Tools to manage trial, phenotyping and marker datasets: FAIRness in the Legume Generation consortium

James Brett

Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UZ, UK
www.earlham.ac.uk

Biotechnology and Biological Sciences Research Council

norwich research park

Legume Generation

Earlham Institute
Decoding Living Systems

Legume Generation

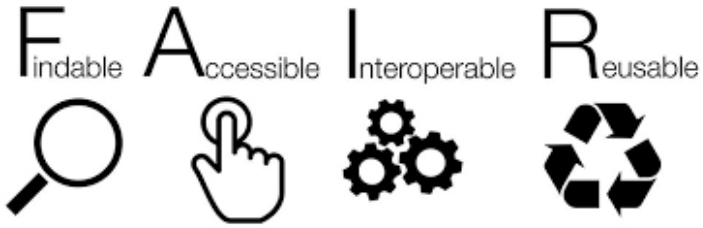
Project Consortium partners

IPK LEIBNIZ INSTITUTE
Euroseeds Embracing Nature
DUNAV SOJA
SERIDA
UNIVERSITÀ POLITECNICA DELLE MARCHE
RAGT 2n
UNIVERSITY OF HOHENHEIM
1880 DANKO
Donal Murphy-Bokern (DMB)
John Innes Centre
IGR pan
JKI Julius Kühn-Institut
Radboud Universiteit
agresearch āta mātai, mātai whetū
Wellspring
Germinal
Earlham Institute
ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ
Agroscope
SZG
AARHUS UNIVERSITY
Lidea
ABI
SaatGut Keyserlingk-Institut
PRIFYSGOL ABERYSTWYTH UNIVERSITY
Ideen um Pflanzen EUKUSA Ideen pflanzen
KWS
UNIVERZITA PALACKÉHO V OLOMOUCI
BOKU
LfL Bayerische Landesanstalt für Landwirtschaft
USDA


Not on
the map:
USDA
AGR
Soya bean
Pea
Lupins
Lentil
Phaseolus beans
Clovers
AU
GER
ABER
JIC
EI
IPP
SRU
DMB
KWS
JKI
IPK
IPG
DANKO
EURS
ESKUSA
UP
UHOH
Keyserlingk
LfL
DS
BOKU
SZG
AGS
LIDEA
SERIDA
RAGT
UNIVPM
ABI
IHU
UNIBAS

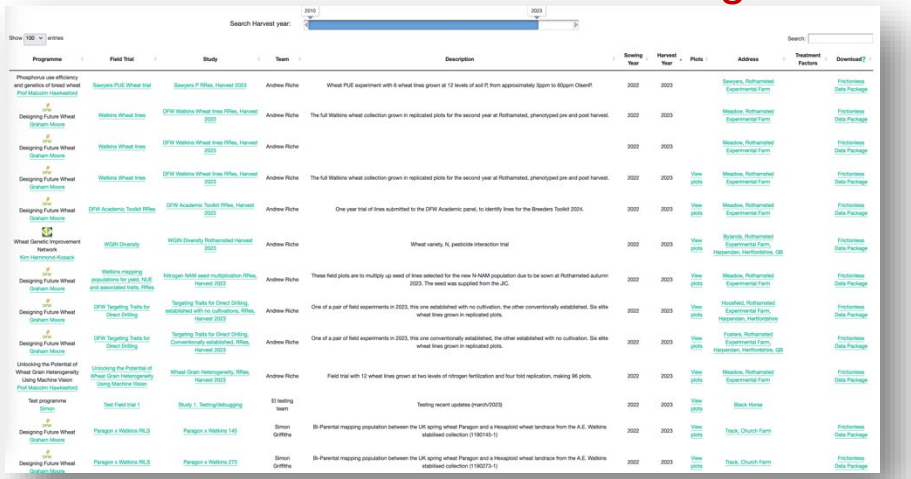
Trial and marker data management:

