Check for updates

SOFTWARE TOOL ARTICLE

Introducing the Brassica Information Portal: Towards integrating genotypic and phenotypic Brassica crop data [version 1; referees: 3 approved]

Annemarie H. Eckes[1], Tomasz Gubała[1,2], Piotr Nowakowski[1,2], Tomasz Szymczyszyn[1], Rachel Wells[3], Judith A. Irwin[3], Carlos Horro [1], John M. Hancock [1], Graham King[4], Sarah C. Dyer[5], Wiktor Jurkowski [1]

[1]Earlham Institute, Norwich, NR4 7UZ, UK
[2]Academic Computer Centre CYFRONET, AGH University of Science and Technology, Kraków, 30-059, Poland
[3]John Innes Centre, Norwich, NR4 7UH, UK
[4]Southern Cross Plant Science, Southern Cross University, Lismore, NSW, 2480, Australia
[5]NIAB, Cambridge, CB3 0LE, UK

Abstract

The Brassica Information Portal (BIP) is a centralised repository for Brassica phenotypic data. Trait data associated with Brassica research and breeding experiments conducted on Brassica crops, used as vegetables, for livestock fodder and biofuels, is hosted on the site, together with information on the experimental plant materials used, as well as trial design. BIP is an open access and open source project, built on the schema of CropStoreDB, and as such can provide trait data management strategies for any crop data. A new user interface and programmatic submission/retrieval system helps to simplify data access for scientists and breeders. BIP opens up the opportunity to apply big data analyses to data generated by the Brassica Research Community. Here, we present a short description of the current status of the repository.

This article is included in the Open knowledge in agricultural development channel.

This article is included in the GODAN channel.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees		
	1	**2**	**3**
version 1 published 12 Apr 2017	✓ report	✓ report	✓ report

1 **Christopher J. Rawlings** , Rothamsted Research UK

2 **Cyril Pommier** , French National Institute for Agricultural Research (INRA) France

3 **Matthew M. Nelson** , Royal Botanic Gardens Kew UK

Discuss this article

Comments (0)

Corresponding author: Wiktor Jurkowski (wiktor.jurkowski@earlham.ac.uk)

Competing interests: No competing interests were disclosed.

Introduction

Brassica is a diverse genus that includes a number of species cultivated as important crops, providing edible roots, leaves, stems, buds, flowers and seed[1]. Due to their wide adaptation and ability to thrive under varying agroclimatic conditions[2], Brassica crops are grown all over the world for food, animal fodder and industrial applications.

Given the importance of Brassica species worldwide, integrated approaches to research and plant breeding for crop improvement are required to address future challenges, such as satisfying increased demand for higher quality or producing reliable yields in a more variable environment with sustainable, lower input cultivation systems. Researchers have quantified many aspects of Brassica phenotypic diversity, including the content of industrially important compounds, yield in response to fertiliser input or degree of resilience to fungal pathogens and pests, such as flea beetles, to aid in crop improvement. The ability to host such data generated from these experiments in an open and accessible database promotes the comparison of trait measurements and calculation of genotype-phenotype (G X E) associations across multiple studies in meta-analyses, potentially identifying underlying common mechanisms or previously unrecorded correlations. This is especially important for plant breeders and researchers involved in pre-breeding, leading to generation of new crop varieties. For example, meta-analysis across different populations or environments can help confirm the location of key areas of the genome identified from multiple studies or assist and help avoid potential trait trade-offs of beneficial or deleterious alleles due to linkage drag. The ability to associate phenotypic data with genomic regions, DNA markers, and genome and transcribed sequence provides the potential to expedite the integration of favorable alleles via marker-assisted or genomic selection. However, few plant phenotype databases exist, despite the importance of phenotypic data to render meaning to sequence data.

Maintenance of easily accessible and reusable plant phenotypic data is challenging because, in contrast to standardised approaches used in sequence analysis, traits can be measured with a variety of protocols whose methods are not captured automatically, and there is mostly no consensus on methodologies to record any given trait. Moreover, there is a lack of controlled vocabularies to describe phenotyping metadata adequately. One possible reason is that, as opposed to publishing sequencing data in databases widely used by the research community, such as ArrayExpress[3], GEO (Gene Expression Omnibus)[4], ENA (European Nucleotide Archive)[5] and SRA (Sequence Read Archive)[6], it is not a requirement to submit phenotyping data to a public repository to accompany peer-reviewed publications. There is, to date, no single plant phenotypic database recommended by Nature's Scientific Data. Compared to sequence data, the formal description of trait and associated meta-data tends to be more complex and crop or species specific.

