Frictionless Data for iRODS

SIMON TYRRELL
RESEARCH SOFTWARE ENGINEER

XINGDONG BIAN
RESEARCH SOFTWARE ENGINEER

ROB DAVEY
GROUP LEADER

Earlham Institute

Decoding Living Systems

Designing Future Wheat (DFW)

The BBSRC funded Designing Future Wheat (DFW) Institute Strategic Programme, spans over 25 groups of scientists across 8 research institutes and universities and aims to develop new wheat germplasm containing the next generation of key traits.

It is anticipated that the world will need to produce 60% more wheat by 2050 to meet global demand. Since it takes between 15 and 20 years for current research to improve wheat varieties grown in farmers' fields, it is imperative that we act now to address problems facing us in the future

Taken from https://designingfuturewheat.org.uk/

DFW Data

DFW produces lots of scientific data

- Field Trial experiments
- Datasets
- Sequences

Target audiences

We have different groups of users

- Breeders
- Academics
- Data Scientists
- Industry

Challenge

To make the data accessible and usable for everyone

Grassroots Infrastructure

A suite that wraps up industry-standard software tools along with our own custom open-source ones

- Consistent JSON-based API
  - Language and platform agnostic
- Can be federated with other Grassroots instances
- Sharing data and services in a FAIR way

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/

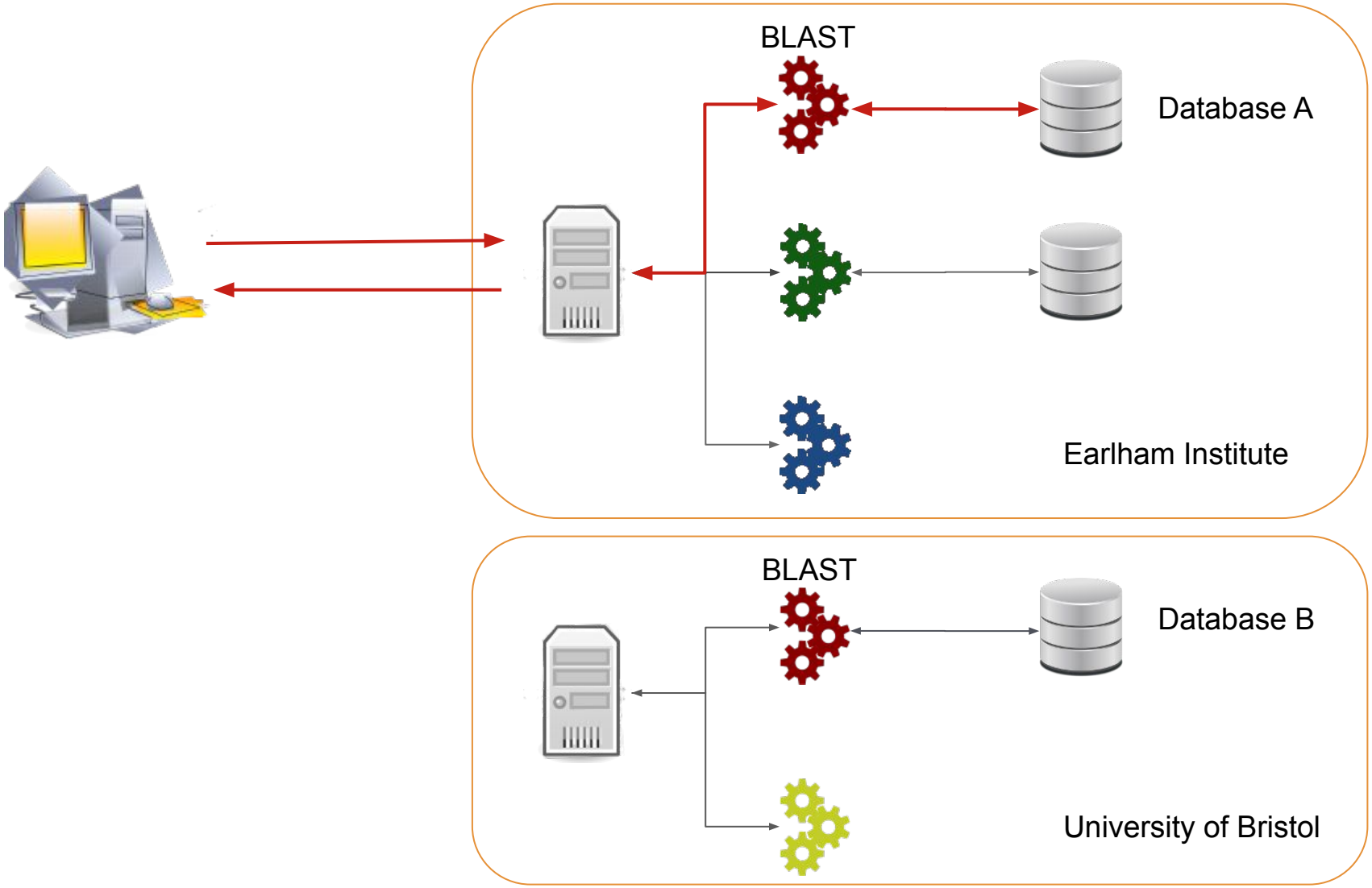
Grassroots Infrastructure

- Grassroots Apache module acts as a bridge between Apache and Grassroots
- A set of cross-platform libraries that can be used by Grassroots components including
  - Networking code to access web resources
  - Server and Service management tools
  - API to & from our web services and their parameters

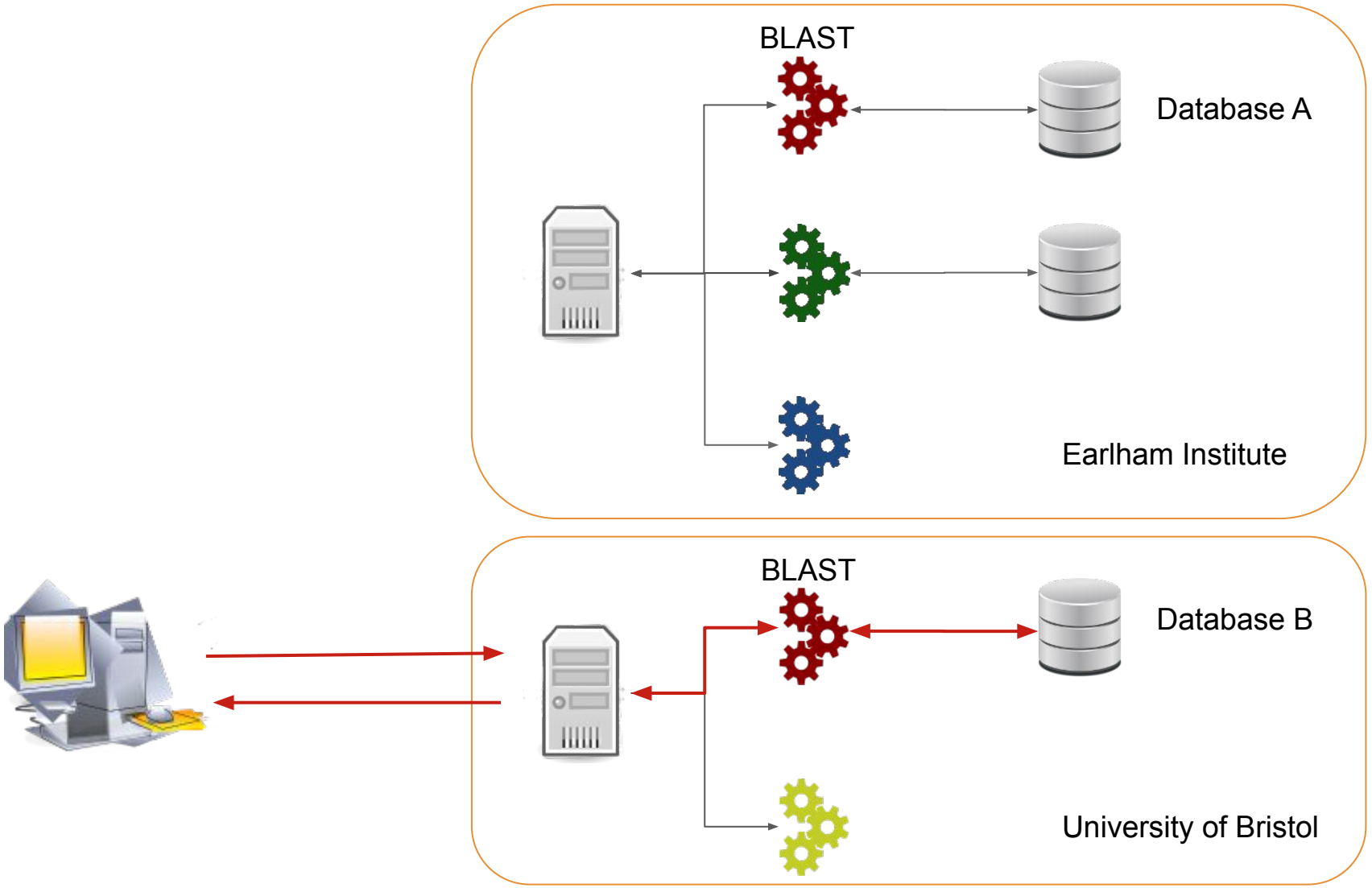
Grassroots Infrastructure - Services

- Components that perform scientific analysis
  - Adapting existing programs
  - Writing our own bespoke tools
- Tools that conform to the Grassroots Services API, which is a well-defined set of standards to access tools and data *e.g.*
  - BLAST
    - - Find areas of similarity between biological sequences
  - Field Trials
  - Unified Search

