PhenomUK Data Strand Report

Report mapping the community requirements on Software and Data

Authors:

Sotirios A Tsaftaris (s.tsaftaris@ed.ac.uk)
Valerio Giuffrida (valerio.giuffrida@nottingham.ac.uk)

4 January 2024

Executive Summary

This report aims to provide recommendations of the software provision required to run and support the digital research infrastructure for the PhenomUK community. To have a better idea of the needs of the UK plant research community, we conducted several interviews and collected information on three major aspects: (i) data collection; (ii) software tools; (iii) FAIR data. After several months of gathering insights from researchers via interviews, we found the following points as the current pressing issues the community faces: (i) Manual software pipelines cause bottlenecks in data analysis; (ii) Bespoke code gets forgotten, becoming legacy in a short timeframe; (iii) Data storage is always insufficient as needs grow; (iv) Data curation and sharing is still an issue for many. To address these community needs, we propose a set of recommendations to inform future directions of the PhenomUK Scoping Project.

I. Introduction

It is undeniable that the UK plant phenotyping & crop research communities require computational power to manage, organise, and analyse data acquired during an experiment. Knowing the typology of data and software these communities currently use, will give us a better understanding of how to shape the proof-of-concept of the Digital Research Infrastructure (hereafter referred to as DRI) we aim to set up in the BBSRC-funded scoping project "*PhenomUK-RI The UK Plant and Crop Phenotyping Infrastructure*".

To this end, we conducted a series of interviews with several researchers and stakeholders within the plant community that allowed us to get better insights into the current data and analysis needs. Furthermore, we also interviewed developers of three major software solutions that have the potential to be utilised within the proof-of-concept DRI, offering researchers and stakeholders a showcase of what our framework will be able to provide at large. Hence, we categorised those interviewed into two main groups:

- **Users:** researchers/stakeholders utilising specific software needed for their experiments;
- **Developers:** research groups that are actively involved in the development of a software resource that can be utilised as a backbone for a DRI.

This report is organised as follows: Section II provides details on the questions we asked the two groups and how interviews were conducted. Section III provides an overall discussion of the information we gathered during the interviews. Finally, we provide specific recommendations in Section IV.

II. Approach

The major goal of interviewing researchers and stakeholders in the plant community can be summarised with two overarching questions:

- *What kind of data do people acquire?*
- *How are these data analysed?*

Therefore, we crafted our interview approach accordingly, by gathering information and insights from users and developers. With the goal defined, we prepared the questions and tailored them on a case-by-case basis. Although the list of questions was adapted to each interviewee, questions were split into three major categories: *Data collection, Software tools, FAIR data* (there was a fourth category called *conclusions,* where we asked any unplanned questions that could have arisen during the meeting). Questions were provided to the participants before the meeting to optimally use the allocated timeslot for each interview. Below, we provide a few sample questions asked (*users* group):

- **Data collection:** *What kind of data do you acquire in the facilities you are involved with? Where do you store data? What data storage infrastructure do you use?*
- **Software tools:** *What software do you use to analyse data? Are these software solutions AI-driven? Do you think that the current software provision you use at the moment satisfies the needs of your team? What are the limitations?*
- **FAIR Data:** *Have you released (or supported the release of) a FAIR dataset? If not, what hindered you to do so?*

During our scoping activity, we identified three major software solutions that are suitable candidates to run the backbone of the PhenomUK-RI:

- **Grassroots:** https://grassroots.tools;
- **PHIS:** http://www.phis.inra.fr;

UK Research and Innovation

Biotechnology and Biological Sciences Research Council

Project funded by BBSRC (BB/Y512333/1)

- **FAIRDOM-SEEK:** https://seek4science.org.