Data re-use and re-analysis can lead to new knowledge and enhancement in understanding. For example, readily available datasets retrieved from a database can serve 1) as evidence for new claims, 2) to explore existing knowledge and identify gaps, 3) to bridge between research domains or 4) to help with selecting research directions. Additionally, databases can serve as tools for collaboration across countries, fields or areas of expertise and link between research and industry.

Existing Brassica-related databases and analysis tools mostly focus on Brassica genomic information. For example, BRAD[7] hosts genomic data on Brassica genomes. Other databases are projector institute-specific, such as PlabiDB[8], a crop genomic platform that, among others, hosts genomic data from *Brassica napus*, and GnpIS that, besides genomic information, hosts some phenotype data that is not publicly available. While these databases are useful in their own context, they do not focus on setting or demonstrate the value of community-wide phenotypic data standards, and therefore do not encourage universal trait sharing and data re-use.

Resources, such as TAIR (arabidopsis.org)[9], or Araport (araport.org)[10] and the 1001genomes project[11], provide data storage and analysis tools for the Arabidopsis Scientific community. Due to the close evolutionary and taxonomic relatedness of the model plant Arabidopsis in the Brassicaceae, Arabidopsis-related databases or platforms for genotype and phenotype information provide valuable insights into the underlying Brassica biology, and potential insights into the crop analysis tools and resources of the future. However, current Arabidopsis resources are not directly applicable to recording genotype-phenotype specific traits of Brassica crops.

CropStoreDB[12], the first systematic Brassica phenotype database, was launched as part of InterStoreDB steming from previous efforts such as Integrated Marker System for Oilseed Rape Breeding[13], and BrassicaDB[14,15]. CropStore hosts Brassica trait data, connecting it with relevant sequences in SeqStore for genotype and phenotype data integration.

The Brassica Information Portal builds on the CropStoreDB schema, and provides the technologies to enable efficient storage, classification and management of phenotyping data. Its unique features are the universally applicable Brassica trait Ontology, a Digital Object Identifier (DOI) for each dataset, login via ORCiD (https://orcid.org/), cross-references to sequence information and primary data, user-friendly wizard-based submissions, and Application Programming Interface (API) based data submission and retrieval. In addition, submitted data can be kept private for a user-defined period of time. Extensions for visual data exploration and tools for associative analysis are intended in the future releases, to provide a desired set of features for the Brassica community.

Methods

To address the challenges facing a phenotypic trait database, the BIP has implemented rigorous use of ontologies and standards. One element is a dictionary of Brassica-specific traits created according to CropOntology[16] for that purpose. The CropOntology phenotype annotation model is based on the combination of "trait method and scale", as it is grounded in breeder's dataset-formats[16]. This makes it especially suitable for describing pre-breeding material traits hosted in the portal. The level of detail in the trait definition makes it possible to select meaningful traits for analysis. Hence, the CropOntology encourages a way of defining

a trait to facilitate meta-analysis, integrated data analysis and data discovery, some of the main aims of the repository besides data storage.

Further incorporation of ontology includes plant anatomy terms from Plant Ontology[17] and taxonomical classification from Gramene's Taxonomy Ontology[18]. Cultivar names are curated, and submissions are checked against this curated list.

Initial minimal requirements for trial, population and trait description have been identified and are made compulsory during submission, to associate meaningful experimental meta-data with the trait scores. BIP development is benefiting from discussions on standard minimal requirements for plant phenotyping experiment (MIAPPE) data currently ongoing as part of the Excelerate project run by ELIXIR consortium. MIAPPE functions as a checklist of attributes that can be used to describe each experimental unit[19,20] and allows the representation of the most useful information captured by researchers to allow comprehensive data sharing and reuse. A similar approach has been applied previously for other data types, including gene expression microarrays (MIAME), sequencing (GSC), metabolomics (MSI), and proteomics (MIAPE).

In BIP, we are focusing on information related to recording Project, Trial, Sample, Location, Treatment and Trial Environment. Our contribution includes the definition of MIAPPE fields and proposal of ontologies to ensure there is standardised content to describe Brassica phenotyping experiments mapped to these requirements. With regard to file types, the choice of standard input and output file is the .csv format, as it is readable by most software tools and humans alike. Further, the API allows for machine readable .json output format.