Standard Web Service Interaction



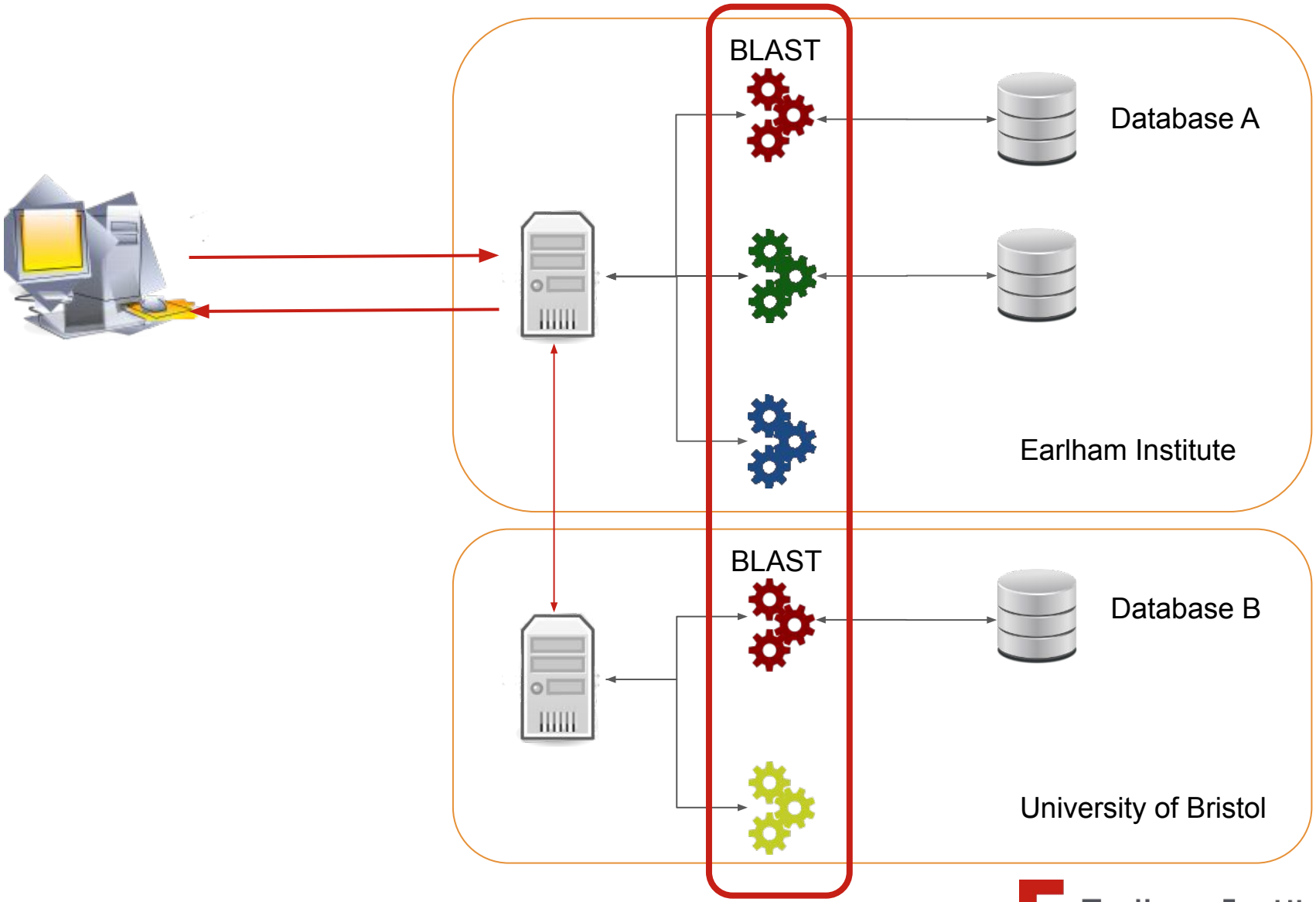
Standard Web Service Interaction



Issues

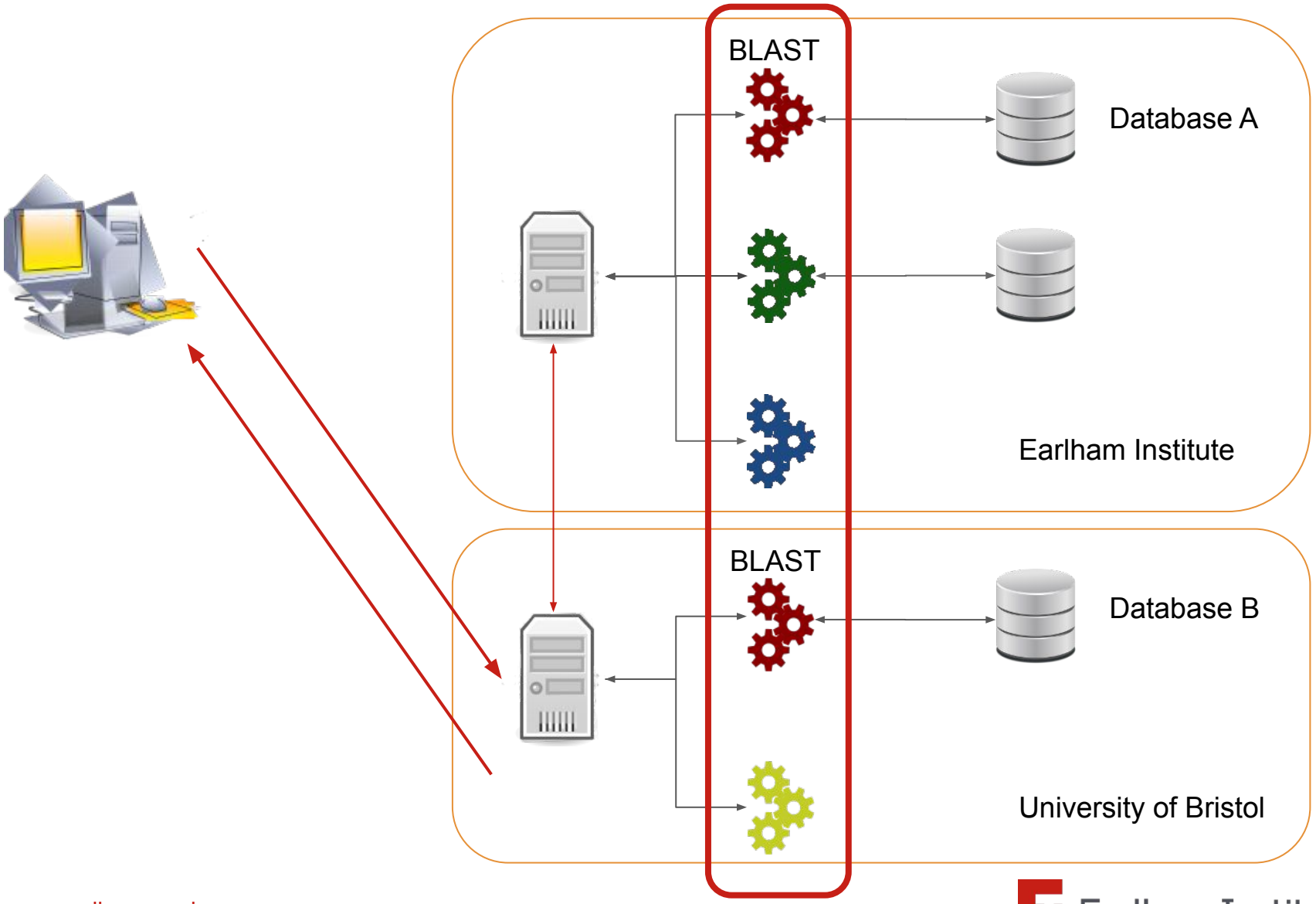
- Manually having to access each Service individually
- Collation of results
- Human error
- Not running each service with the same parameters
- Mistakes when putting the results together
- Time consuming