> **Other alternatives:** We came across other software solutions. However, we did not explore them further as we considered them unsuitable to run a nationwide digital research infrastructure. Below, we report their name with a justification that led us to discard them as suitable options:
>
> **PIPPA:** Database with web interface for plant phenotyping data management. Initially developed by Ghent University, its development was taken over by VIB Agro-Incubator. It runs already as a web service, but the source code seems to be unavailable. Given the new developers are from the industry, we suspect PIPPA would become a proprietary product in the future.
>
> **GridScore-Next:** Web and smartphone app for plant phenotyping data collection developed by the James Hutton Institute and the University of Dundee. It is focused on assisting researchers with in-field data acquisition. It is unclear if data acquired in the past can be ingested into the database. Moreover, given it is built using mainly web technologies, we fear it may not be suitable to run as the backbone of a nationwide research infrastructure.
>
> **FAIDARE:** Web application for phenotyping data search using BrAPI.[1] Code is available, but its current development status is uncertain (most links referenced within are unresolved). We will consider in the future whether to adopt it or not to improve the findability of data maintained in our infrastructure.

We interviewed lead developers of suitable software frameworks adapting questions as needed as well as user groups. The number of interviewees varied, as we invited several people from each institute or research group. This arrangement broadened the spectrum and diversity of the research conducted in the UK in plant and crop research.

III. Discussion

We summarise the core points (CP) highlighted during the interviews with the *users* group. From a digital research infrastructure point of view, we found that:

CP1. Data storage is always insufficient as needs grow.
CP2. Data curation and sharing is still an issue for many groups and teams.
CP3. Manual software pipelines cause bottlenecks in data analysis.
CP4. Bespoke code gets forgotten, becoming legacy in a short timeframe.

Below, we will expand on each of the following points.

► **CP1 – Insufficient data storage:** Research projects generate a huge amount of data, especially imaging, sensors, and genetic data, and such generated data are usually archived for many years after the end of a research project. This data *preservation* occurs for two main reasons: (i) acquired data can be useful in the long-term for other (related) future projects; (ii) institutional and/or funder policies.[2] Therefore, there is a pressing need for where to store data, because institutional data storage services are under constant pressure.

► **CP2 – Lack of data curation & difficult data sharing:** Once data are collected, several research groups do not have a standardised protocol to curate it. Some institutions can benefit from dedicated professionals supporting data curation. However, this is not something everyone can currently benefit from. Even though data get curated and well organised, sharing is not always simple, especially when such sharing needs to be done with external collaborators.[3]

[1] Further details on BrAPI are provided later in this report.

[2] UK Research Councils require data to be accessible for (at least) 10 years after their release.

[3] When we refer to data sharing, we refer to the practice of sharing data to external collaborators (e.g., co-Is) before their release to the public (e.g., a publication or other dissemination means).

UK Research and Innovation

Biotechnology and Biological Sciences Research Council

Project funded by
BBSRC (BB/Y512333/1)

► **CP3 – Manual software pipelines:** If we use plant image analysis as an example, plant researchers need to undertake the following steps[4]: (i) plant segmentation; (ii) data extraction; (iii) data analysis. Plant segmentation is the computer vision task needed to identify a plant (or specific organs) from an image. A visual example of plant segmentation is offered in Figure 1. When this operation is completed, each delineated part of a plant is used to extract quantitative data (typically in the form of measurements). When data is extracted from all the images in the experiment at hand, statistical analysis is undertaken to prove (or disprove) a biological hypothesis that was previously set. These steps are performed with different pieces of software, generating a data analysis pipeline that is mostly done by hand. This means that images are segmented with one or more tools; then the outputs are provided to the next tool(s) to get phenotyping information; lastly, phenotyping information is analysed for statistical hypothesis testing, by typically using yet another piece of software.

