

# 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](http://www.earlham.ac.uk)



Decoding Living Systems



Legume  
Generation

# Project Consortium partners



Donal Murphy-Bokern  
(DMB)



Radboud Universiteit

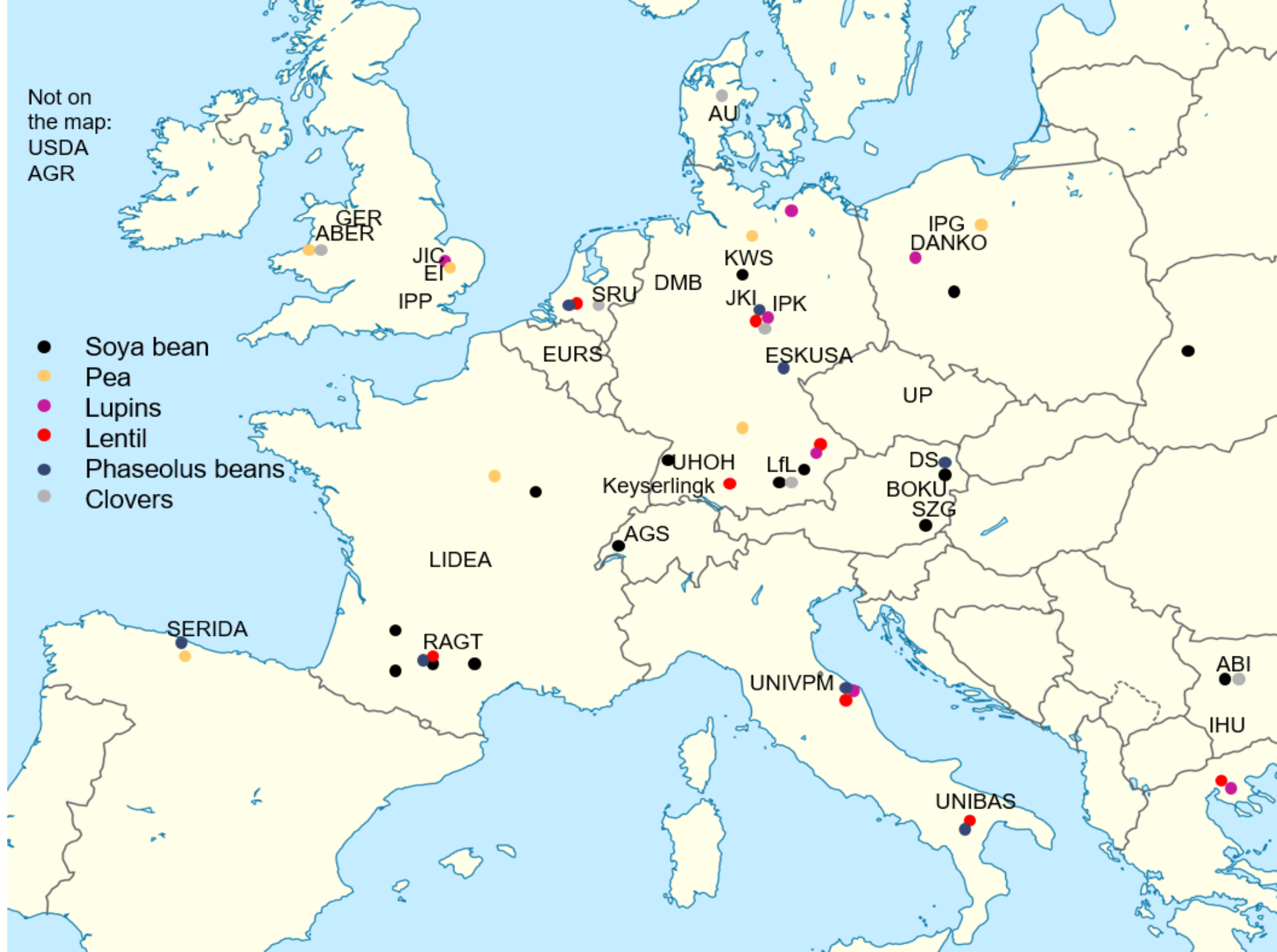


Agroscope



Not on  
the map:  
USDA  
AGR

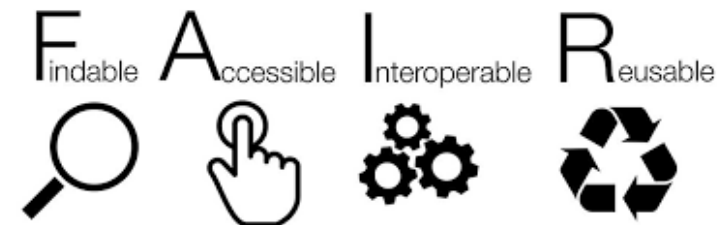
- Soya bean
- Pea
- Lupins
- Lentil
- Phaseolus beans
- Clovers



# 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”

Programme	Field Trial	Study	Team	Description	Breeding Year	Harvest Year	Plots	Address	Treatment Factors	Download?
Phosphorus use efficiency and genetic of bread wheat (Prof. Malcolm Haslam)	Seaxons-PH2 wheat trial	Seaxons P-Phos, Harvest 2023	Andrew Ritchie	Wheat PH2 experiment with 6 wheat lines grown at 12 levels of soil P from approximately 1ppm to 10ppm (Oxley)	2020	2023	View plots	Seaxons, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	The full Widdows wheat collection grown in replicated plots for the second year of Rothamsted, phenotyped pre and post harvest.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	The full Widdows wheat collection grown in replicated plots for the second year of Rothamsted, phenotyped pre and post harvest.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	The full Widdows wheat collection grown in replicated plots for the second year of Rothamsted, phenotyped pre and post harvest.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	The full Widdows wheat collection grown in replicated plots for the second year of Rothamsted, phenotyped pre and post harvest.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	The full Widdows wheat collection grown in replicated plots for the second year of Rothamsted, phenotyped pre and post harvest.