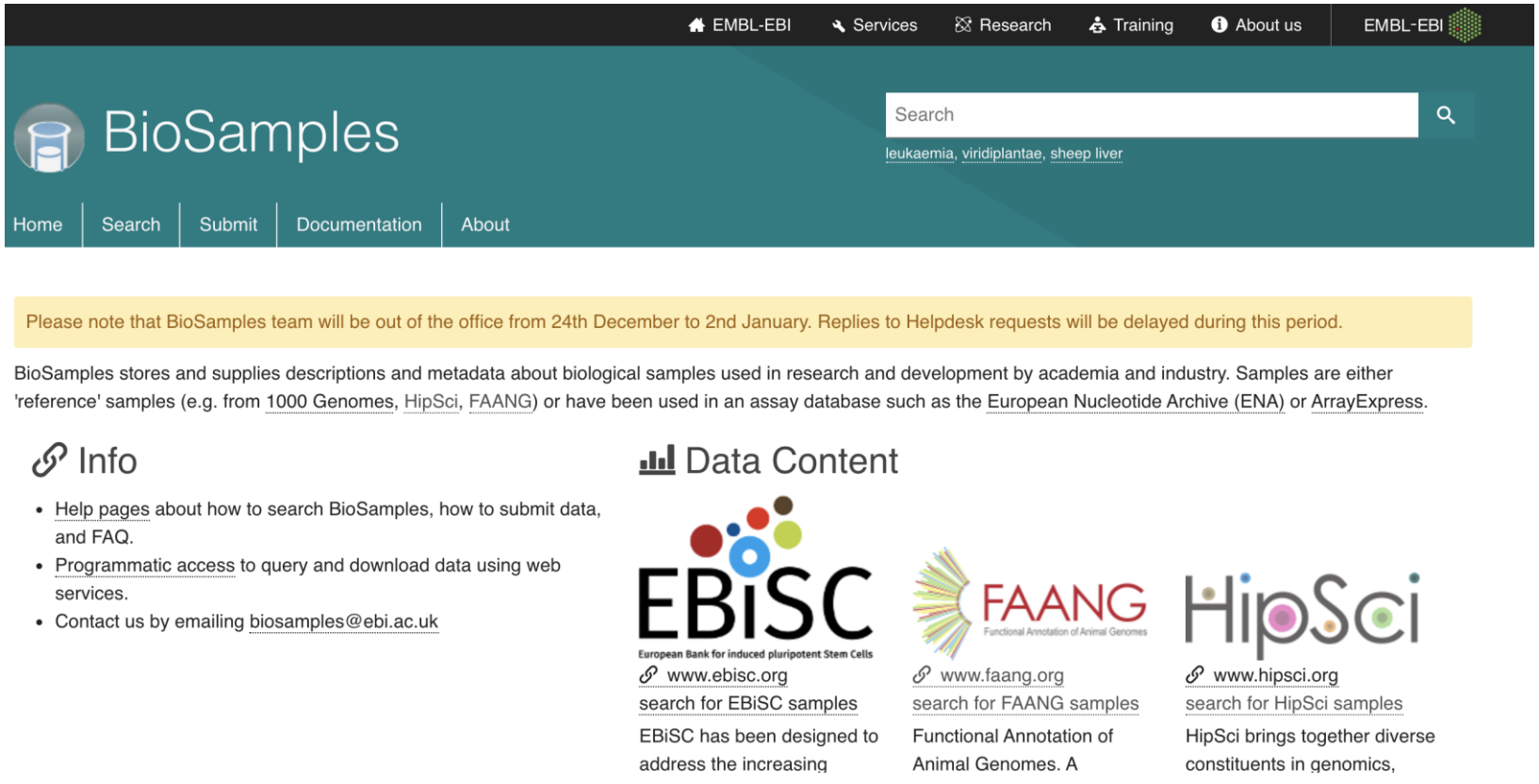


Validation

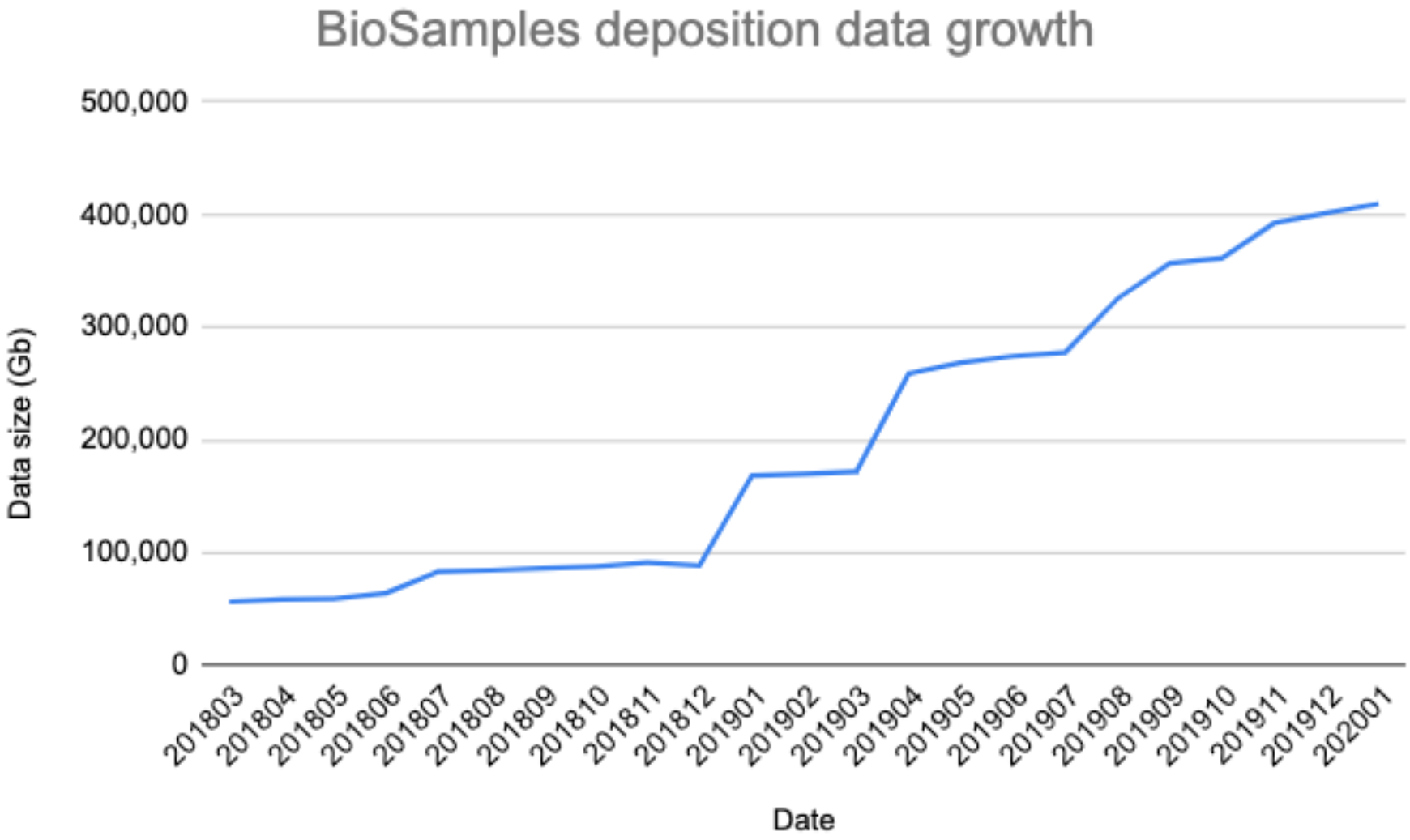
- Rule-based contextual validation
- Standardises language, terminology, units and ontologies
- Ensures minimum mandatory fields have been completed correctly
- Prevents errors and duplications
- Feeds back errors clearly for correction through Inject tool interface



BioSamples Database



IMAGE


BioSamples deposition data growth
Data size (Gb)
500,000
400,000
300,000
200,000
100,000
0
201803
201804
201805
201806
201807
201808
201809
201810
201811
201812
201901
201902
201903
201904
201905
201906
201907
201908
201909
201910
201911
201912
202001
Date