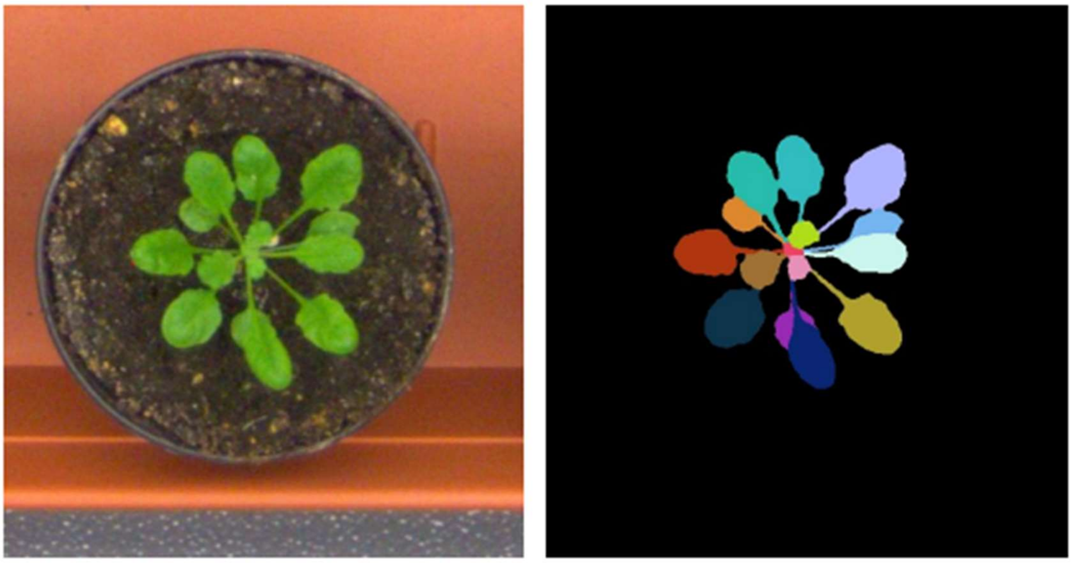

Figure 1. An example of image segmentation in a plant, where each leaf is individually delineated (segmented) and identified.

► **CP4 – Legacy code:** Most of the time, research institutions make their own bespoke codes that are focused on solving one or more simple tasks in the context of a specific project. When the project is over, the code developed is likely not used anymore, becoming quickly obsolete as the computing technology progresses. It is rather common for research institutes to have a repository with legacy codes. This causes the problem of *reinventing the wheel*, as well as adds economic costs to pay for professionals to make such bespoke programs.

In the next section, we provide some recommendations aimed at providing solutions to these highlighted issues, within the budgetary and time constraints of the PhenomUK Scoping Project.

IV. Proposed Recommendations

The proposed recommendations are based on the interviews discussed in Section **Error! Reference source not found.**, as well as the result of a Mentimeter poll we ran during the first PhenomUK Conference held in September 2023.[5] A detailed report on the responses collected on Mentimeter is offered as a separated document.

[4] These steps are examples and not the general protocol followed by any group.

[5] Further information is available at: https://phenomuk.org/uk-plant-phenomics-2023-conference/

Project funded by BBSRC (BB/Y512333/1)

► **Suggestions addressing CP1:** It is important to define what the PhenomUK-RI wants to be like in the future. From a DRI perspective, what should be our major aim? Since this decision cannot be solely made by the strand leaders, we propose the following options:

- **Option 1** [Full-scale Centralised Data Hosting]: With this option, a DRI hosts large quantities of data. Data quotas can be allocated at the user or group level, such that all the participants will have a fair share of archival space. It is important to highlight that any data storage assurance has to be doubled to accommodate for data backup.[6] Catastrophic archival backup is not considered here.
- **Option 2** [Decentralised Data Hosting]: This option asks each participating institute to outsource a portion of their data storage for the PhenomUK-RI. We then leverage all of these collective space contributions to support the proof-of-concept DRI.
- **Option 3** [Limited Centralised Data Hosting]: Taking also into consideration the Mentimeter response in Figure 2: Mentimeter response on what to prioritise more between Computational Power and Data Storage. Figure 2 (data storage had more votes than computational power by a slight margin), this option aims to balance data storage and computational power. Considering the scoping exercise, data storage will be delivered to accommodate the adopted data exemplar used during this project to showcase the viability of the PhenomUK DRI.