Federating Services



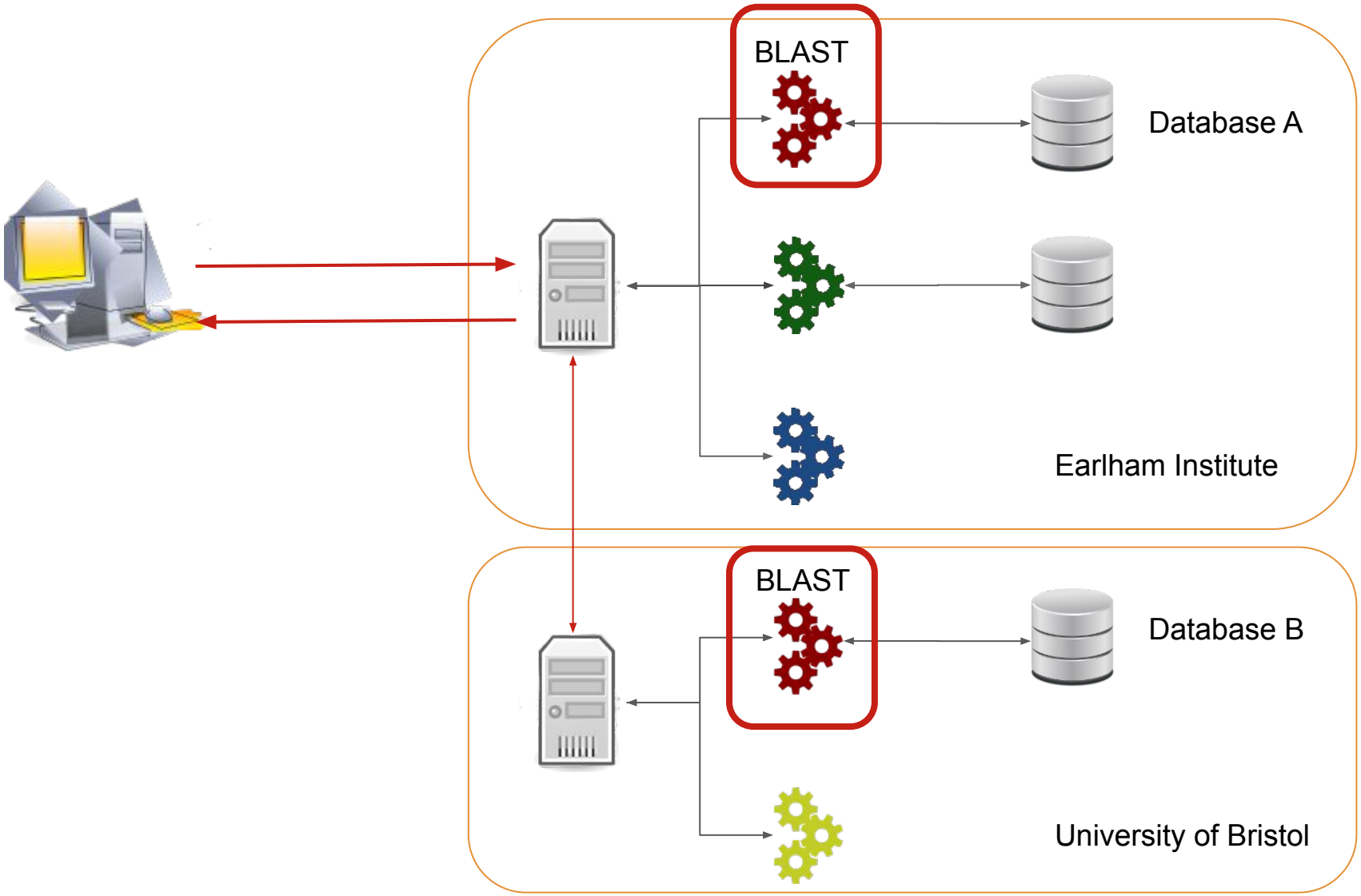
Earlham Institute

Different Server, Same List of Services

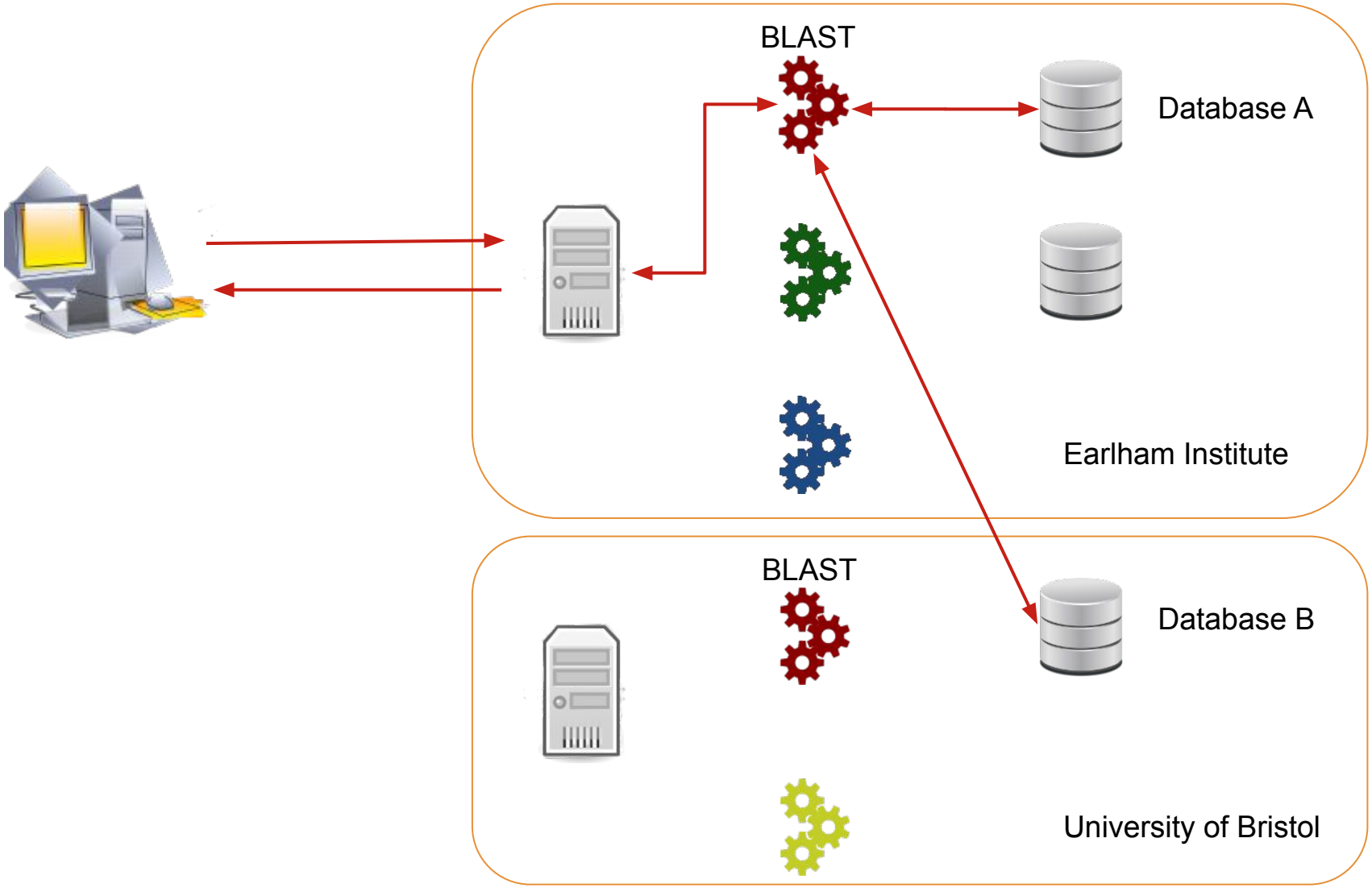


Earlham Institute

Same Services...

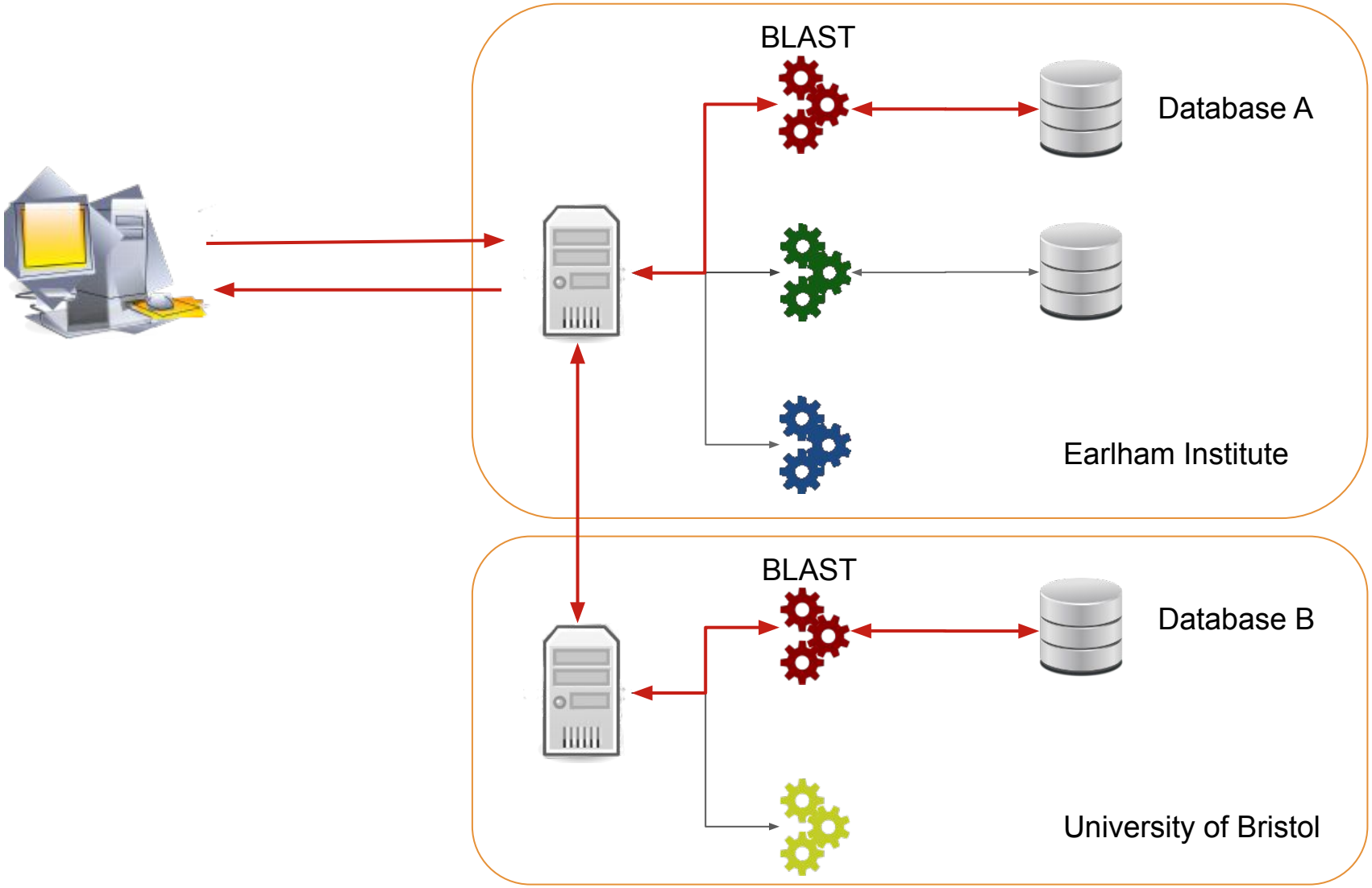


… Get Amalgamated



Earlham Institute

Under the Hood



DFW Data Portal

Available at
https://opendata.earlham.ac.uk/wheat/under_license/toronto/

- A repository for all data generated within DFW
- Based upon the Toronto data agreement
  - Prepublication data sharing
  - This agreement does not expire by time but only upon publication of the first global analysis by the data producers and contributors.
- Hosted on iRODS using mod_eirods_dav

mod_eirods_dav

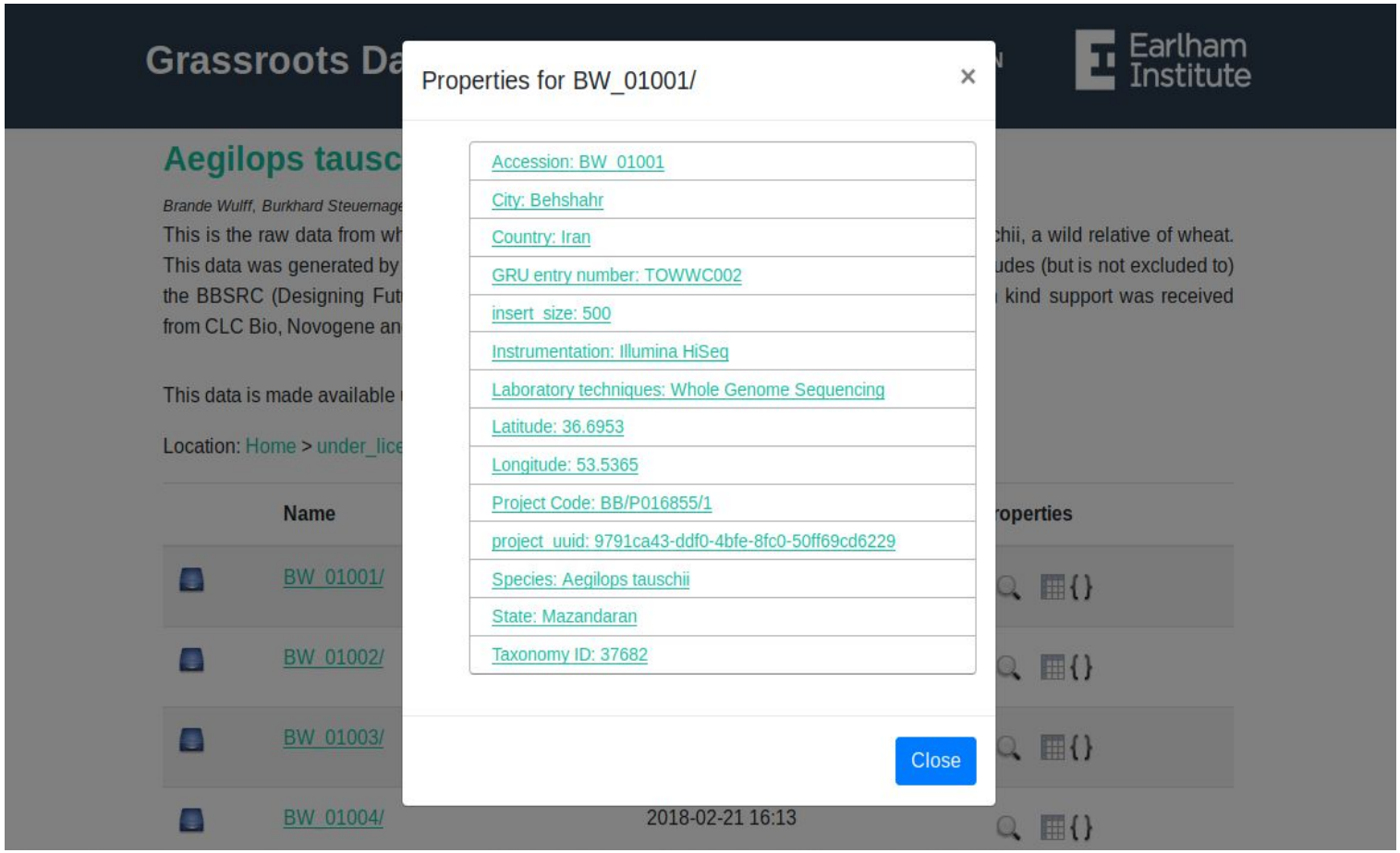
Our open source Apache module, forked from https://github.com/UtrechtUniversity/davrods to access iRODS repositories using standard web technologies.