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Wheat Genetic Improvement Network (Ken Heckmond-Graham)	WGH Diversity	WGH Diversity Rothamsted Harvest 2023	Andrew Ritchie	Wheat diversity, N <sub>2</sub> fertilizer interaction trial	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	Three field plots are to multiply up seed of lines selected for the new N <sub>2</sub> NRM population due to be sown at Rothamsted autumn 2023. The seed was supplied from the JIC.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	One of a pair of field experiments in 2023, this one established with no cultivation, the other conventionally established. Six sites wheat lines grown in replicated plots.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	One of a pair of field experiments in 2023, this one conventionally established, the other established with no cultivation. Six sites wheat lines grown in replicated plots.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Unlocking the Potential of Wheat Genes Using Machine Vision and Genetic Characterisation (Test programme Simon)	Test Field trial 1	Study 1, Test/Genotyping	Simon Goffree	Testing recent varieties (2020-2023)	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	Field trial with 12 wheat lines grown at two levels of nitrogen fertilization and four field replicates, making 96 plots.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package
Deepening Future Wheat (Graham Moore)	Widdows Wheat lines	DFW Widdows Wheat lines 1916a, Harvest 2023	Andrew Ritchie	Field trial with 12 wheat lines grown at two levels of nitrogen fertilization and four field replicates, making 96 plots.	2020	2023	View plots	Widdows, Rothamsted Experimental Farm	Phosphorus	Fetch Data Package

### Genotyping Data Project Search

Show 10 entries

Search:

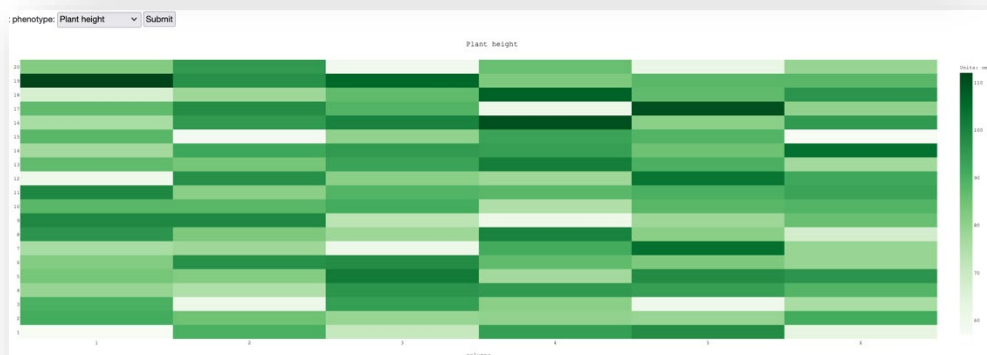
Genotyping Data Project Name	Description	Breeding program	Folder	Year	Location	Genotyping Facility
Tinker_QTL_2021	SNP	QTL mapping	AAFC Ottawa	2021		
Metabolomics-Oat6K	SNP	Infinium	Cornell University	2020		
Spain_2020	SNP	Instituto De Agricultura Sostenible	Spanish Council for Scientific Research	2020		
POGI_2019	SNP	re-called CORE	AAFC Ottawa	2019		
POGI_2017	SNP	SNPs interpreted from haplotype calls	POGI	2017		
SDSU_2017	SNP					
CornellMetabolomic_2016	SNP					

695 results found in 3.98 seconds

Showing results 1 to 30 of 695. [Or download these results as a text file]

« Prev 30 | Next 30 »

Marker	Protocol	Map	Chromosome	Position	Confidence
<a href="#">solcap_snp_sl_15058</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.05	uncalculated
<a href="#">solcap_snp_sl_60635</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.05	uncalculated
<a href="#">solcap_snp_sl_60604</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.19	uncalculated
<a href="#">solcap_snp_sl_60586</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.24	uncalculated
<a href="#">solcap_snp_sl_15056</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.26	uncalculated
<a href="#">solcap_snp_sl_15055</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.30	uncalculated
<a href="#">solcap_snp_sl_15054</a>	SNP	Tomato - Kazusa and SolCAP markers mapped to genome	1	0.30	uncalculated



## 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.

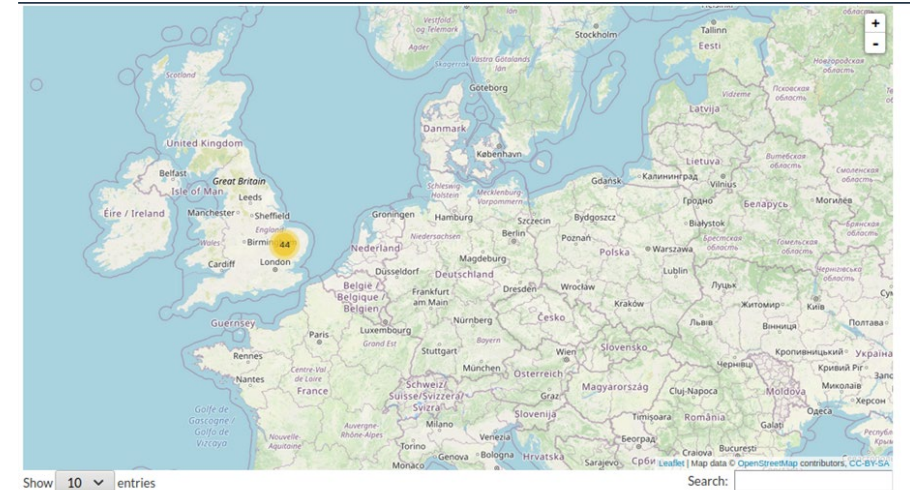
- Standardised template for submitting the genotype (the genetic material of the crop) and the phenotype (the characteristics that you want to measure) data
- **To facilitate publishing of data compliant with FAIR sharing principles**
- Main DFW/DSW goal is to make all data openly available

# 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

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



# 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	<a href="#">View Study</a>
2	Field Trial	DFW WP3 - DFW Academic Toolkit Trials	DFW WP3	<a href="#">View Field Trial</a>
3	Field Trial	DFW WP3 - DFW Breeders Toolkit Trials	DFW WP3	<a href="#">View Field Trial</a>
4	Field Trial	Andrew Riche - DFW Academic Toolkit RRes	Andrew Riche	<a href="#">View Field Trial</a>
5	Study	DFW Academic Toolkit Trial H2019	Black Horse St Albans United Kingdom AL3 7PX	<a href="#">View Study</a>
6	Study	DFW Toolkit lines 2nd year	Black Horse St Albans United Kingdom AL3 7PX	<a href="#">View Study</a>
7	Study	DFW Academic Toolkit RRes Harvest 2020	Meadow, Rothamsted Experimental Farm Redbourn	<a href="#">View Study</a>