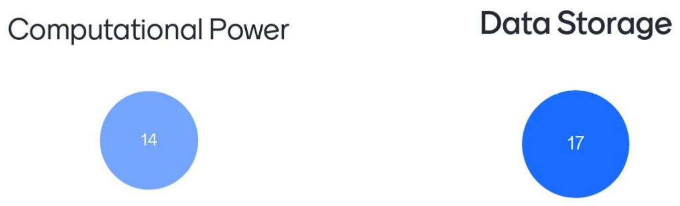

Figure 2: Mentimeter response on what to prioritise more between Computational Power and Data Storage.

All these outlined options will require different hardware; specs (together with costings) will be provided in the *Hardware Report* that will follow this report.

Taking into consideration the interviews and the engagement with the community, as well as the constraints of this scoping project, we recommend **Option 3.**

► **Suggestions addressing CP2:** We are currently scoping three potential software solutions that can provide support for data sharing and data curation. Each solution has strengths and limitations, which are summarised in the following tables.

GRASSROOTS[7]	
Strengths	**Weaknesses**
Modular design	Limited features
Uses iRODS	No user authentication
Includes a gene sequence service	Different databases are used, making data retrieval/linkage cumbersome and hard
Development assured for the next 5-6 years	
PHIS[8]	
Strengths	**Weaknesses**
Large development team	Difficulty in uploading large datasets
Expected to be used by EMPHASIS	Data ingestion enforces metadata
Provides a wide range of features	Unstable
FAIRDOM-SEEK[7]	
Strengths	**Weaknesses**

[6] If our DRI provides 1PB of data storage to all users, we have to account for an additional 1PB for backup.

[7] Details on GRASSROOTS and FAIRDOM-SEEK pros and cons are in Appendix A & Appendix B.

[8] At the time of the writing, we have not had the opportunity to test PHIS ourselves. The provided list of strengths and weaknesses are based on the users' feedback we received during interviews.

User authentication/profile available	Not specific to plants/unneeded features
Good search system	Unclear how modular or customisable it is
MIAPPE Compliant	Search via MIAPPE metadata seems not doable

A detailed overview of all these software solutions is provided in Appendix C. Based on the limitations outlines above, we propose the following:

- **Option 1** [GRASSROOTS]: Under this option, we will adopt GRASSROOTS to run our digital infrastructure. We will invest the time of our software engineers to develop necessary features needed to deliver a proof-of-concept software framework.
- **Option 2** [FAIRDOM-SEEK]: Under this option, we will instead run FAIRDOM-SEEK. As this software includes a wide range of features, we will find that some of these might not be necessary for the plant community. Hence, we will invest resources to strip it down and include any necessary features needed to deliver a proof-of-concept.
- **Option 3** [PHIS]: Although we are aware of its major limitations, with this option we will adopt PHIS in the digital research infrastructure. Resources will be invested to strip unnecessary features and find solutions to interoperability and data ingestion, as these two aspects were highlighted as potential issues by interviewed PHIS users.
- **Option 4** [Multiple solutions]: As we are intrigued by the modular design characterising both GRASSROOTS and FAIRDOM-SEEK, this option will explore the possibility to adopt certain features of each to run the digital research infrastructure. Resources will be focusing on identifying key useful features on each of the two framework and devise a way to facilitate interoperability between these two frameworks.

Taking into consideration the interviews and the engagement with the community, as well as the constraints of this scoping project, we recommend **Option 4.** However, if any software limitations from both software frameworks hinder the whole development, we will switch to either **Option 1** or **Option 2**, planning based on:

- Focus groups with stakeholders aimed at gaining further expert knowledge.
- Data exemplars provided by the community.