- Themeable listings similar to mod_autoindex
- Metadata display and editing
- Show public or authenticated user data
- Full REST Web Service API

https://github.com/billyfish/eirods-dav

DFW Data Portal - Metadata

Metadata based upon the [Minimum Information About a Plant Phenotyping Experiment (MIAPPE)](#) standard

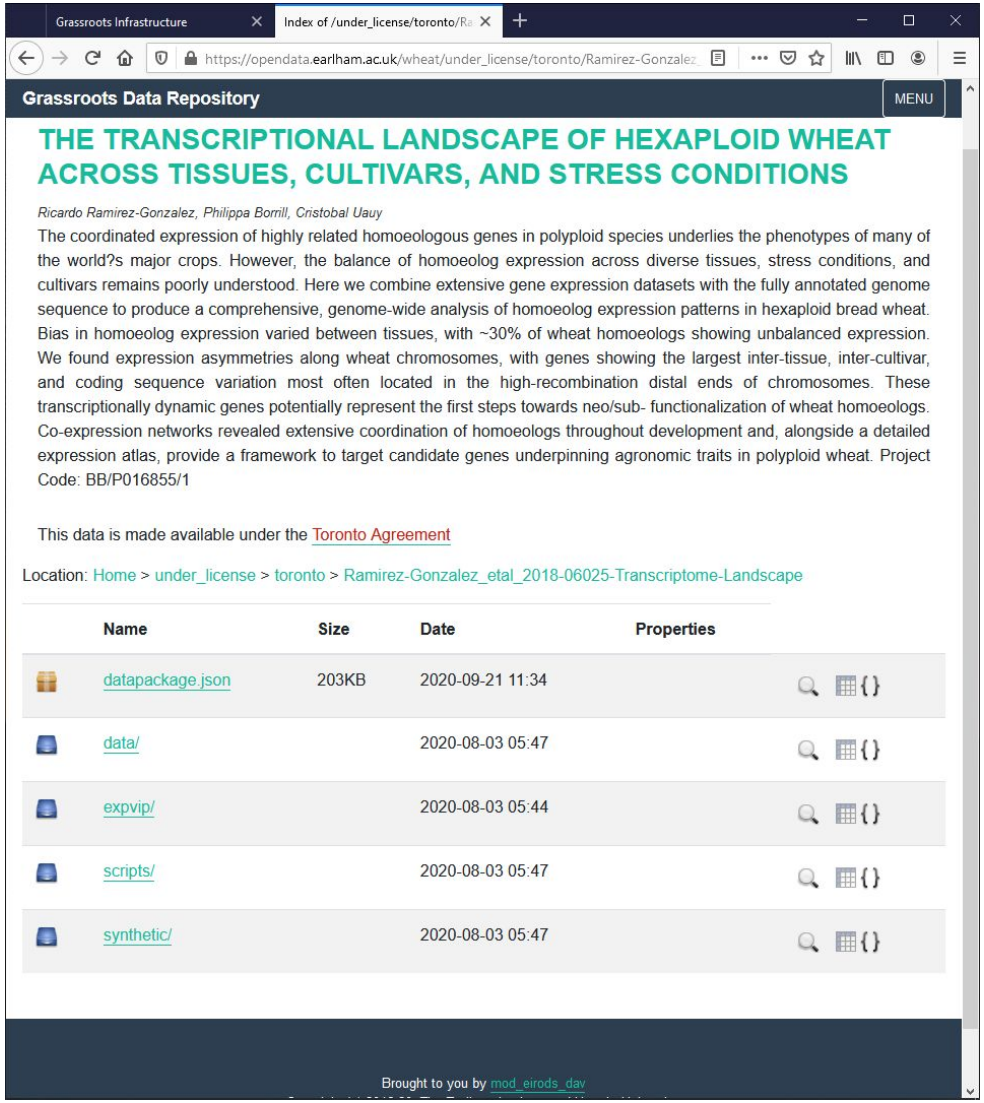


MIAPPE

MIAPPE is a Minimum Information (MI) standard for plant phenotyping.

- A list of attributes that might be necessary to fully describe a phenotyping experiment
- Meaningful data and metadata for the interpretation and potential replication of the research.

DFW Data Portal - Project



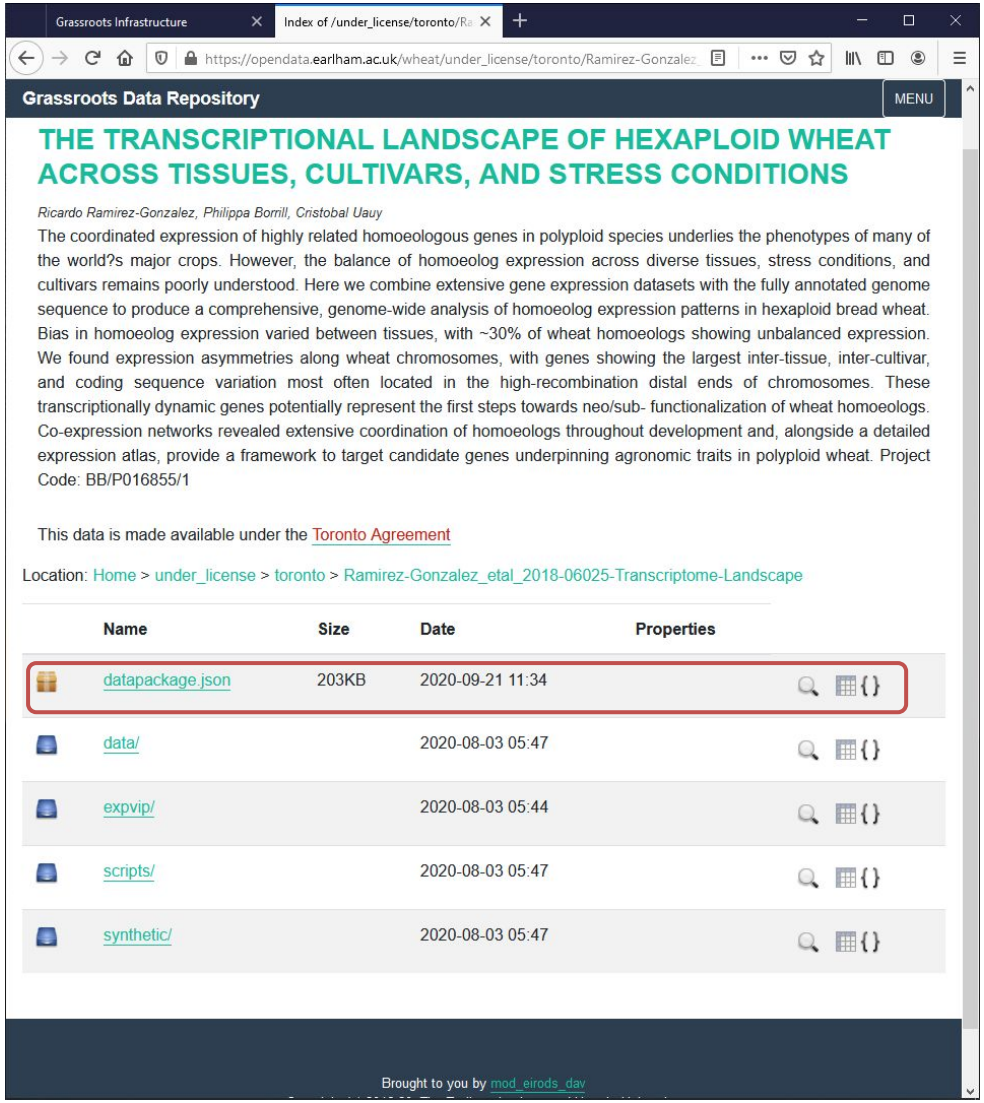
Projects have

- Titles
- Authors
- Descriptions
- License details
- Data

All indexed and searchable using our Lucene-based text search engine

Earlham Institute

DFW Data Portal - Frictionless Data



[Open Knowledge](#) Tool Fund to expose our DFW Data Portal datasets and publications as [Frictionless Data Packages](#)

Earlham Institute

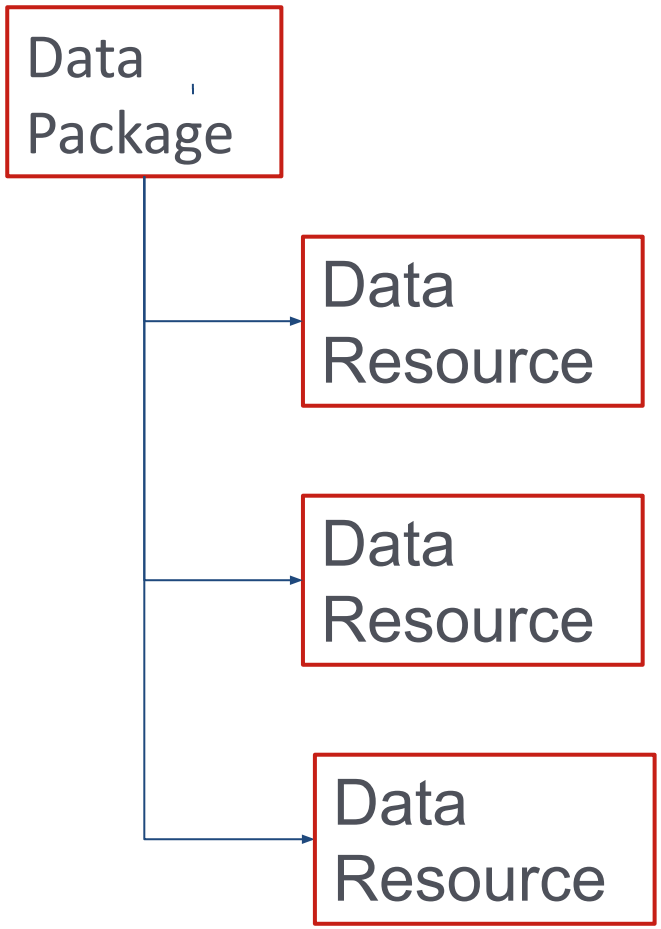
Frictionless Data

A Frictionless Data Package is a simple container format used to describe and package a collection of data.

- Can package any kind of data.
- Simple
- Extensible
- Metadata that is human-editable and machine-usable
- Reuse of existing standard formats for data
- Language, technology and infrastructure agnostic

https://frictionlessdata.io/data-package/

Frictionless Data - Data Package



A Data Package is a container consisting of one or more Data Resources.

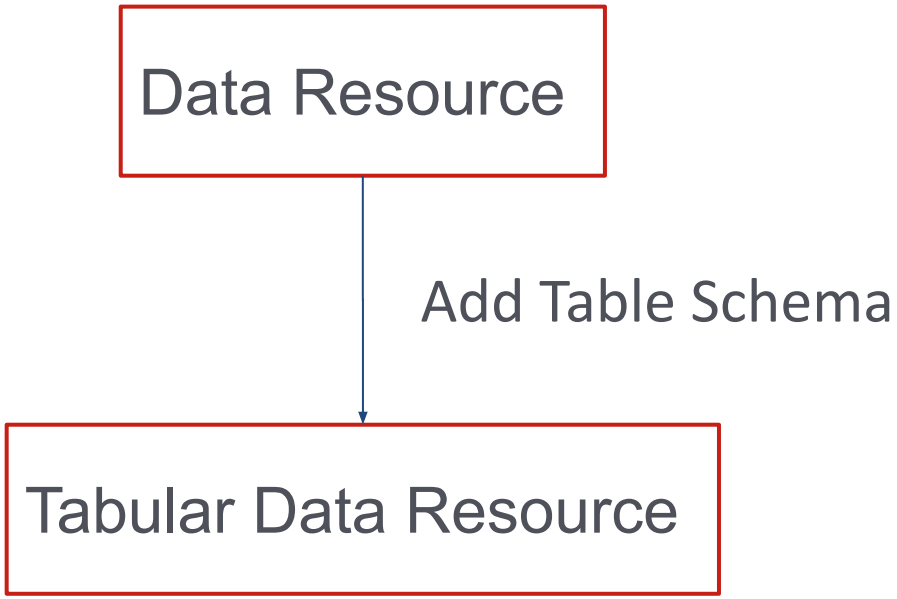
Frictionless Data - Data Resource

Data Resources describe a data resource such as an individual file or table.