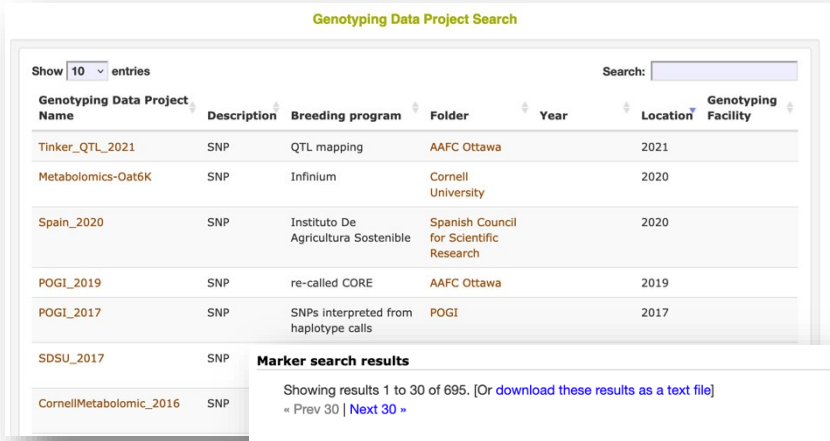
"Data access plays a crucial role in aiding breeders to make informed decisions in their breeding programmes, which results in improved and accelerated crop improvement."

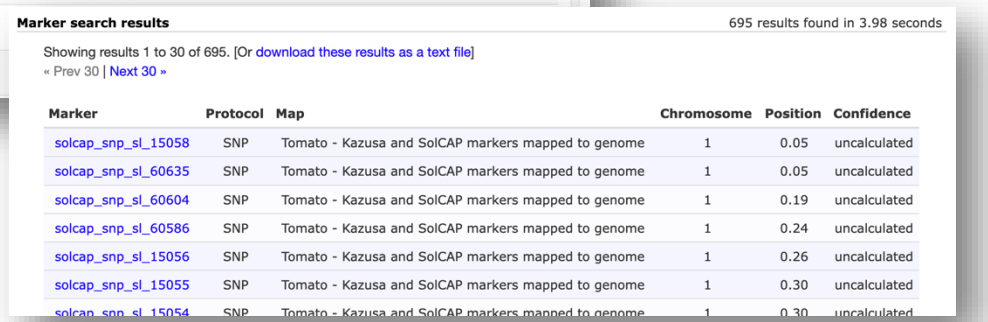


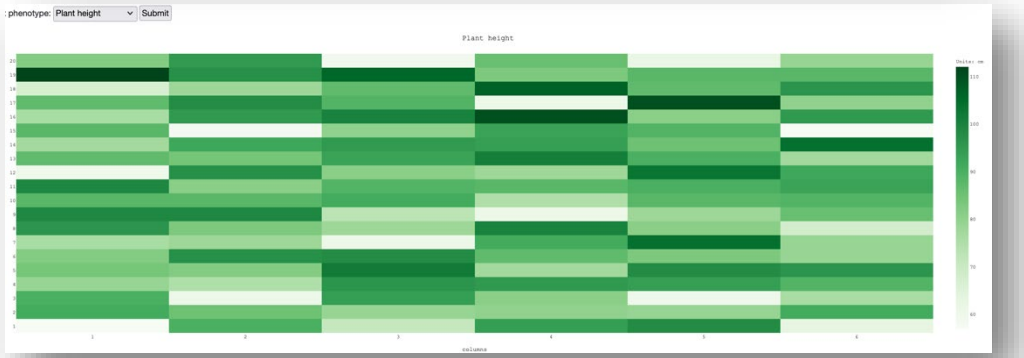
"…extracting useful knowledge requires integration and contextualisation"

the "Legume Generation Knowledge centre"









Data problems

- Increased computational capacity is leading to more data getting produced
- **Complex problem to store and make this data accessible**
- How to make it appropriate for increasing rise of machine learning?

Solutions for data storage – phenotyping and field trials

Here at EI we tackle this problem by developing software and services to address these challenges using FAIR principles

- Grassroots

Dr Simon Tyrrell

Wheat Initiative Software Engineer, Data Science Group

FAIR data principles - Findable

The first step in (re)using data is to find them.

- Data are described with rich metadata
- Metadata and data should be easy to find for both humans and computers.
- Machine-readable metadata are essential for automatic discovery of datasets and services

Taken from https://www.go-fair.org/fair-principles/

FAIR data principles - Accessible

Once the user finds the required data, they need to know how can they be accessed, possibly including authentication and authorization.