► **Suggestions addressing CP3 & CP4:** We believe that current advancements in AI can be leveraged to alleviate the manual pipelining/workflow, as well as limit the proliferation of legacy code in the future. We see an opportunity to leverage foundation models [1] to address these two core points, with a visual example displayed in Figure 3.

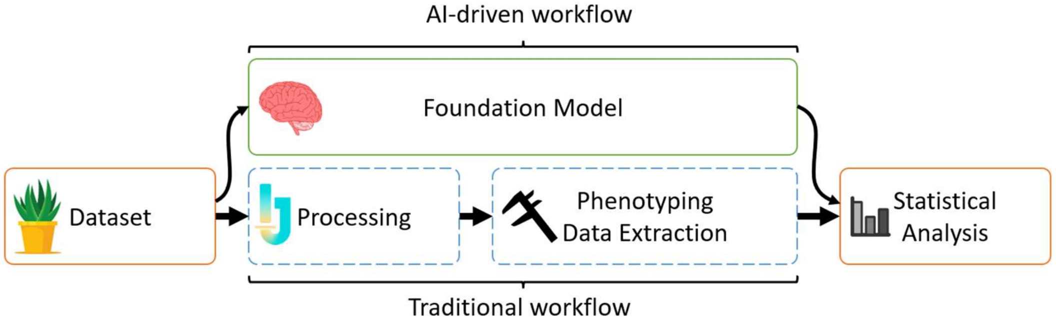

Figure 3: AI replacing traditional workflows.

Once a dataset (*e.g.* a stack of images) is acquired during an experiment, in a traditional workflow they are processed and data extracted using either off-the-shelf programs (e.g., ImageJ) or bespoke code.[9] However, this collection of software tools has been identified as a bottleneck during our interviews. We argue that we can replace all of them with a single AI model, that extracts actionable data from images ready to be used for further statistical analysis (*e.g.*, with the use of R).

[9] It's not excluded that several researchers may employ off-the-shelf programs and bespoke scripts at the same time within the same research project.

Such an AI-based approach built upon a foundation model can also replace existing legacy code, minimising its proliferation in the future.

In the first 6 months of the PhenomUK Scoping Project, we have explored this opportunity, and we collected several preliminary results that will be soon published and shown during the *Computer Vision in Plant Phenotyping and Agriculture* (CVPPA)[10] workshop, held in conjunction with ICCV 2023. This milestone will lay the path towards the mitigation of these two core problems.

Acknowledgements

We thank all the collaborators for their tireless and invaluable contribution to the Data Strand: Feng Chen & Andrew Wood. We also thank to all the people who accepted our invitations and replied for 1hr to all our questions which constituted a big portion of our scoping activity in the first 6 months of the project. Without their insights, most of this work could not have been achieved. Finally, we also thank all the people who replied to our Mentimeter poll during the first PhenomUK conference.

References

[1] Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

[10] https://cvppa2023.github.io/

PhenomUK Data Strand Report

Appendix: Report mapping the community requirements on Software and Data

Authors:

Sotirios A Tsaftaris (s.tsaftaris@ed.ac.uk)
Valerio Giuffrida (valerio.giuffrida@nottingham.ac.uk)

Appendix A. Detailed report on GRASSROOTS

Grassroots is made up of a core system from which different services can be added. Currently the main services are:

- BLAST – protein/nucleotide sequence queries and searches
  - BLASTN
  - BLASTX
  - BLASTP
- Field Trials – plot data for crop experiments
  - Search
  - Submit
  - Search treatments
  - Search measured phenotype
  - Search Locations
- Data Portal – stores genomic (and any other) data from experiments
- CKAN – stores papers etc

Each of these is stored in separate databases at the minute, although there is a separate 'Search' service which is able to look for data in any of the locations. All services communicate with each other through JSON files, meaning that new services can be written in any language, which should make it relatively easy to extend or expand and needs change.