- A locator for the data it describes.
  - Path to file
  - Url
- Other properties can be declared to provide a richer set of metadata.

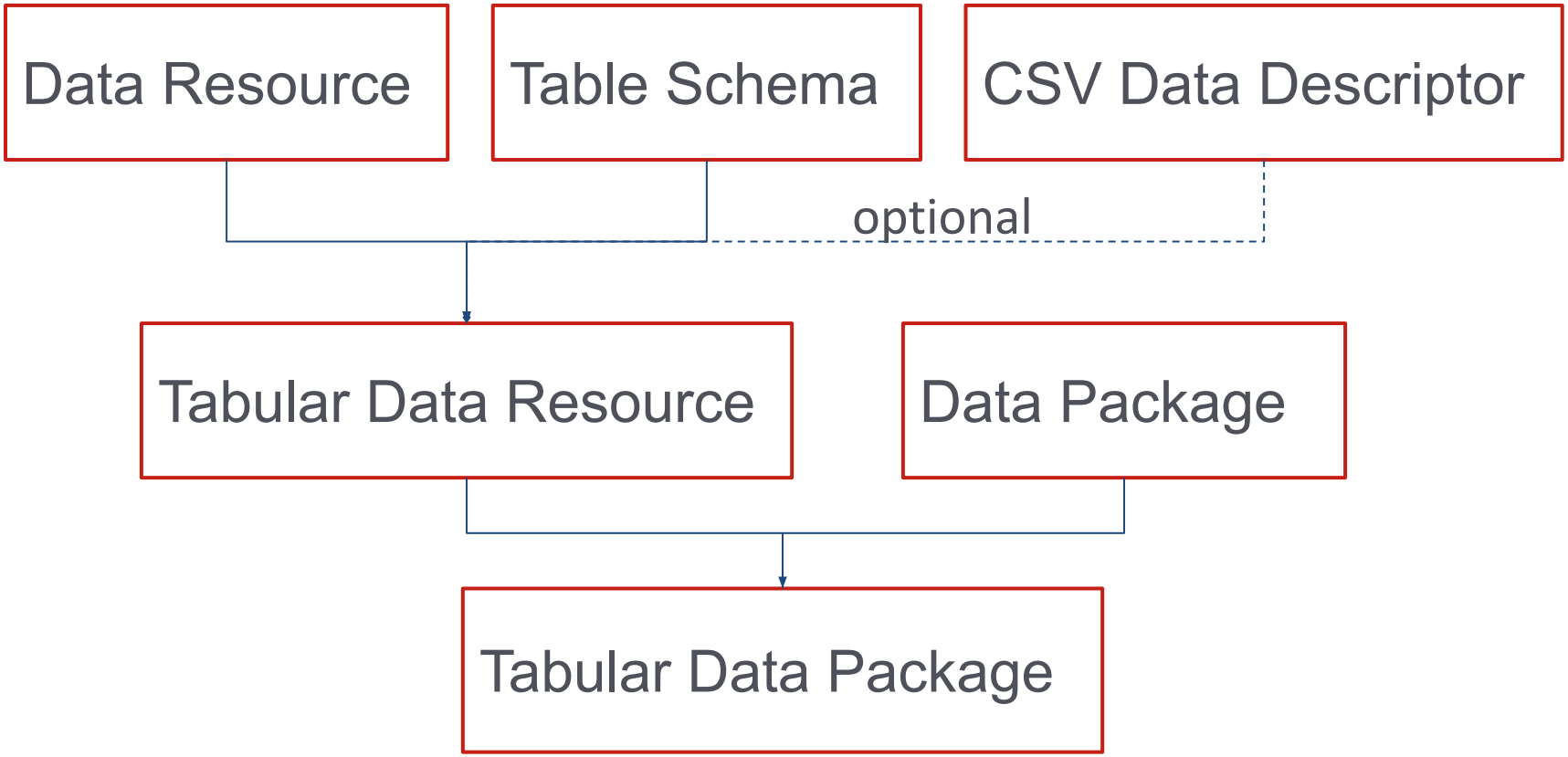
Frictionless Data - Tabular Data Resource

Tabular Data Resources represent data tables such as spreadsheets

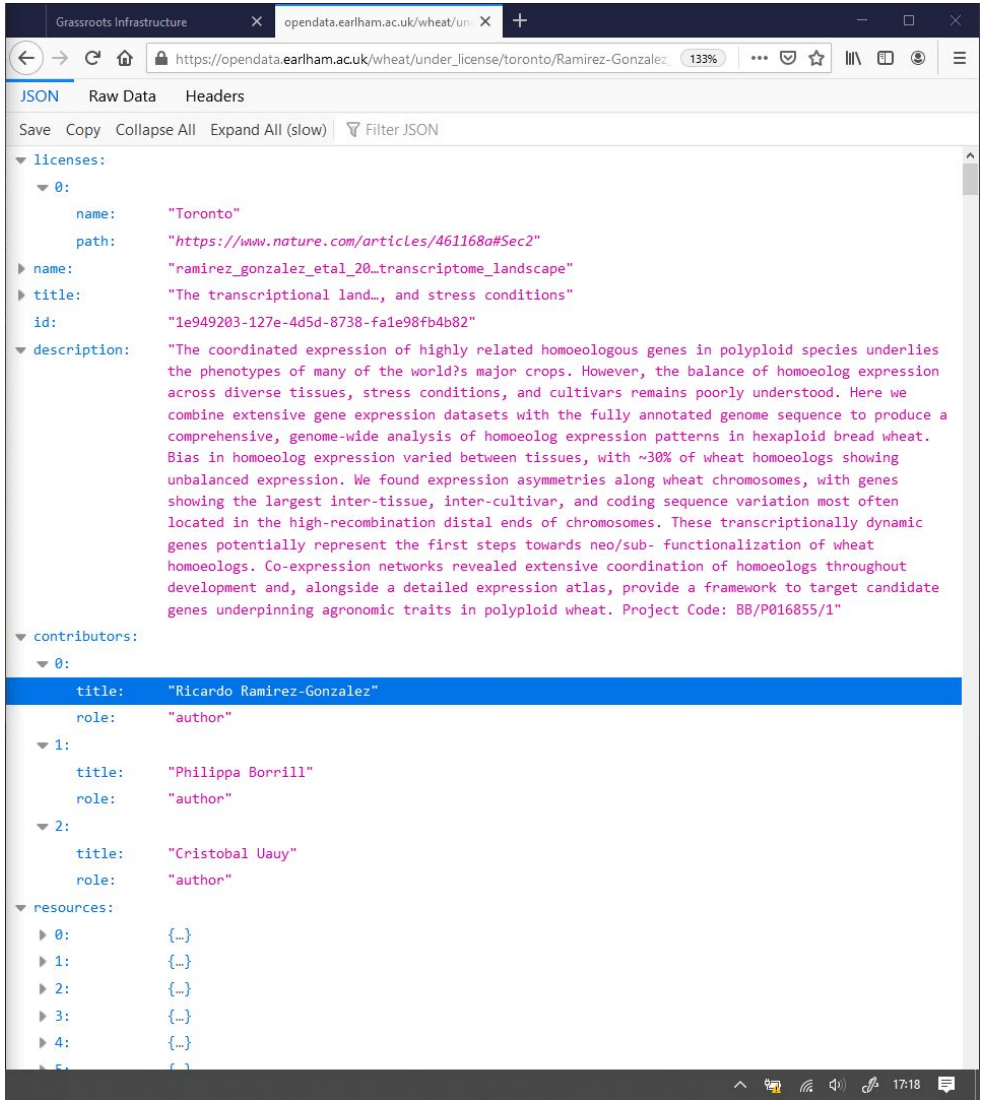


Frictionless Data - Tabular Data Package

Tabular Data Packages contain one or more Tabular Data Resources



DFW Data Portal - Frictionless Data Package



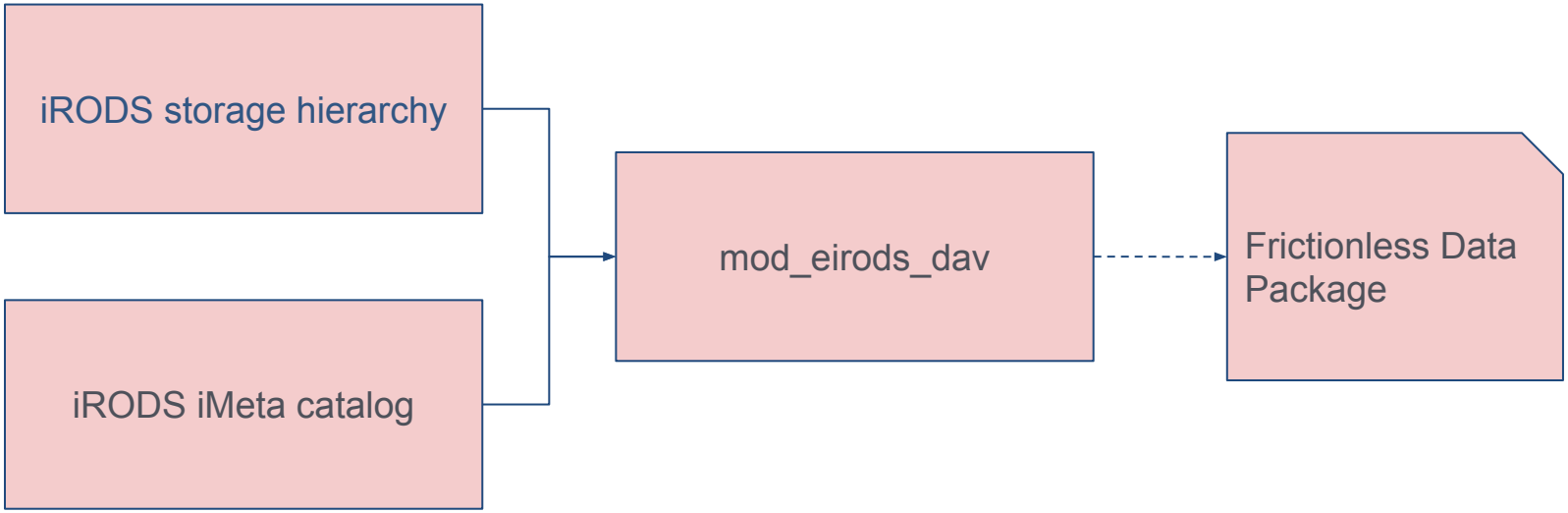
Each Project within the Data Portal has a Frictionless Data Package currently containing

- License
- Name
- Description
- Authors
- Title
- Id

Earlham Institute

mod_eirods_dav - Frictionless Data additions

Dynamic creation of Frictionless Data Packages by generating values from the iRODS metadata values



mod_eirods_dav - Frictionless Data configuration

Enabled using the *DavRodsFrictionlessData* configuration directive

```
# Generate Data Packages for all child directories directly below /data
<LocationMatch "/data/[^\/]+/">
  DavRodsFrictionlessData true
  DavRodsFDDataPackageImage /images/archive
</LocationMatch>

# Exclude all directories further down
<LocationMatch "/data/[^\/]+/[^\/]+/">
   DavRodsFrictionlessData false
</LocationMatch>
```

mod_eirods_dav - Frictionless Data additions

Completely configurable mappings between the iRODS metadata and the Frictionless Data values

Data Package field	**Default iRODS metadata key**
license_name	license
license_url	license_url
description	description
name	name
authors	authors
title	title
id	id

mod_eirods_dav - Frictionless Data additions

- Metadata keys are completely configurable

  E.g. if the value that you wish to use for the description is the *short_info* iMeta key value, then the configuration would be.

```
DavRodsFDResourceDescriptionKey short_info
```

- Values can be combined

  For example, to use the combination of *short_info* and *detailed_info* metadata keys for the description:

```
DavRodsFDResourceDescriptionKey short_info,detailed_info
```

mod_eirods_dav - Frictionless Data additions

- Whitespace, full stops/periods, newlines can also be be used

  For example, if you would like to have the

  - *short_info* metadata value
  - a full stop / period
  - 2 blank lines
  - *detailed_info* metadata value
  - a space
  - *footnote* metadata value

```
DavRodsFDResourceDescriptionKey
short_info,.,\n,\n,detailed_info, ,footnote
```

mod_eirods_dav - Saving generated packages

- By default, the datapackage.json files are virtual and generated on the fly. Although this may be fine for smaller datasets, you may find that the time that is taken to generate these files is too long. So you can configure to store the datapackage.json file within the relevant collection.
    - Equivalent to running the iRODS command iput.

mod_eirods_dav - Tabular Data Resources

Any csv or tsv files in a Frictionless Data package can be configured to display their tabular-specific data fields using the iMeta catalog.

