Frictionless Data for iRODS

SIMON TYRRELL RESEARCH SOFTWARE ENGINEER

XINGDONG BIAN RESEARCH SOFTWARE ENGINEER

ROB DAVEY GROUP LEADER

f 🎔 🞯 in 🛗 🔊 🐱



Decoding Living Systems

Designing Future Wheat (DFW)

The <u>BBSRC</u> 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 <u>https://www.go-fair.org/fair-principles/</u>



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



Different Server, Same List of Services



Same Services...



... Get Amalgamated



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 <u>Toronto data agreement</u>
 - 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 <u>https://github.com/UtrechtUniversity/davrods</u> 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 <u>Minimum Information About a Plant</u> <u>Phenotyping Experiment (MIAPPE)</u> standard

Grassroots Da	Properties for BW_01001/ ×	Earlham Institute
Aegilops tausc Brande Wulff, Burkhard Steuemage This is the raw data from wh This data was generated by the BBSRC (Designing Fut from CLC Bio, Novogene an	Accession: BW_01001 City: Behshahr Country: Iran GRU entry number: TOWWC002 insert_size: 500 Instrumentation: Illumina HiSeg	chii, a wild relative of wheat. udes (but is not excluded to) kind support was received
This data is made available Location: Home > under_lice Name	Laboratory techniques: Whole Genome Sequencing Latitude: 36.6953 Longitude: 53.5365 Project Code: BB/P016855/1 project, unid: 9791ce43-ddfp-4bfe-9fc0-50ff69cd6229	operties
BW 01001/	Species: Aegilops tauschii State: Mazandaran	् ा {}
BW 01002/	Taxonomy ID: 37682	् ा{}
BW 01003/	Close 2018-02-21 16:13	् ≣{} ् ≣{}



MIAPPE

MIAPPE is a <u>Minimum Information (MI)</u> 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

The coordinated expression of highly related homoeologous genes in polypoid species underlies the phenotypes of many of the world?s major crops. However, the balance of homoeolog expression across diverse tissues, stress conditions, and cultivars remains poorly understood. Here we combine extensive gene expression datasets with the fully annotated genome sequence to produce a comprehensive, genome-wide analysis of homoeolog expression patterns in hexaploid bread wheat. Bias in homoeolog expression varied between tissues, with ~30% of wheat homoeologs showing unbalanced expression. We found expression asymmetries along wheat chromosomes, with genes showing the largest inter-tissue, inter-cultivar, and coding sequence variation most often located in the high-recombination distal ends of chromosomes. These transcriptionally dynamic genes potentially represent the first steps towards neo/sub- functionalization of wheat homoeologs. Co-expression networks revealed extensive coordination of homoeologs throughout development and, alongside a detailed expression atlas, provide a framework to target candidate genes underpinning agronomic traits in polyploid wheat. Project Code: BB/P016855/1

This data is made available under the Toronto Agreement

Location: Home > under_license > toronto > Ramirez-Gonzalez_etal_2018-06025-Transcriptome-Landscape

Name	Size	Date	Properties	
datapackage.json	203KB	2020-09-21 11:34		ୣ ∎{}
data/		2020-08-03 05:47		■{}
expvip/		2020-08-03 05:44		् ■{}
scripts/		2020-08-03 05:47		■{}
synthetic/		2020-08-03 05:47		ୣ ∎{}
	Bro	ught to you by <u>mod_eirods_dav</u>		

Projects have

- Titles
- Authors
- Descriptions
- License details

Data

All indexed and searchable using our <u>Lucene-based</u> text search engine



DFW Data Portal - Frictionless Data

Grassroots Infrastructure

Index of /under_license/toronto/Ra 🗙 🕂

- 0

MENU

🔶) 🔶 😋 🕼 🔽 🔒 https://opendata.earlham.ac.uk/wheat/under_license/toronto/Ramirez-Gonzalez 🗉 🚥 😎 🏠 航 🗉 🕲

Grassroots Data Repository

THE TRANSCRIPTIONAL LANDSCAPE OF HEXAPLOID WHEAT ACROSS TISSUES, CULTIVARS, AND STRESS CONDITIONS

Ricardo Ramirez-Gonzalez, Philippa Borrill, Cristobal Uauy

The coordinated expression of highly related homoeologous genes in polyploid species underlies the phenotypes of many of the world?s major crops. However, the balance of homoeolog expression across diverse tissues, stress conditions, and cultivars remains poorly understood. Here we combine extensive gene expression datasets with the fully annotated genome sequence to produce a comprehensive, genome-wide analysis of homoeolog expression patterns in hexaploid bread wheat. Bias in homoeolog expression varied between tissues, with ~30% of wheat homoeologs showing unbalanced expression. We found expression asymmetries along wheat chromosomes, with genes showing the largest inter-tissue, inter-cultivar, and coding sequence variation most often located in the high-recombination distal ends of chromosomes. These transcriptionally dynamic genes potentially represent the first steps towards neo/sub- functionalization of wheat homoeologs. Co-expression atlas, provide a framework to target candidate genes underpinning agronomic traits in polyploid wheat. Project Code: BB/P016855/1