# Grassroots - Findable

...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

```
▼ metadata:
  ▼ pagination:
    currentPage: 1
    pageSize: 44
    totalCount: 44
    totalPages: 1
    datafiles: []
    status: []
  ▼ result:
    ▼ data:
      ▼ 0:
        studyName: "1st vs 3rd wheat take-all resistance trial"
        studyDbId: "5dd8009ade68e75a927a8274"
        locationName: "Stackyard RES"
        locationDbId: "5d67a6f124ce205d7f6bbc53"
        ▼ additionalInfo:
          study_design: "Randomised block design"
          ▼ phenotype_gathering_notes: "Sponsors to take plant samples. Farm to record yields."
          ▼ trialName: "DFW - Designing Future Wheat - Work package 2 (WP2) - Added value and resilience"
          trialDbId: "5d5ac41c24ce20420b23322a"
      ▼ 1:
        studyName: "2017 DFW Paragon x Watkins Mapping Populations 6th Year"
        studyDbId: "5ef1d9de02700f433d408463"
        locationName: "Meadow, Rothamsted Experimental Farm"
        locationDbId: "5ef1dbb702700f447d624323"
        commonCropName: "wheat"
        startDate: "2016-10-19"
        endDate: "2017-08-15"
        active: "false"
        ▼ additionalInfo:
          study_design: "Split plot randomised & blocked"
          ▼ so:description: "7 PxW Mapping populations grown at 2 N levels plus 2 Robigus x Watkins mapping populations"
```

# 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

Plot data is standardized using ontological terms for each plot

## PLOT DETAILS

x

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			<a href="#">Germplasm Resources Unit</a>
3 (Plot Row:3 - Col:23)	1	DFW SEL 0208			<a href="#">Germplasm Resources Unit</a>
2 (Plot Row:14 - Col:15)	1	DFW SEL 0208			<a href="#">Germplasm Resources Unit</a>

## PHENOTYPES

Close



# 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.

The screenshot shows the CASSAVABASE website. At the top, there is a navigation bar with the logo, the name 'CASSAVABASE', and links for 'Search', 'Manage', 'Analyze', 'Maps', and 'About'. A cookie consent banner is visible below the navigation bar. The main content area features a 'SGN SlideShare' section with a title 'Slides from conferences and courses'. Below this, there are several slide thumbnails. One slide is titled 'CASSAVABASE, a lab perspective: Manage Genotypes' and another is 'From Phenotype to Genotype to Breeding: Harvesting the fruits of CASSA'. A third slide is titled 'Web-based Solution for Genomic Selection' and mentions 'Naama Menda, Guillaume het, Lukas Mueller'. At the bottom of the slide thumbnails, there are navigation tabs for 'Genomics', 'Breeding', 'Genomic selection', and 'Community', along with a 'Slideshare' button. A 'New to the database?' link is also present.

The screenshot shows the KnowPulse website. The header features the 'KnowPulse' logo with the tagline 'pulse crop breeding & genetics' and a background image of green plants. On the right side of the header, there are links for 'Sign in' and 'Create an Account'. Below the header, there is a main navigation area with a 'New to KnowPulse: Genetic Maps' section and three dropdown menus for 'Phenotypes', 'Genotypes', and 'Germplasm'. The main content area is titled 'Crops' and displays five crop cards, each with a photograph and the crop name and scientific name: Chickpea (*Cicer arorientinum*), Lentil (*Lens culinaris*), Dry Bean (*Phaseolus vulgaris*), Faba Bean (*Vicia faba*), and Pea (*Pisum sativum*). At the bottom, there is a 'Tools' section and a link to 'View other species of interest'.



# 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](http://www.earlham.ac.uk)



Decoding Living Systems