The current Field Trial set up is close to what we are looking for although currently only applicable to large (in-field) trials. Studies are created with a defined number of plots, which can be associated with different treatment options. Plot data is for a given study is easily downloadable as a CSV file, or as frictionless data (json) files. As the system was designed for crop research, there are already a large number of things in place which would be very useful for other plant phenotyping work. The Field Trials data follows BrAPI v2 standards (compatible with MIAPPE), and Grassroots aim to keep in contact with BrAPI to add new fields etc. Plant Treatments use existing ontologies (https://browser.planteome.org/amigo), which would still apply to a wider range of plants and lab settings. Finally, Grassroots already have the BLAST gene query services set up, and they have indicated plans to tie this further into the Field Trials service. For example the genome of a particular Plot could be compared to the existing databases if the measured phenotype value is particularly notable.

Currently, we think there are 2 major problems:

- **User Authentication** – Anyone can edit any of the Field Trial data, and all data is publicly shared immediately. A user authentication system (including associations to project groups or institutions) would allow much more security and flexibility in what can be seen. This would also solve some of the issues which are likely to arise as Grassroots scales up – namely that they will add more fields and options to cater to more people, but currently they are seen by everyone. If new fields can be toggled for different user groups or studies depending on their work then this won't be an issue as much.
- **Data Storage and Findability** – It is very unclear what information is stored where. There are many field trials which do not have any extra data in the data portal, and there are many items in the data portal which do not appear to be associated with a particular field trial (This may be the case if they are looking at BLAST things perhaps?). Similarly, it is very difficult to find/know which (if any) field trials have papers associated with them stored on CKAN. Currently uploads to the Data Portal can only be done by Grassroots, so other data may be hosted elsewhere on private servers. This means that it would not appear in search results and is therefore harder to access by other users. Furthermore, if users are interested in exploring work looking at a particular treatment, this is currently difficult, as the search results will show you the treatment, but there is no way to see which studies have recorded that treatment. Finally, the current search system does not seem to be particularly robust, as all words appear

to be treated equally (including 'the' and 'and'). This leads to receiving many more results than may be wanted, so it is harder to find the information you are looking for.

All (identified) problems with the current setup:

- There is no user authentication required (other than an ORCID), and all data is public. This means that data can be edited or overwritten by any user.
- It is unclear what is stored where, and what is searchable. Items in the CKAN database do not seem to appear in search results.
- Currently users cannot upload to the data portal directly – things are sent to grassroots and uploaded manually.
- Items other than plot data cannot be uploaded directly into the field trial system. Instead links to their location are uploaded. Often this is a link to the data portal location, but this can also be to a privately hosted location.
- As the extra files are stored wherever people choose to store them, there is less guarantee of their accessibility. Currently the plot results themselves have metadata associated, but accompanying pictures or other files may not.
- The current search method is very simplified – every word appears to be weighted equally and gives many more results than may be desired because of this.
- While it is possible to link from a trial/study to the data portal or CKAN to get more information, it is not possible the other way round. If you have found a dataset in the data portal, there is no clear way to see the Field Trial which it is associated with.
- Searching for the sub-items of a Field Trial (Location, Treatment, Measured Variable) does not allow you to see the Studies which contain those items.
- It is very unclear which data in the Data Portal is associated with field trials and which is stored on the system associated with other work.
- There is no version control in the system – if data has been changed or overwritten, it is not possible for users to see that.
- The system is currently designed for field data, not lab data, so many of the data inputs would be different. Without user authentications and an ability to filter/select the desired fields it is likely that users would skip through things.
- Currently every study requires a 'Location', which must first be added to the database. There is not any duplication checking on this resulting in the same location being present multiple times.
- The frictionless data requires a separate tool to extract and view it which only runs from the command line. I'm not sure how viable this is for a larger scale release.
- The search and submit services are separated at the minute. This means once you have found a Study in the search, you need to select it again the in the Submit service to be able to edit it.
- The grassroots documentation is poor at the minute, which may slow down any future development.