The content of the database includes population data with member lines and diversity sets with lines representing cultivars or other *ex-situ* genetic resource material. Linked to this hierarchical relationship of cultivars and lines are trait scores with associated experimental and organisational metadata. As this repository has inherited its structure and content from CropStoreDB, it also hosts historical QTL and linkage map data linked to experimental segregating populations and trait data. For obvious reasons we do not intend to cater directly for sequencing data, yet we continue to host genotyping data generated through modern DNA, RNA sequencing or genotyping arrays. The BIP currently hosts data from numerous pre-breeding programmes that include 22 plant trials, with 139,000 trait scores from 188 different varieties and 26,100 plant lines.

While the repository lays the foundation for more extensive integration of genotype and phenotype information and tools, it is already integrated with TGAC Browser[21] to allow cross-linking of single nucleotide polymorphisms (SNPs) of interest to lines/cultivars in the BIP that may carry the corresponding SNP.

Implementation
There are three main building blocks that constitute the BIP. The first element is the relational database, which stores and manages all BIP data. It uses PostgreSQL RDBMS technology. BIP's initial data and the database schema were obtained from the CropStoreDB system and then underwent several transformations, including translation from the original MySQL format to the target PostgreSQL format, i.e. introduction of formal foreign keys and presence constraints, and a series of data curation procedures to ensure referential integrity. The BIP database is the core of the system, enacting all data management operations, issued both through the web interface and the REST API.

The decision to transfer data to PostgreSQL was taken as it delivers more advanced, modern capabilities of data management than MySQL, while still being an open source and a free-of-charge solution. The overview of the database schema is presented in Figure 1.

The second element is built on top of the database and is a modern web application. Its purpose is threefold: (i) it delivers advanced data browsing interfaces, which allow for data ordering, filtering and export to CSV documents (available to all users without registration; see Figure 2) (ii) it implements data submission wizards, explained in detail in the section below (for registered users only) (iii) and it gives the possibility of managing data through the API, which supports reading, loading, querying, searching, creating and publishing data.

The web application was created using the Ruby on Rails framework. All web requests are channeled through the Nginx web server, which in turn routes them to a web application logic server (a so-called reverse proxy setup). All traffic is secured with an encrypted HTTPS protocol. User registration is implemented using the third party authentication solution (orcid.org).

The third constituent element of the BIP is the data indexing and full-text search engine. It analyzes all records in every table of the database, including all new data being submitted by users, and creates a search index. The BIP web search functionality (present on the front page of the website; see Figure 2) and its API search capability (available via programmatic access to registered users) is integrated with the search engine in order to deliver the arbitrary term search. For every submitted search term, the engine first looks the term up in the index, and if found, responds with a set of datatable records that include this term. This element of the BIP is created using Elasticsearch technology.

All three elements are presented in Figure 3, together with respective data transfer channels between them. Deployment-wise, the database system, the web application logic server, and the indexing and search engine all reside within a single system. If, in the future, the scale of the data managed by the BIP becomes too high for current machine capabilities, it is possible to decouple these elements in order to distribute the load on several physical servers.

Operation
BIP operates at bip.earlham.ac.uk/ and is publicly accessible at this address, as a web resource. However, it is possible to use BIP open source to deploy one's own BIP instance. In such a case, BIP will require a server, either physical or virtual, with at least two
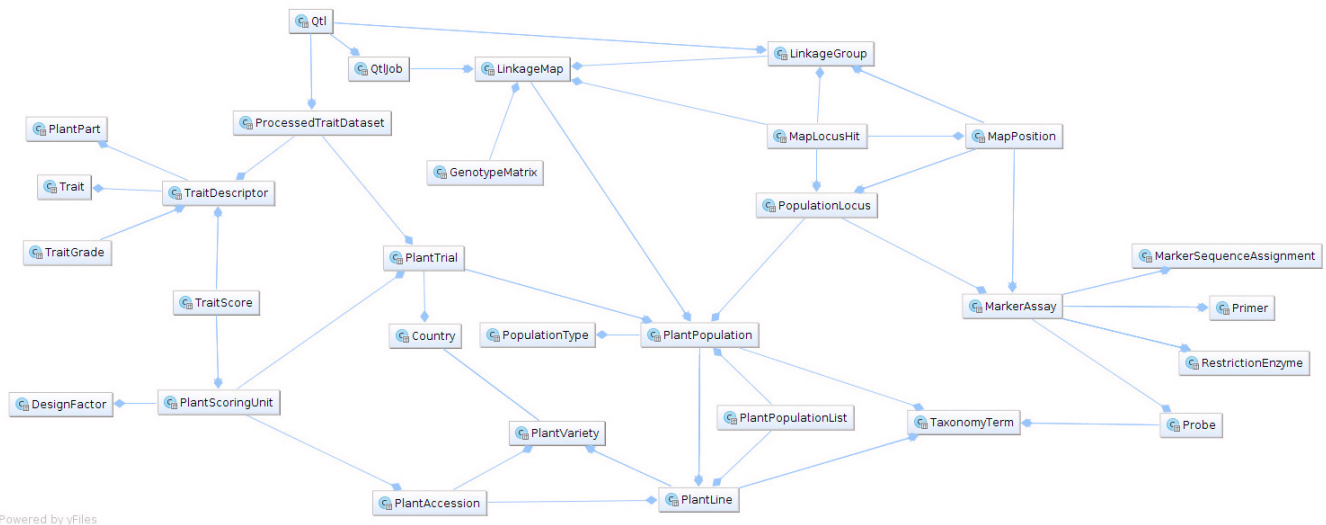
Figure 1. BIP database model. The diagram provides an overview on all tables in the BIP database, and the relations between them. Some "operational" tables, which are used only internally by the system
(e.g. for access rights), were omitted.