- (Meta)data are retrievable by their identifier using a standardized communications protocol
- Metadata are accessible, even when the data are no longer available

Taken from https://www.go-fair.org/fair-principles/

FAIR data principles - Interoperable

The data usually need to be integrated with other data.

- Able to be easily integrated with applications or workflows for analysis, storage, and processing.
- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Taken from https://www.go-fair.org/fair-principles/

FAIR data principles - Reusable

The ultimate goal of FAIR is to optimize the reuse of data.

- Metadata and data should be well-described so that they can be replicated and/or combined in different settings
- Metadata and data are associated with detailed provenance

Taken from https://www.go-fair.org/fair-principles/

Grassroots

- Grassroots infrastructure is a lightweight architecture to share both distributed data and services across multiple servers.

- The scientific functionality of the Grassroots is provided by services
  - **Field Trials**
  - Parental Genotypes
  - Field Pathogenomics
  - Blast
  - SamTools

Field Trial Experiments

Experiments where different crops are planted in plots within a field, differing treatments applied and then traits are measured.

- **Standardised template for submitting the genotype (the genetic material of the crop) and the phenotype (the characteristics that you want to measure) data**
- To facilitate publishing of data compliant with FAIR sharing principles

Field Trial Experiments

Experiments where different crops are planted in plots within a field, differing treatments applied and then traits are measured.

- Standardised template for submitting the genotype (the genetic material of the crop) and the phenotype (the characteristics that you want to measure) data
- **To facilitate publishing of data compliant with FAIR sharing principles**
- Main DFW/DSW goal is to make all data openly available

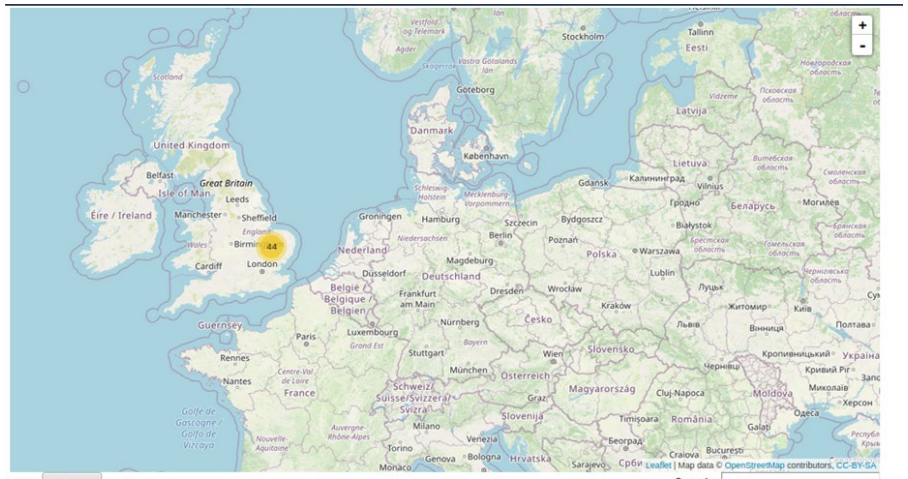
Grassroots - Programmes

- The high level organization or group that is responsible for conducting trials and studies.
  - Designing Future Wheat, Delivering Sustainable Wheat, Wheat Genetic Information Network (WGIN), BBSRC Low Protein Wheat, **Legume Generation**.

- Metadata such as
  - Name
  - Crop
  - Objective
  - Principal Investigator
  - Etc.

Grassroots - Findable

The experimental data can be accessed using a map-based view and a searchable table of the data...