- *column_headings*: A comma-separated list of the column headings for the tabular file

Each of these headings an additional key-value pair specify the type of data in the given column of the file. The keys for these are the column name with a *_type* suffix and the values being ones of the types defined here.

mod_eirods_dav - Tabular Data Resource example

For example, a file *data.csv* which has three columns containing a string, an integer and a floating point number respectively:

var1,var2,var3

A,1,2.1

B,3,4.5

- column_headings: var1,var2,var3
- var1_type: string
- var2_type: integer
- var3_type: number

mod_eirods_dav - Tabular Data Resource example

billy@desktop:~/$ imeta ls -d datasets/tabular/data.csv

attribute: column_headings

value: var1,var2,var3

---

attribute: var1_type

value: string

---

attribute: var2_type

value: integer

---

attribute: var3_type

value: number

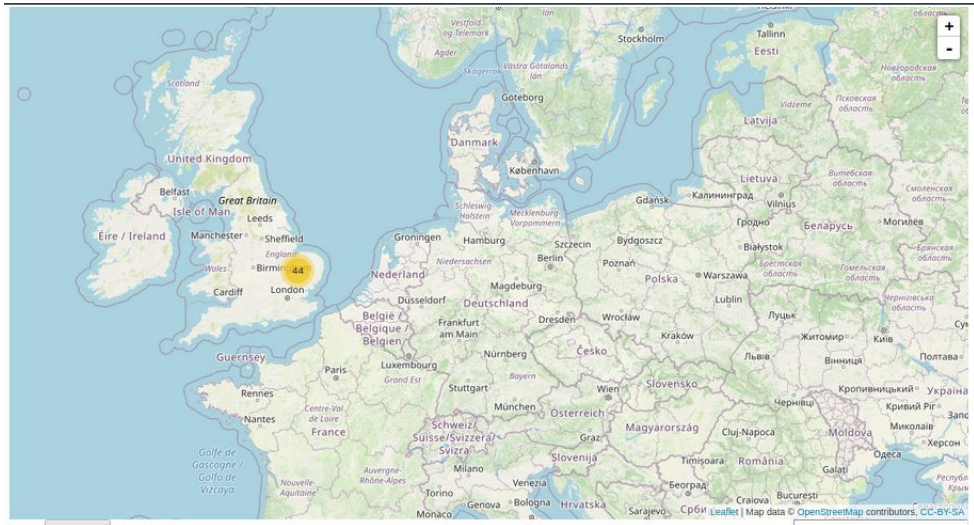
Field Trials

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 Trials - Findable

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



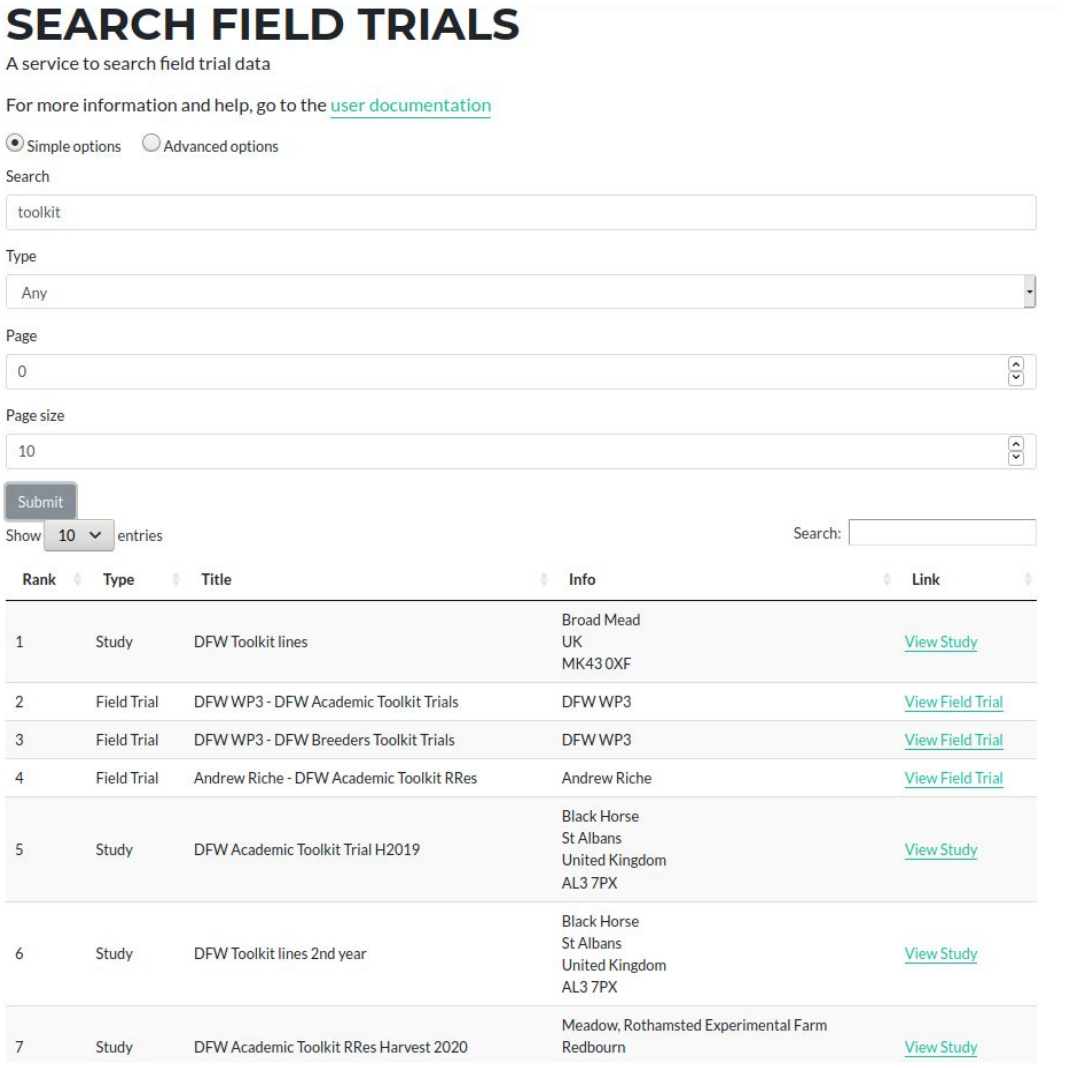
Show 10 entries Search:

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

Earlham Institute

Field Trials - 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:

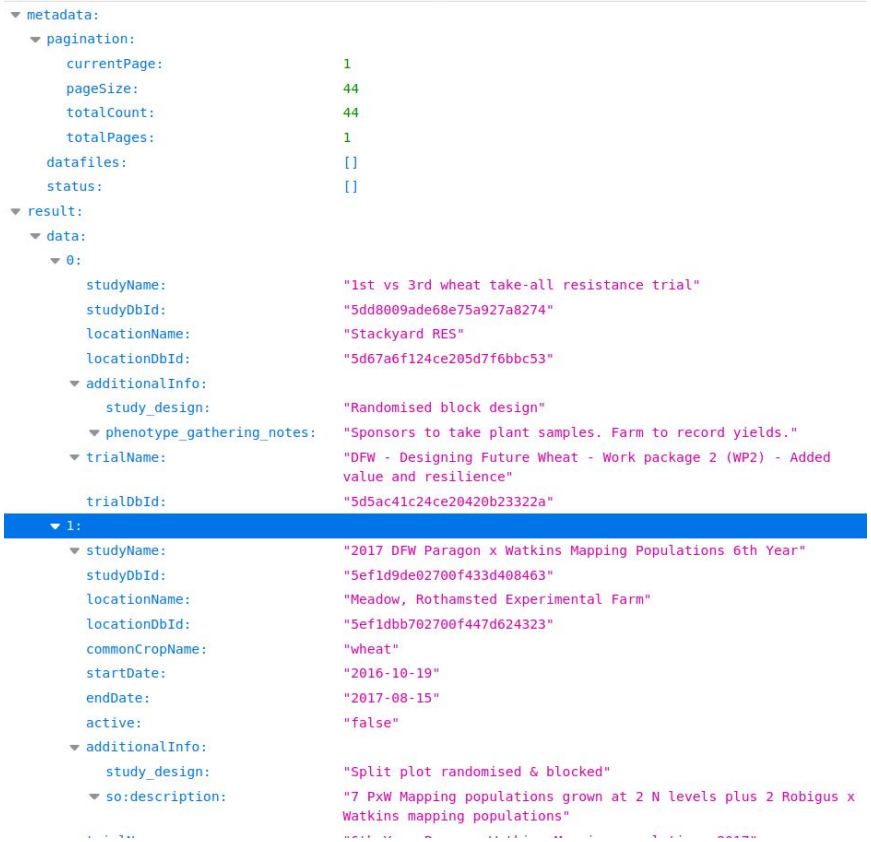
Rank	Type	Title	Info	Link
1	Study	DFW Toolkit lines	Broad Mead UK MK43 0XF	View Study
2	Field Trial	DFW WP3 - DFW Academic Toolkit Trials	DFW WP3	View Field Trial
3	Field Trial	DFW WP3 - DFW Breeders Toolkit Trials	DFW WP3	View Field Trial
4	Field Trial	Andrew Riche - DFW Academic Toolkit RRes	Andrew 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