EMBL-EBI BioSamples archive

- Inject tool manages submission of IMAGE data to BioSamples archive
- BioSamples assigns universally recognized unique identifier to each organism and specimen e.g. SAMEA5159537
- Provides sustainable and secure long term storage of information
- Fully supports IMAGE metadata standards

Organism
SAMEA6265168

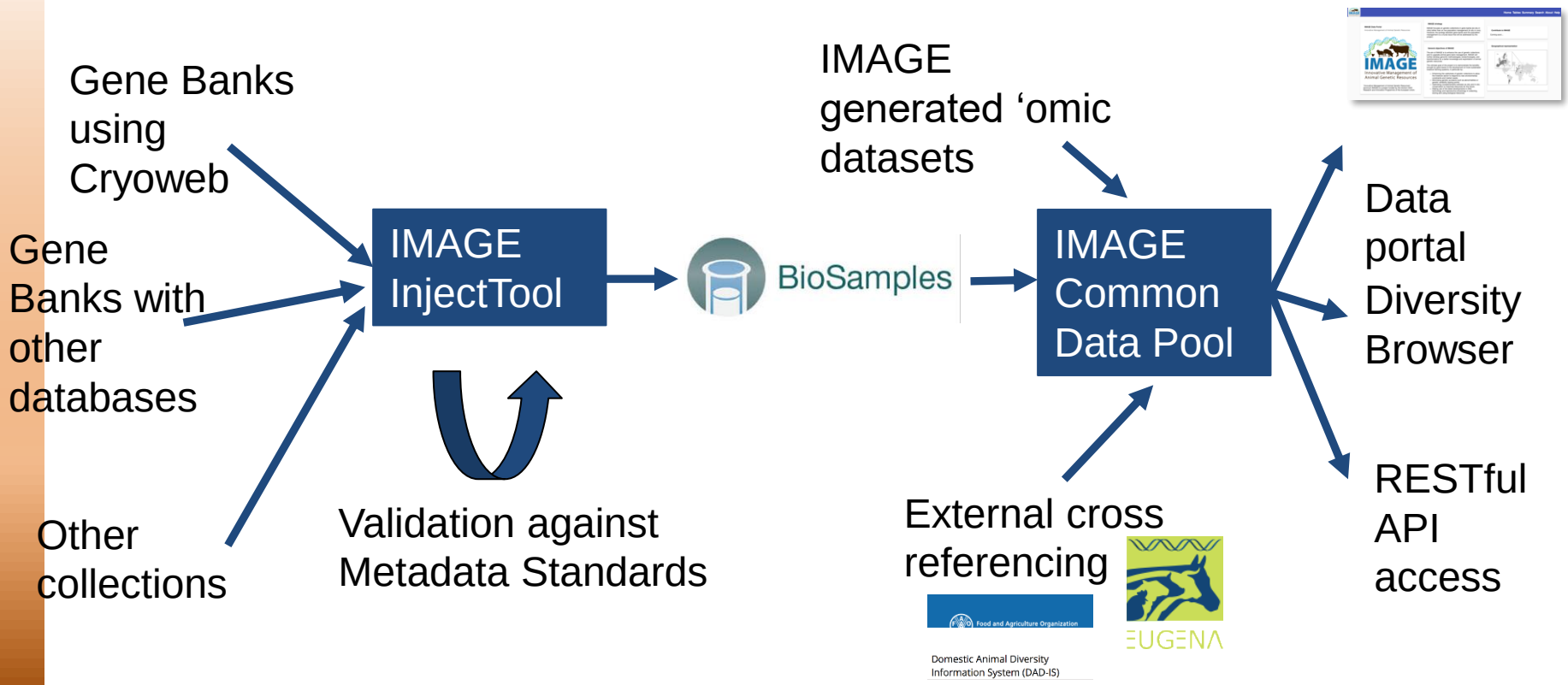
Multiple Specimens
SAMEA5079418
SAMEA103886757

Genotype Data to European Variation Archive
PRJEB24746

Photo: Peter Harrison

IMAGE Data model

- Simplified schematic of IMAGE data model.