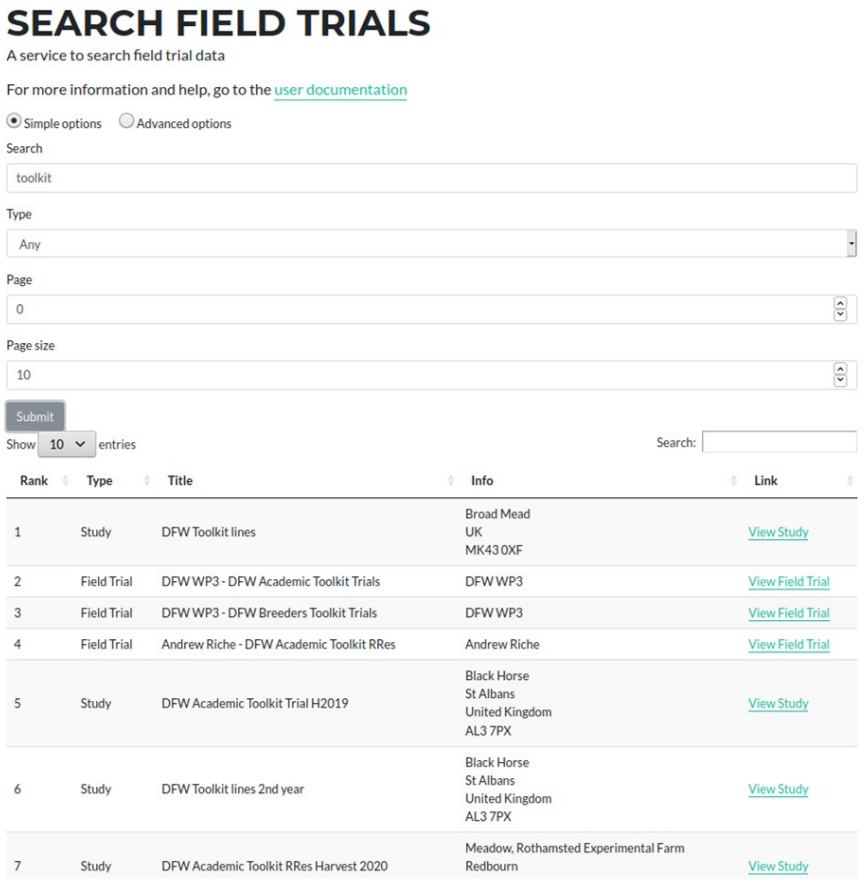


Show 10 entries Search:

Name	Team	Description	Sowing Date	Harvest Date	Plots	Address	Popup Info	Links
Drilling Date Robigus x Claire Experiment	Simon Griffiths						Study Info	• Study
DFW TKNIL Set 2	Simon Griffiths		2017-10-11	2018-08-04	View plots	Coppers Bawburgh GB NR9 3QL	Study Info	• Study • Field Trial
DFW Field Phenotyping Facility - Wheat	Rothamsted Research					Great Field 1/2 Phenotyping Area St Albans UK AL5 2GT	Study Info	• Study • Field Trial

Grassroots - Findable

…or via a text-based search web page…



SEARCH FIELD TRIALS

A service to search field trial data

For more information and help, go to the user documentation

Simple options Advanced options

Search

toolkit

Type

Any

Page

0

Page size

10

Submit

Show 10 entries Search:

Rank	Type	Title	Info	Link
1	Study	DFW Toolkit lines	Broad Mead UK MK43 0XF	View Study
2	Field Trial	DFW WP3 - DFW Academic Toolkit Trials	DFW WP3	View Field Trial
3	Field Trial	DFW WP3 - DFW Breeders Toolkit Trials	DFW WP3	View Field Trial
4	Field Trial	Andrew Riche - DFW Academic Toolkit RRes	Andrew Riche	View Field Trial
5	Study	DFW Academic Toolkit Trial H2019	Black Horse St Albans United Kingdom AL3 7PX	View Study
6	Study	DFW Toolkit lines 2nd year	Black Horse St Albans United Kingdom AL3 7PX	View Study
7	Study	DFW Academic Toolkit RRes Harvest 2020	Meadow, Rothamsted Experimental Farm Redbourn	View Study

Grassroots - Findable

Innovate 2023

…or programmatically

- Curl
- Python
- C/C++
- R
- etc.