Appendix B. Detailed report on FAIRDOM-SEEK

<u>PROS</u>

- Has user profiles to filter access by project/institution/user.
- Already used by EOSC-Life Plant A+ Demonstrator project (ELIXIR). For this Fairdom seek was extended with MIAPPE metadata support.
- It can be used on small scale (<1TB) through public FAIRDOMHub, or built on a custom private Hub.
- Adding support for European Nucleotide Archive.
- Good search system to find known studies/ experiments.
- It looks like they provide support and consultancy options to help with projects.

<u>CONS</u>

- Not plant specific – has MIAPPE metadata as an option, but has to always be selected. Likely to be other extraneous features as well.
- Fairdom Hub response times can be quite slow – I don't know if this would also happen on a private server.
- Cannot search by MIAPPE metadata fields.

Appendix C. Overview of all software solutions

Grassroots:

Current benefits:

- Already has set up:
  - treatment ontology list
  - measured variable ontologies
  - accession seed bank
  - follow BrAPI
- Existing CSV plot data visualisation
  - The data is already being read, which should hopefully make filtering/searching it easier
- Current Field Trial system should be extendable to other growth settings

Current downsides:

- The existing search methods are weak and would need large overhaul
  - It is difficult to find data/experiments for a given search term
  - Having things spread across multiple databases makes it harder to know where to search
- There is not a strong link between the various databases.
  - Currently while looking at a dataset, it is difficult to see what studies used it
- It is not possible to upload additional documents (i.e. images) to the database directly – need to contact the grassroots team, or store them in a private database
- User identification and user groups are not implemented
  - This is currently being added, but we don't know what form this will take
- Not sure there are checks in place to ensure the ontology codes are valid

Unknown:

- Uses MongoDB
  - More flexible due to lack of need for fixed schema
  - May make searching more difficult if users can use their own terms?

FairdomSeek:

UK Research and Innovation

Biotechnology and Biological Sciences Research Council

Project funded by
BBSRC (BB/Y512333/1)

Current benefits:

- Has good user/project interfacing to control data permissions
- Single database and good search functionality (missing MIAPPE search terms though)
- I believe we should be able to utilise the existing Sample methods, or extend the existing Assays methods to allow for more data/experiement types
- Easy to upload/download large data files

Current downsides:

- Addition of MIAPPE metadata is incomplete – MIAPPE batch template exists, but upload doesn't work. This includes ontologies for treatments and observations
- Somewhat confusing regarding what to store where – ie samples, strains collections, organisms etc. (Note – this may be easier for users with more knowledge of the field)

Unknown:

- Uses MySQL
    - More rigid schema control should make searching and comparing results easier
    - Requires schema creation

Phis:

Current downsides:

- Limit on file size of additional data (i.e. images), may require secondary database
- Old public version of PHIS regularly runs into internal errors when searching. This seems to be resolved on the new version.
- OpenSilex (new version) is down right now at time of writing – not sure if that represents an underlying issue

Unknown:

- Uses hybrid database system

All three methods do not allow for searching/filtering the data to download specific treatments within a dataset for example. Currently, Grassroots is closest to this, as it reads the data and presents plot visualisations. However, comparing cross-study datasets may be easier to implement if everything were stored in an SQL database with a fixed schema, which would suggest using Fairdom-SEEK. Furthermore, the Mentimeter output also suggests that ease of use is a very important factor. Currently, fairdom seek is the only system whereby all files associated with a particular experiment could be uploaded to the same source. (Note: I have only tested this on files up to 10GB).

The data fields in grassroots are currently specific to field trial data, so would need to be tailored to different experiment types to prevent overwhelming users with unnecessary fields.

UK Research and Innovation

Biotechnology and Biological Sciences Research Council

Project funded by BBSRC (BB/Y512333/1)