Diversity Browser **Notebook**

- PLINK
- Python
- Building and testing on existing dataset
- Five consecutive steps

1) Run PLINK on reference dataset

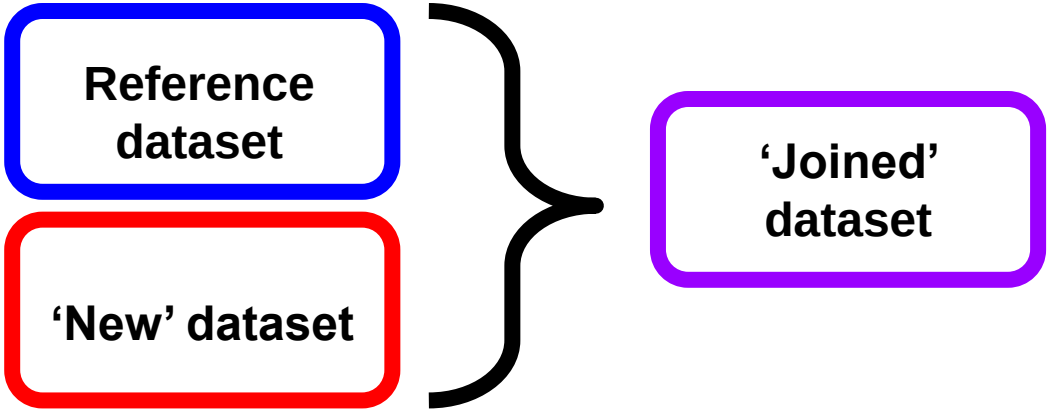
- Public dataset: https://doi.org/10.5061/dryad.30tk6
- Created subset of 623 animals

Table 1. Overview of pig reference dataset

Origin	# of Families	# of Animals
Asia	11	174
Asia Wild boar	2	10
Europe	24	290
Europe Wild boar	12	149

- Extracted the +/- 10,000 SNPs (which are on the *multispecies SNP-chip*)
  - Created the reference dataset

2) Join the reference dataset and the new dataset in PLINK



3) Run PCA in PLINK and generate output file with eigenvectors

- Principal Component Analysis (PCA)

  Statistical procedure

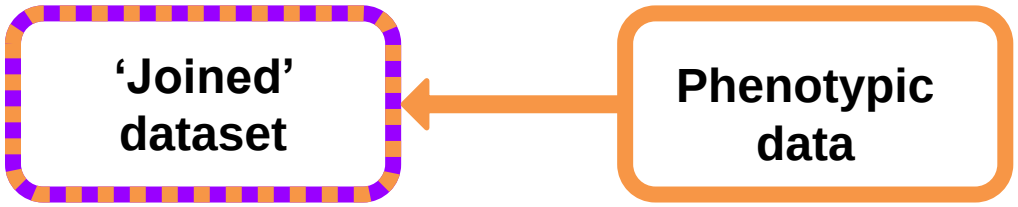
  Orthogonal transformation

- Used for:

  Exploratory data analysis

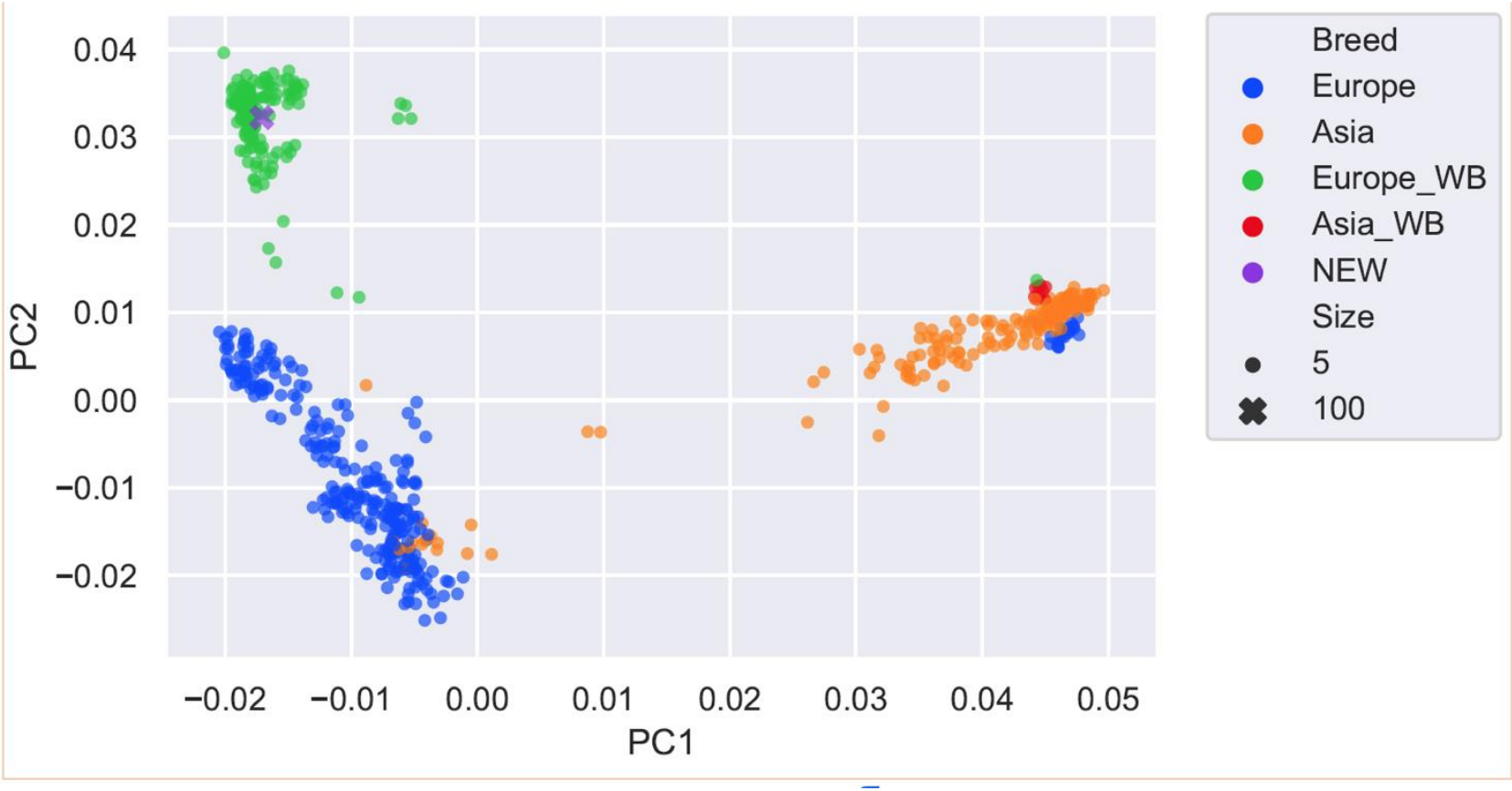
  Making predictive models

4) Enrich the PCA output file with phenotypic data



- Species; pig, poultry, cattle, ...
- Breed; Landrace, Duroc, ...
- Origin; geographical location

Diversity Browser



What went well vs what went so and so

- IMAGE Web Portal up and running
  - InjectTool aids in submissions
    - Still assistance is needed
  - > 36k samples uploaded
    - 18k animals
- Issues of privacy of data
    - Who is in charge of decision to submit
- Delays with submission of genomic data
  - Privacy issues with genotype data

What went well vs what went so and so

- GIS data scarce
  - Stimulate interest in the usefulness of geoclimatic infos
- Alignment issues among IMAGE common data pool and Biosample and other EBI archives
  - Need to be solved by opening specific tickets with different references
- Portal management and curation

IMAGE

Jun Fan
Alexey Sokolov
Peter Harrison

EMBL-EBI

Paolo Cozzi
Eildert Groeneveld
Ale Stella

UNIVERSITAS CATHOLICA SACRI CORDIS MEDIOLANI

Sipke Hiemstra

Sylvain Marthey

INRAE
science for people, life & earth

Oliver Selmoni
Elia Vajana
Stephane Joost

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

DirkJan Schokker
Richard Croijmans
Rayner Gonzalez Prendes

WAGENINGEN
UNIVERSITY & RESEARCH