JSON request to run a given service

```
{
  "services": [{
    "so:name": "Search Grassroots",
      "start_service": true,
      "parameter_set": {
        "parameters": [{
          "param": "SS Keyword Search",
          "current_value": "Paragon"
        }]
      }
  }]
}
```

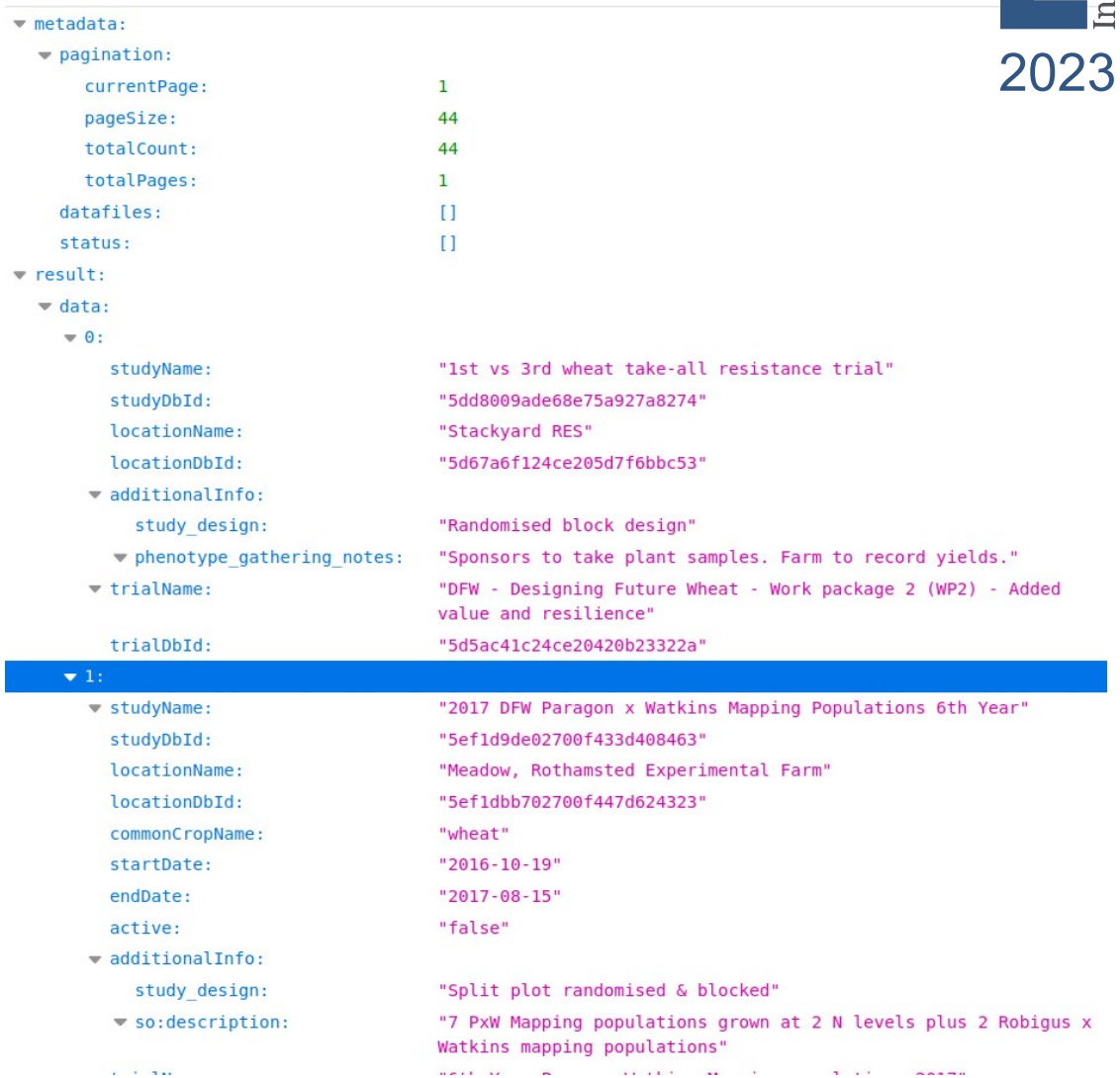
Grassroots - Accessible

- All data is openly available
- All Field Trials, Studies, etc. have a unique identifier and are accessible through standard web technologies