This data is made available under the Toronto Agreement

Location: Home > under_license > toronto > Ramirez-Gonzalez_etal_2018-06025-Transcriptome-Landscape

Name	Size	Date	Properties		
datapackage.json	203KB	2020-09-21 11:34		Q	₩{}
data/		2020-08-03 05:47		Q,	₩{}
expvip/		2020-08-03 05:44		Q	■{}
scripts/		2020-08-03 05:47		Q	
synthetic/		2020-08-03 05:47		Q	₩{}
	Brou	ight to you by <u>mod_eirods_dav</u>			

Open Knowledge Tool Fund to expose our DFW Data Portal datasets and publications as <u>Frictionless</u> <u>Data Packages</u>



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
 - o 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



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>



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



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



- 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...





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				
Туре				
Any				
Page				
0				•
Page size				
10				× •
Submit	1			
Show 10) 🗸 entries		Search:	
Rank	Туре	Title	Info	🕴 Link 🛛 🕚
			Broad Mead	
1	Study	DFW Toolkit lines	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
			Black Horse	
5	Study	DFW Academic Toolkit Trial H2019	St Albans	View Study
			AL3 7PX	
			Black Horse	
6	Study	DFW Toolkit lines 2nd year	St Albans	View Study
			AL3 7PX	
-	C 1		Meadow, Rothamsted Experimental Farm	
/	Study	DEW Academic Toolkit RRes Harvest 2020	Kedbourn	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 <u>BrAPI</u> which is a community-driven standardized RESTful web service API specification to enable interoperability among plant breeding databases.

▼ metadata:	
<pre>> pagination:</pre>	
currentPage:	1
pageSize:	44
totalCount:	44
totalPages:	1
datafiles:	П
status:	П
<pre>v result:</pre>	
▼ data:	
▼ 0:	
studyName:	"1st vs 3rd wheat take-all resistance trial"
studyDbId:	"5dd8009ade68e75a927a8274"
locationName:	"Stackyard RES"
locationDbId:	"5d67a6f124ce205d7f6bbc53"
<pre> additionalInfo:</pre>	
study_design:	"Randomised block design"
<pre>w phenotype_gathering_notes:</pre>	"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"
in the second	



Field Trials - Plots Geolocations

Grassroots Infrastructure × +							_		
→ C ^I (a) (b) (b) https://grassroots.tools/dev/f	fieldtrial/study/5f8	eea740270	0f64852ae	e91 🗉	80%	⊘	☆	lii\ C	
Grassroots Infrastructure	SERVICES	DOCS	ABOUT	BLOG	CONTAC	т	Ι	Earlf Insti	nam tute
My Location			3						+
10 m 50 ft					eaflet Map dat	ta 🖲 OpenStre	eetMap o	ontributors	. CC-BY-S/
10 m 50 ft Study Info				,	eaflet Map dat Values	ta @ OpenStre	eetMap o	ontributors	, CC-BY-S/
10 m 50 ft Study Info Study Name:		Natkins map Harvest 202	oping popula 0	ations for	eaflet Map dat Values • yield, NUE	ta OpenStre	eetMap o	ontributors raits, RR	. CC-8Y-8/
10 m 50 ft Study Info Study Name: Study Description:		Atkins map Harvest 2020	oping popula 0	ations for	eaflet Map dat Values	ta @ OpenStre	eetMap o	ontributors raits, RR	, CC-8Y-8/
10 m 50 ft Study Info Study Name: Study Description: Programme:		Watkins map Harvest 2020	pping popula 0 ture Wheat	ations for	vafiet Map dat	ta © OpenStre	eetMap o	ontributors raits, RR	. CC-BY-SA
10 m 50 ft Study Info Study Name: Study Description: Programme: Field Trial Name:		Natkins map Harvest 202 Designing Fu Graham Moo Watkins map	oping popula 0 ture Wheat	ations for t	vield, NUE	and associ	iated to	ontributors raits, RR	. CC-BY-S4 es,
10 m 50 ft Study Info Study Name: Study Description: Programme: Field Trial Name: Study Design:		Vatkins map Harvest 2021	oping popula 0 ture Wheat	ations for t	values Values vjeld, NUE	and associ	iated tr	ontributors raits, RR	. CC-8Y-S4 les,
10 m 1 50 ft Study Info Study Name: Study Description: Programme: Field Trial Name: Study Design: Team:		Watkins map Harvest 2021	oping popula 0 ture Wheat ore e	ations for t	vield, NUE	and associ	iated ti	ontributors raits, RR	es,
10 m 50 ft Study Info Study Name: Study Description: Programme: Field Trial Name: Study Design: Team: Phenotype Gathering Notes:		Watkins map Harvest 2021 Designing Fu Graham Moo Natkins map	oping popula 0 ture Wheat oping popula	t t	vield, NUE	and associ	iated tr	ontributors raits, RR	. cc-87-5/
10 m 50 ft Study Info Study Name: Study Description: Programme: Field Trial Name: Study Design: Team: Phenotype Gathering Notes: Slope:		Andrew Rich	oping popula 0 ture Wheat ping popula e	ations for t	vield, NUE	and associ	iated tr	ontributors raits, RR	. CC-BY-S4