Field Trials - Accessible

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

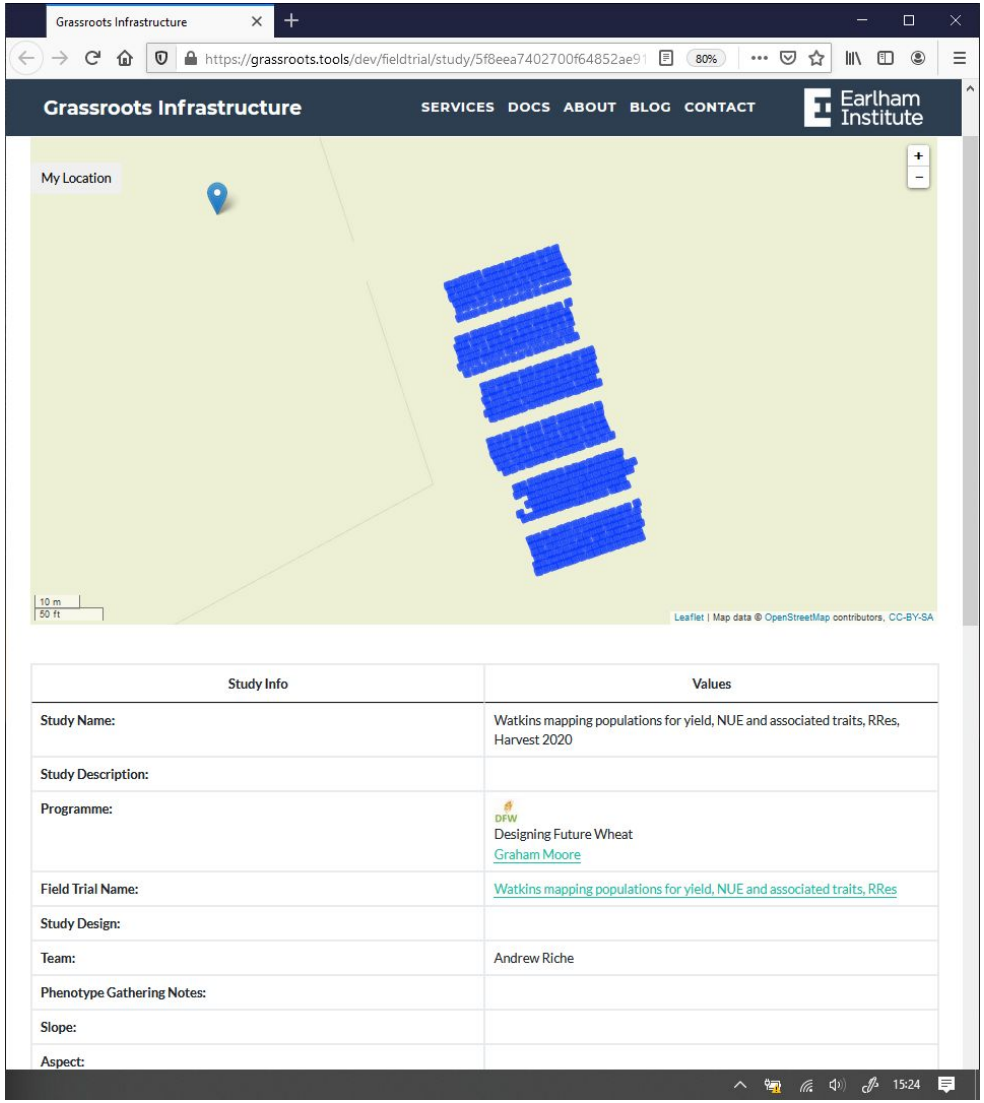
Field Trials - Interoperability

The Field Trials data and metadata is exposed using both BrAPI which is a community-driven standardized RESTful web service API specification to enable interoperability among plant breeding databases.



Earlham Institute

Field Trials - Plots Geolocations



The geolocations of each plot within a study, coupled with automatic location updating, allows the scientists to walk around a study and see which plot they are within in real time.

Earlham Institute

Field Trial – Reusable data

Plot data is standardized using ontological terms for each plot

PLOT DETAILS ×

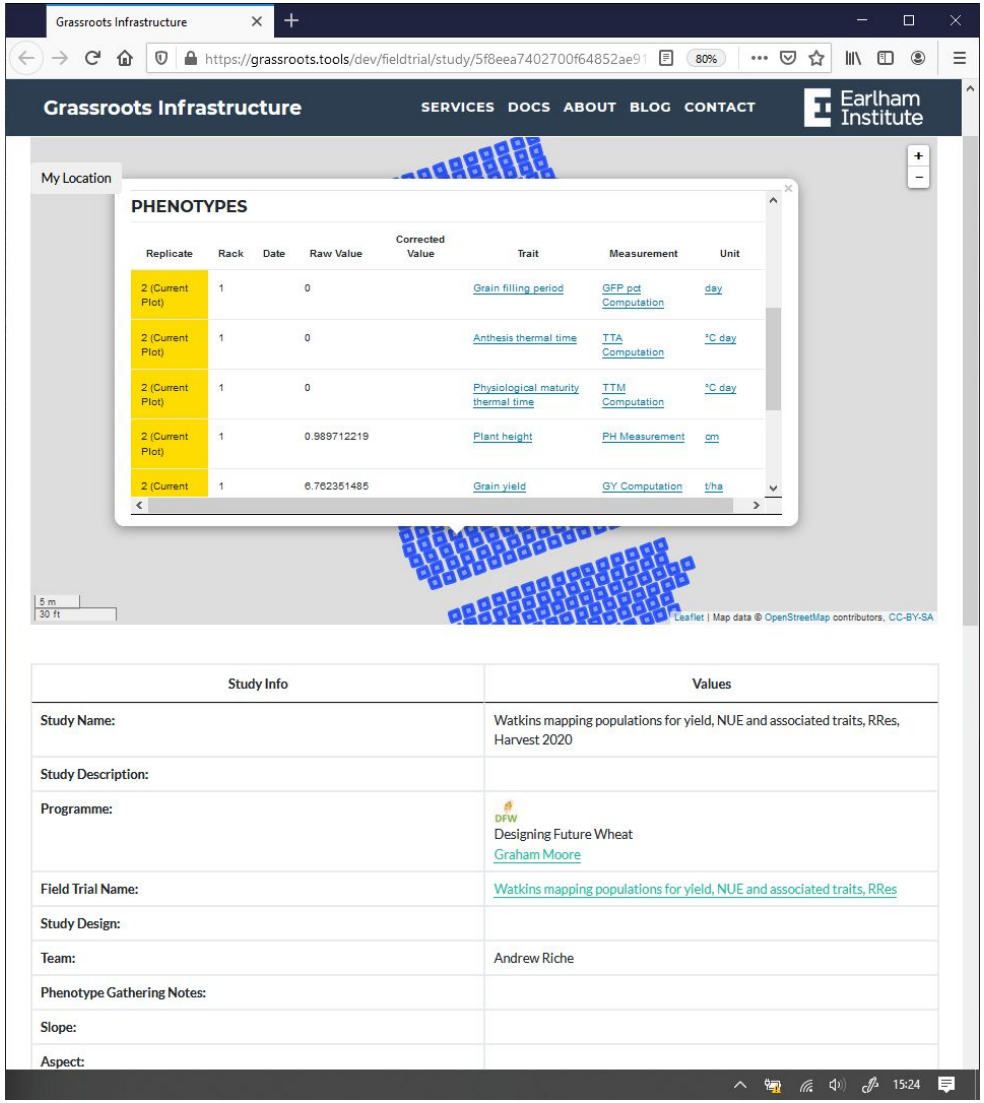
Row: 20
Column: 1
Length: 3.594m
Width: 1.8m
Study Design:
Sowing Date: 2019-10-30
Harvest Date: 2020-08-10
Treatment:
Comment: Slight height segregation

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

Field Trials - Plot Phenotypes



Phenotypes stored as

- Trait
  - What to Measure
- Method
  - How it was measured
- Unit
- Value(s)
- Date

All of these are well-defined terms from the [Crop Ontology]

Earlham Institute

Field Trials - Exposing Data

All of the data and metadata are available via Web Service APIs

- Grassroots
- Partial BrAPI support

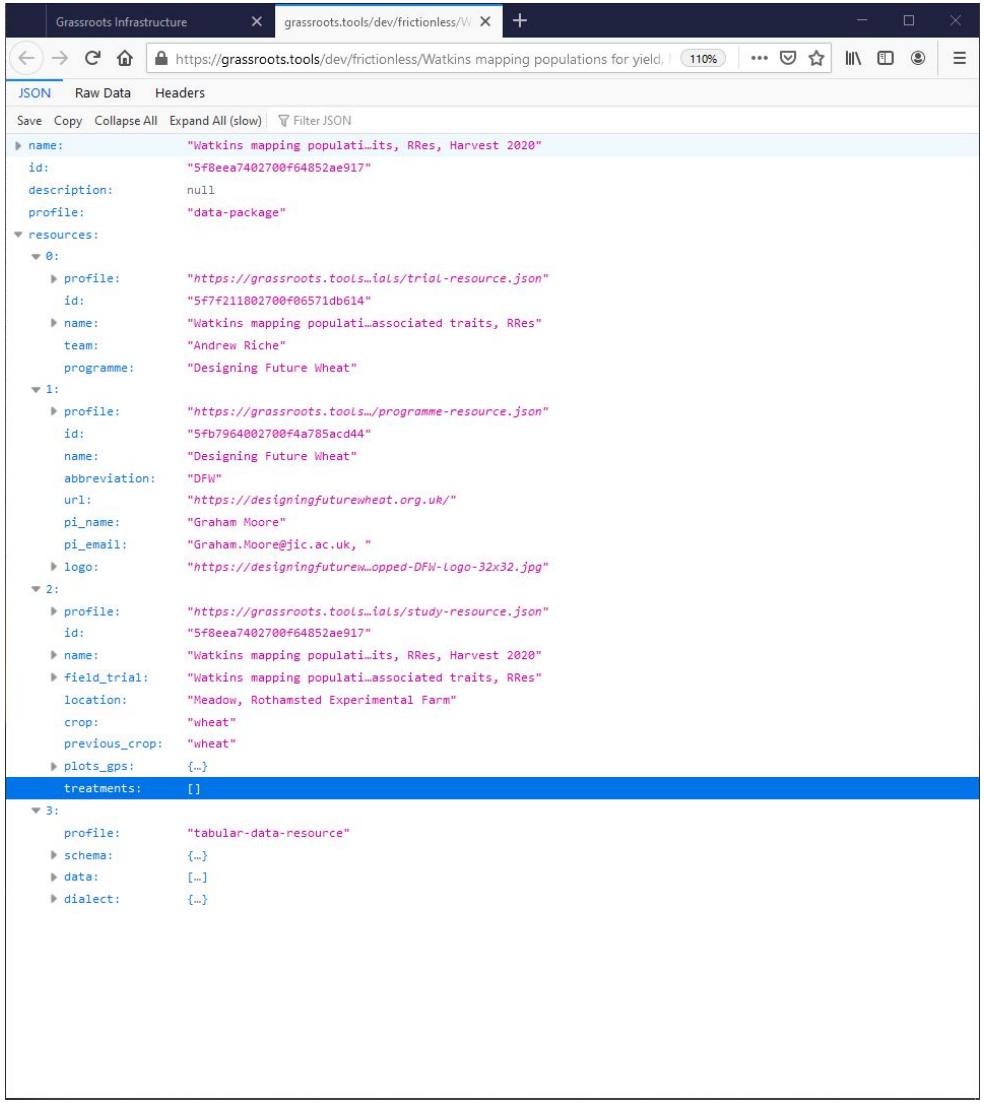
Field Trials - Exposing Data

Q: APIs work for people comfortable scripting and programming, but what about people who just want the basic data...

A: Frictionless Data!