Grassroots - Interoperability

The Field Trials data and metadata is exposed using many APIs such as

- Grassroots
- BrAPI which is a community-driven standardized RESTful web service API specification to enable interoperability among plant breeding databases.
- Frictionless Data
  - Schemas published at https://grassroots.tools/frictionless-data/
- CSV files



Grassroots - Reusability

- Using as many standard ontologies as possible
  - Schema.org
  - Crop Ontology
  - Plant Experimental Conditions Ontology
  - Environment Ontology
  - Software Ontology
  - Agronomy Ontology
- Custom ontological terms that will be submitted to the Crop Ontology

Grassroots - Reusable data

Innovate 2023

Plot data is standardized using ontological terms for each plot

PLOT DETAILS ×

Row: 20
Column: 1
Length: 3.594m
Width: 1.8m
Study Design:
Sowing Date: 2019-10-30
Harvest Date: 2020-08-10
Treatment:
Comment: Slight height segregation

Replicate	Rack	Accession	Pedigree	Gene Bank	Links
1 (Current Plot)	1	DFW SEL 0208		Germplasm Resources Unit	
3 (Plot Row:3 - Col:23)	1	DFW SEL 0208		Germplasm Resources Unit	
2 (Plot Row:14 - Col:15)	1	DFW SEL 0208		Germplasm Resources Unit	