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.



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 <u>(Plot Row:3 - Col:23)</u>	1	DFW SEL 0208		Germplasm Resources Unit	
2 <u>(Plot Row:14 - Col:15)</u>	1	DFW SEL 0208		Germplasm Resources Unit	

PHENOTYPES

Close

×



Field Trials - Plot Phenotypes

Grassroots	Infrastructure		× +	-							- 0	×
 	@ ₪ ₽	https://	grassro	ots.tools/dev/	fieldtrial/study	//5f8eea74027001	f64852ae91 🗉 🤇	80%	• 🛛	☆ II	\ 🗉 🔹	Ξ
Grassro	oots Infra	stru	cture	•	SERVI	CES DOCS A	BOUT BLOG C	ONTACT		I E	arlham Istitute	^
My Location	_				ADDA						+	
	PHENOT	YPES							^			
	Replicate	Rack	Date	Raw Value	Corrected Value	Trait	Measurement	Unit				
	2 (Current Plot)	1		0		Grain filling period	GFP pct Computation	<u>day</u>	Ы			
	2 (Current Plot)	1		0		Anthesis thermal time	TTA Computation	°C day				
	2 (Current Plot)	1		0		Physiological maturity thermal time	<u>TTM</u> Computation	°C day				
	2 (Current Plot)	1		0.989712219		Plant height	PH Measurement	<u>cm</u>				
	2 (Current	1		6.762351485		Grain yield	GY Computation	<u>t/ha</u>	~			
5 m 30 ft	1				0		Land Lea	let Map data (D OpenStre	eetMap contri	butors, CC-BY-S.	A
		Stuc	ly Info					Values				
Study Name:						Watkins mapp Harvest 2020	ing populations for y	eld, NUE ar	id associ	iated trait	s, RRes,	
Study Descri	ption:											
Programme:						Designing Futu Graham Moore	ure Wheat					
Field Trial Na	ime:					Watkins mapp	ing populations for y	eld, NUE ar	id associ	iated trait	s, RRes	
Study Design	1:											
Study Design Team:						Andrew Riche						
Study Design Team: Phenotype G	n: Siathering Notes:	:				Andrew Riche						
Study Design Team: Phenotype G Slope:	:: athering Notes:	:				Andrew Riche						

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 <u>Crop</u> <u>Ontology</u>



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 <u>https://grassroots.tools/frictionless-data/</u>
- Other DFW work on Frictionless Data by Richard Ostler at Rothamsted Research



Field Trials - Study Frictionless Data Package

Grassroots Infrastr	ucture 🗙 grassroots.tools/dev/frictionless/\\\ 🗙 🕂 👘 👘 👘
< → ℃ @	🔒 https://grassroots.tools/dev/frictionless/Watkins mapping populations for yield, 110% 🚥 😒 🏠 💵 🗈 🗧
JSON Raw Data	Headers
Save Copy Collapse Al	I Expand All (slow) 🗑 Filter JSON
▶ name:	"Watkins mapping populati…its, RRes, Harvest 2020"
id:	"5f8eea7402700f64852ae917"
description:	null
profile:	"data-package"
resources:	
▼ 0:	
<pre>profile:</pre>	"https://grassroots.tools…ials/trial-resource.json"
id:	"5f7f211802700f06571db614"
name:	"Watkins mapping populati…associated traits, RRes"
team:	"Andrew Riche"
programme:	"Designing Future Wheat"
v 1:	
<pre>> profile:</pre>	"https://grassroots.tools…/programme-resource.json"
id:	"5fb7964002700f4a785acd44"
name:	"Designing Future Wheat"
abbreviation:	"DFW"
url:	"https://designingfuturewheat.org.uk/"
pi_name:	"Graham Moore"
pi_email:	"Graham.Moore@jic.ac.uk, "
logo:	"https://designingfuturew…opped-DFW-logo-32x32.jpg"
▼ 2:	
<pre>> profile:</pre>	"https://grassroots.tools…ials/study-resource.json"
10:	"5+8eea/402/00+64852ae91/"
name:	"Watkins mapping populati…its, RRes, Harvest 2020"
<pre>> field_trial:</pre>	"Watkins mapping populatiassociated traits, KKes"
location:	"Meadow, Kothamsted Experimental Farm"
crop:	wheat"
previous_crop	is wheat
<pre>procs_gps:</pre>	(w) [1]
= 3.	
y p;	"tabulan data nangunga"
provine.	
b data:	
<pre>dialect:</pre>	
- didiceet	(m)