- Grassroots Schemas published at https://grassroots.tools/frictionless-data/
- Other DFW work on Frictionless Data by Richard Ostler at Rothamsted Research

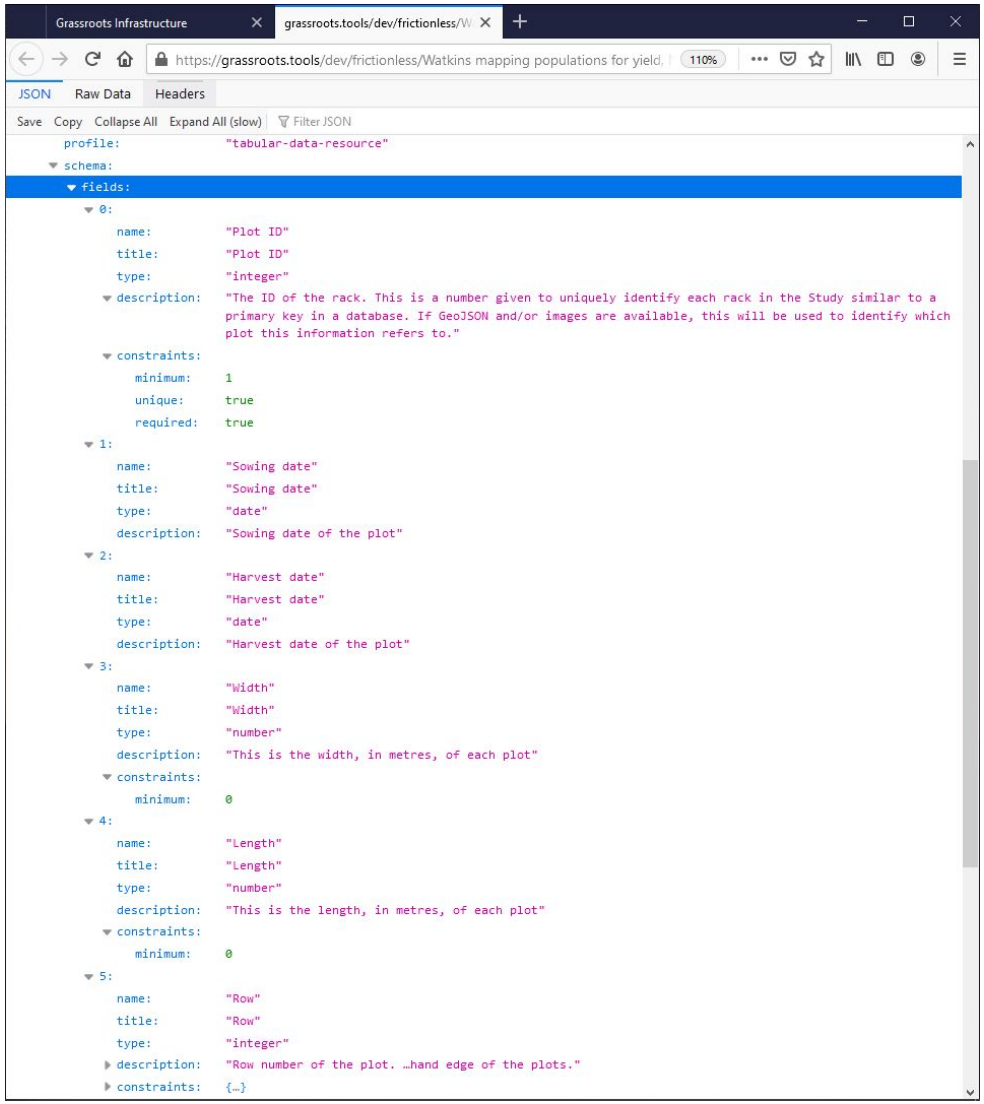
Field Trials - Study Frictionless Data Package



All of the details of each individual study are stored in a single Frictionless Data Package

Earlham Institute

Field Trials - Plots Schema Fields



The data for the study's plots is tabular with dynamically generated schemas

Standard attributes

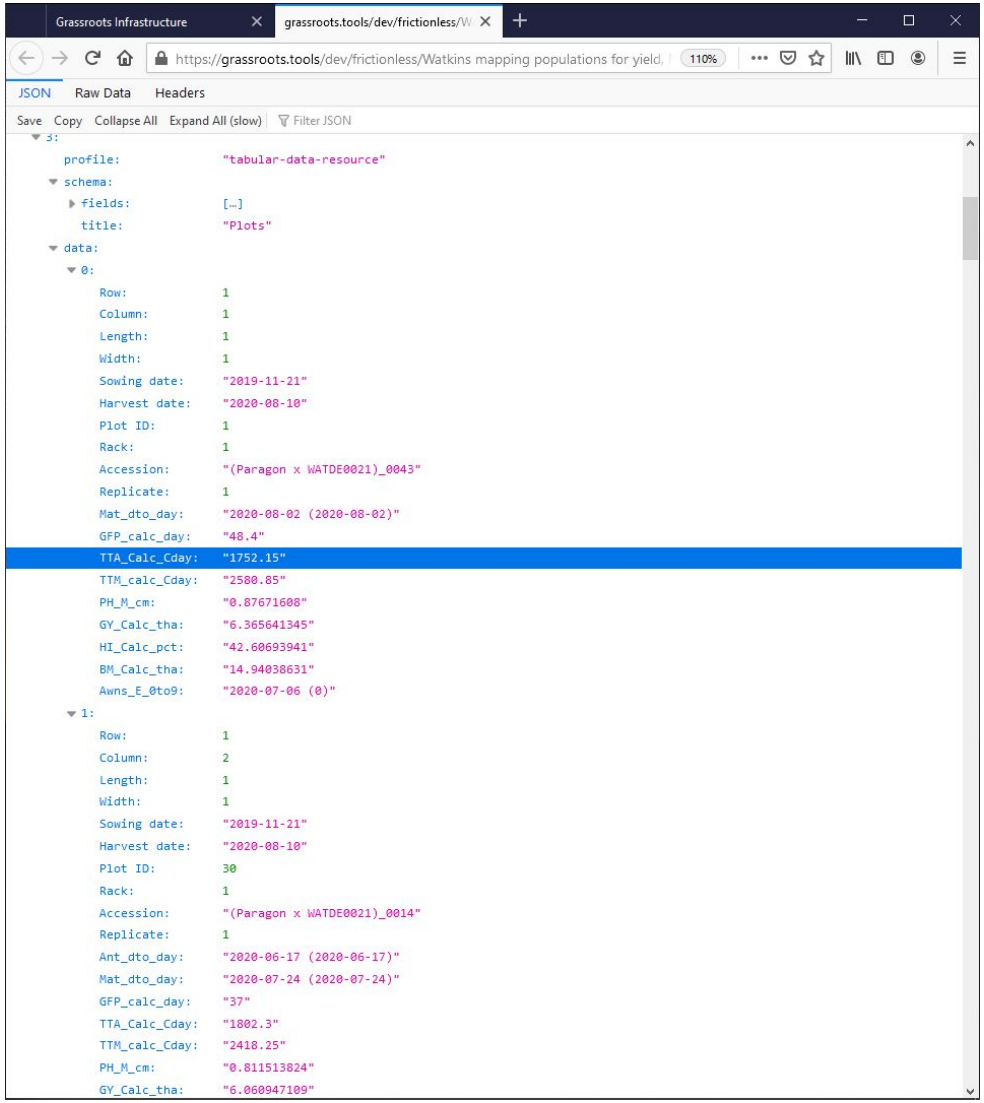
- Length
- Width
- Position
- *etc.*

Custom attributes

- Treatments *e.g.* fertilizers
- Phenotypes

Earlham Institute

Field Trials - Plots Tabular Data



Earlham Institute

Field Trials - Exposing Data

So people can download a field trial as a Frictionless Data Package and Frictionless comes with a great API to extract the data

Can we make it simple for them to unpack the data?

Field Trials - Exposing Data

A tool to extract the resources within a Frictionless Data Package

- Downloads and uses the schemas specified by the profile value of each Data Resource
- Converts Data Resources
  - Markdown
  - HTML
- Converts Tabular Data Resources
  - CSV
- Cross Platform and Open Source
- Not wheat-specific, works with any Data Package

More information at

https://grassroots.tools/frictionless-data/grassroots-fd-client.md

Frictionless Data


```
{
  "name": "WGIN Diversity Rothamsted Harvest 2021",
  "id": "603e290902700f3758557214",
  "description": "Wheat variety, N, pesticide interaction trial",
  "profile": "data-package",
  "resources": [{
      "profile":
"https://grassroots.tools/frictionless-data/schemas/field-trials/trial-resource.json",
      "id": "603e12a602700f380d2e01d5",
      "name": "WGIN Diversity",
      "team": "Andrew Riche",
      "programme": "Wheat Genetic Improvement Network"
  },
  …
}
```

Field Trials - Download profile


```
{
  "name": "WGIN Diversity Rothamsted Harvest 2021",
  "id": "603e290902700f3758557214",
  "description": "Wheat variety, N, pesticide interaction trial",
  "profile": "data-package",
  "resources": [{
      "profile":
"https://grassroots.tools/frictionless-data/schemas/field-trials/trial-resource.json",
      "id": "603e12a602700f380d2e01d5",
      "name": "WGIN Diversity",
      "team": "Andrew Riche",
      "programme": "Wheat Genetic Improvement Network"
  },
  …
}
```

Field Trials - Frictionless Data

https://grassroots.tools/frictionless-data/schemas/field-trials/trial-resource.json

```
{
    "$schema": "http://json-schema.org/draft-04/schema#",
    "title": "Field Trial Data Resource",
    "description": "Field Trial Data Resource is a specification for detailing a field trial containing one or more studies.",
    "type": "object",
    "required": [
        "profile",
        "name"
    ],
    "properties": {
        "profile": {
    …
}
```

Field Trials - Frictionless Data

An output file is generated for each data resource

WGIN Diversity

- **profile ***: https://grassroots.tools/frictionless-data/schemas/field-trials/trial-resource.json
- **id**: 603e12a602700f380d2e01d5
- **name ***: WGIN Diversity
- **programme**: Wheat Genetic Improvement Network
- **team**: Andrew Riche

Parsed ~/Downloads/WGIN Diversity Rothamsted Harvest 2021.json using profile https://grassroots.tools/frictionless-data/schemas/field-trials/trial-resource.json

Further Work

- Add more information to the generated Frictionless Data Packages
- Add Machine Learning to detect phenotypic values from media such as photos taken by drones
- Further collaboration with Richard Ostler refining a common Frictionless Data standard for both single-year and long-term field trial wheat studies

Acknowledgements

- Earlham Institute
  - Alice Minotto
  - Felix Shaw
  - Anthony Etuk
  - Nicola Soranzo
  - Martin Ayling
  - Catherine Hunter
  - Anil Thanki
  - Jon Wright
  - Luis Yanes
  - Anthony Hall
  - Laura Jayne Gardiner
  - Ben White
  - Joshua Ball
- John Innes Centre
  - Luzie Wingen
  - Simon Griffiths
- Rothamsted Research
  - Andrew Riche
  - Chris Rawlings
  - Richard Ostler
- University of Bristol
  - Paul Wilkinson
  - Mark Winfield
  - Gary Barker

And a BIG thank you to the iRODS UGM 2021 organizers!

SIMON TYRRELL

Research Software Engineer

simon.tyrrell@earlham.ac.uk

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