PHENOTYPES

Close

Earlham Institute

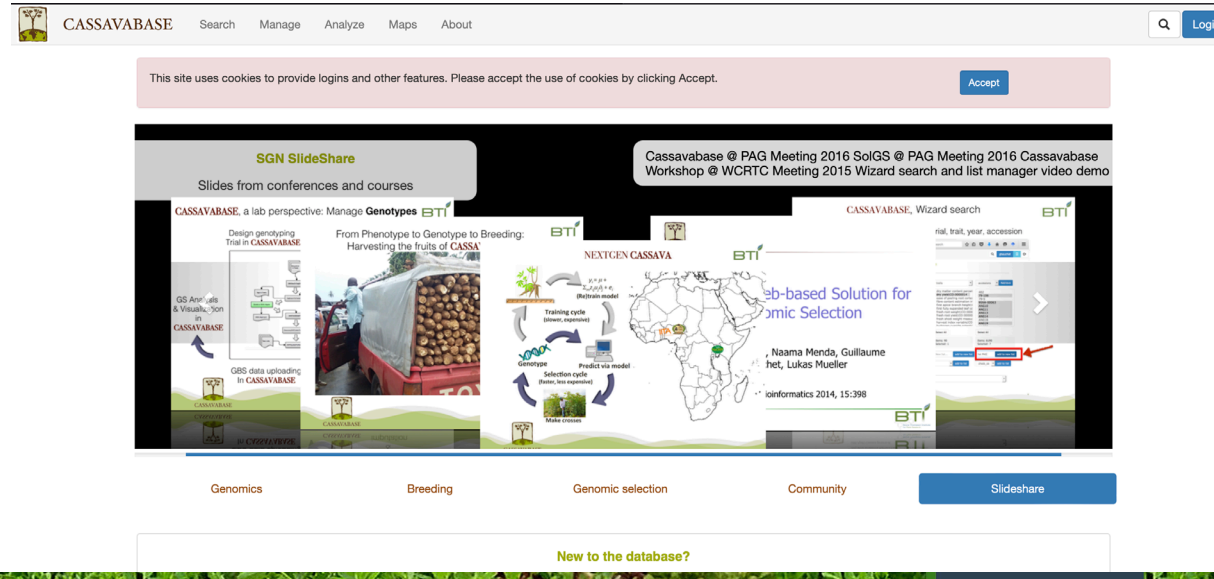
Databases Storing Genetic Data

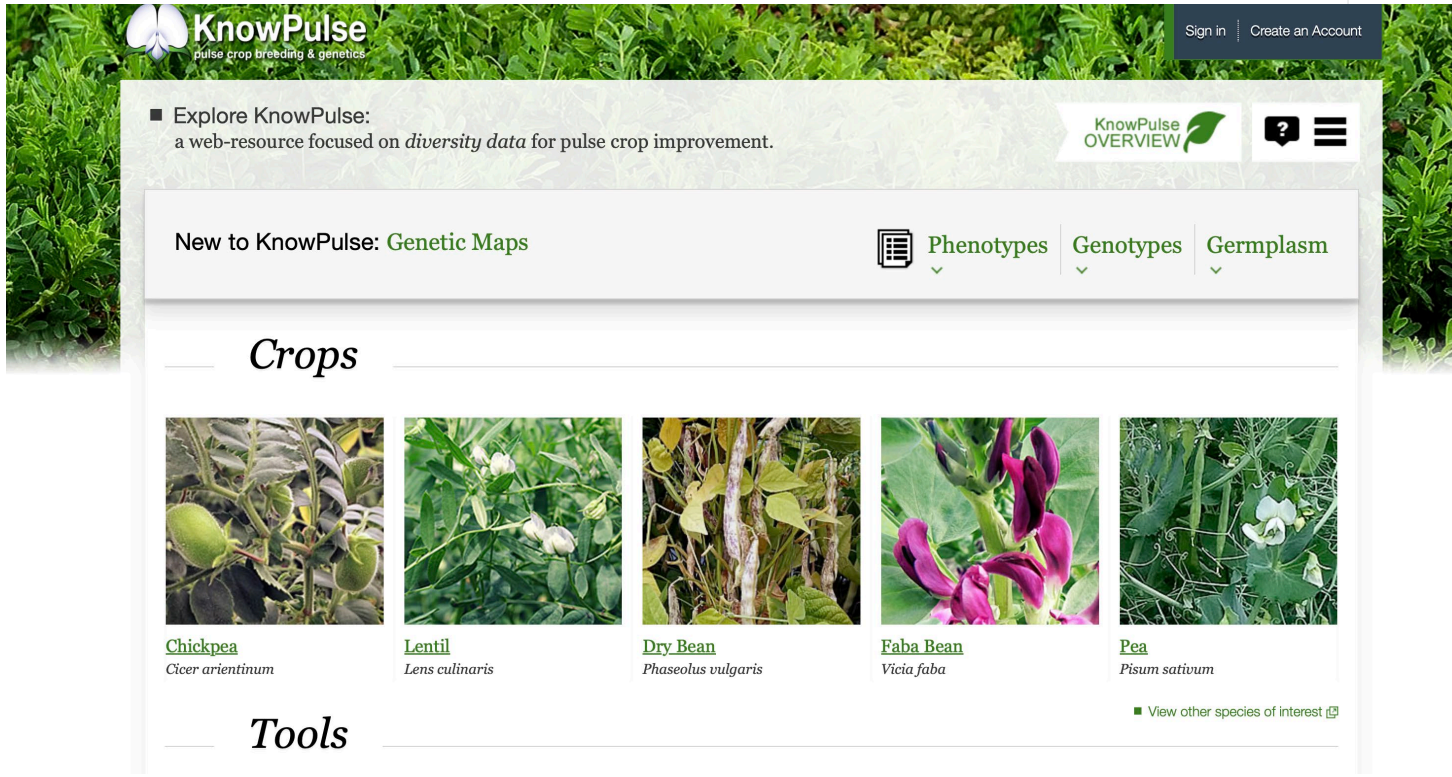
- Still deciding which database software to work with.

Two that we are considering and have already been used successfully to house genetic information on various crops:

- Breedbase -> https://cassavabase.org
- Tripal3 -> https://knowpulse.usask.ca

Integrate with Grassroots database and can perform downstream analyses on data e.g. BLUPS/BLUEs and genomic selection.





Acknowledgements

Earlham Institute

- Jose De-Vega
- Simon Tyrell
- Daniel Olvera
- Robert Davey
- Xingdong Bian
- Nicola Soranzo
- Felix Shaw
- Catherine Hunter
- Anil Thanki
- Alice Minotto
- Jazz Urog

John Innes Centre

- Luzie Wingen
- Simon Griffiths

Rothamsted Research

- Andrew Riche
- Chris Rawlings
- Richard Ostler

University of Bristol

- Paul Wilkinson
- Mark Winfield
- Gary Barker

James Brett

Legume Breeding Data Manager

James.brett@earlham.ac.uk

Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UZ, UK
www.earlham.ac.uk