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



Field Trials - Plots Schema Fields

\leftrightarrow \rightarrow C \textcircled{a} https://g	grassroots.tools/dev/frictionless/Watkins mapping populations for yield, 110% 🚥 🗵 🛔 📗 🗉
JSON Raw Data Headers	
Save Copy Collapse All Expand All	I (slow) Trilter JSON
profile:	"tabular-data-resource"
🔻 schema:	
✓ fields:	
▼ 0:	
name:	"Plot ID"
title:	"Plot ID"
type:	"integer"
<pre></pre>	"The ID of the rack. This is a number given to uniquely identify each rack in the Study similar to a primary key in a database. If GeoJSON and/or images are available, this will be used to identify which plot this information refers to."
<pre></pre>	
minimum:	1
unique:	true
required:	true
v 1:	
name:	"Sowing date"
title:	"Sowing date"
type:	"date"
description:	"Sowing date of the plot"
▼ 2:	
name:	"Harvest date"
title:	"Harvest date"
type:	"date"
description:	"Harvest date of the plot"
▼ 3:	
name:	"Width"
title:	"Width"
type:	"number"
description:	"This is the width, in metres, of each plot"
<pre> constraints: </pre>	
minimum:	
• 4:	11 an abh 1
name:	Length "
title:	Length
description.	"This is the length in meteor, of each plat"
= constraints:	This is the tength, in metres, of each piot
minimum:	8
× 5:	ž.
name:	"Row"
title:	"Row"
type:	"integer"
description:	"Row number of the plothand edge of the plots."
constraints:	{_}

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

Standard attributes

Length

Ξ

- Width
- Position
- etc.

Custom attributes

- Treatments *e.g.* fertilizers
- Phenotypes



Field Trials - Plots Tabular Data

Grassroots Infrastructure	X grassroots.tools/dev/frictionless/W X + ×
← → C û 🔒 https:/	//grassroots.tools/dev/frictionless/Watkins mapping populations for yield. 110% 🚥 🗟 🛝 🗊 🕥 🚍
JSON Raw Data Headers	
Save Copy Collapse All Expand	All (slow) 🛛 Filter JSON
▼ 3:	"tabulan_data_parounce"
v schema:	cabular -uaca-resource
<pre>> fields:</pre>	[]
title:	"Plots"
🔻 data:	
▼ 0:	
Row:	1
Column:	1
Length:	1
Width:	1
Sowing date:	"2019-11-21"
Harvest date:	"2020-08-10"
Plot ID:	1
Rack:	I #/Decesses
Accession:	(Paragon x walbcoozi)_0045"
Nat dto day:	"2020-08-02 (2020-08-02)"
GEP calc day:	"48.4"
TTA_Calc_Cday:	"1752.15"
TTM_calc_Cday:	"2580.85"
PH_M_cm:	"0.87671608"
GY_Calc_tha:	"6.365641345"
HI_Calc_pct:	"42.60693941"
BM_Calc_tha:	"14.94038631"
Awns_E_0to9:	"2020-07-06 (0)"
w 1:	
Row:	1
Column:	2
Length:	1
Width:	1
Sowing date:	"2019-11-21"
Harvest date:	2020-08-10"
Piot ID:	1
Accession:	1 "(Paragon x WATDE0011) 0014"
Replicate:	1
Ant dto dav:	"2020-06-17 (2020-06-17)"
Mat_dto_day:	"2020-07-24 (2020-07-24)"
GFP_calc_day:	"37"
TTA_Calc_Cday:	"1802.3"
TTM_calc_Cday:	"2418.25"
PH_M_cm:	"0.811513824"
GY_Calc_tha:	"6.060947109"



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

f 🍠 🞯 in 🛗 🗟 🐱

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





Decoding Living Systems