Figure 2. Access to the database via the web interface by using a universal wildcard search (upper red circle) or by targeting specific content/tables by following the link on "Browse database" (lower red circle).

contemporary CPU cores, and at least four gigabytes of operational memory. Due to its internal architecture, described in the previous section, two processor cores are required to achieve a smooth user experience along with simultaneous maintenance of fast database and indexing engine responses. Since the DB and the web application itself makes a heavy use of disk caching mechanisms to provide faster response to user queries, it is also beneficial, though not absolutely required, to equip the server with fast hard drives. The storage space needed depends on the amount of data one plans to deposit in the instance of BIP, but the application itself has no
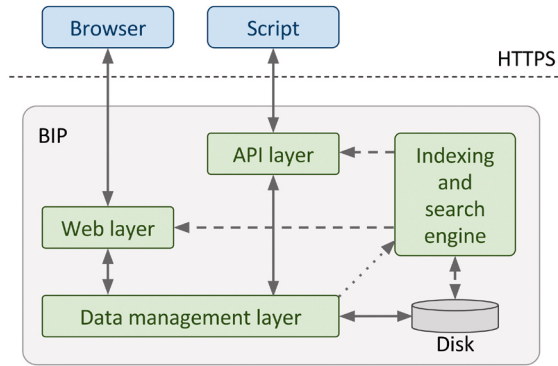
Figure 3. BIP architecture diagram. Users initiate actions with either their web browsers or API clients (blue elements). All requests go through respective access control layers (these ensure e.g. that a given set of data is accessible to the user), before they contact the data management layer to retrieve (or commit) data from (to) the database. Access mode is read/write (solid arrows). Additionally the search engine could be required to search through gathered data (dashed arrows) - it's a read-only access mode. Finally, the data layer informs the search engine about any new data to be indexed (the dotted arrow).

high requirements in this respect, and 6 GB of storage space should be sufficient for both the operating system and all the components of BIP. Any up-to-date Linux operating system will be adequate to host BIP.

Use of the BIP
0.1 Content navigation
Several tables hosting interlinked population, trial, trait score, linkage map, QTL and marker assay data are accessible via the web-interface, either from a universal search, or by browsing the database for specific tables (Figure 2).

Accessing related data is facilitated by following the "Related" button in case of the Populations table, or by clicking onto a cross-linked field within the table (these appear green;

Figure 4A). On top of each table, a search bar can query for specific components within the table (Figure 4A). Organisational and provenance metadata are made visible by clicking the blue information button for each row (Figure 4B). Each column can be sorted, resulting in global reordering across the whole table (Figure 4B, red arrow). The visible content of the tables can be downloaded as a .csv file, by clicking on the respective button on top of the table (Figure 4B).

Submission and download of data
Experimental data is generally recorded in spreadsheet format. Therefore, both the web form wizard provided within the tool, and the available Ruby script client, use spreadsheet formats as templates to assist the user in uploading the data. Data are submitted in two separate logical batches. First, information about the experimental plant population needs to be submitted. Second, trait data from Trials performed on these plants are submitted during Trait Scoring Submission. This enables the submission of multiple trials per experimental Plant Population.

A wizard-based submission of Population and Trial Data data guides the user through all compulsory and optional field contents that can be submitted to the database. During Population Submission (Figure 5A), a four-step process asks the user to enter population related information, including the submission of the actual lines and associated metadata in a .csv template designed in step 3. The purpose of the template is to make it easier to appropriately format large data sets and avoids spelling mistakes in the uploaded file's header. The standard template used during Population Submission is shown in Supplementary Figure 1B. Plant variety names are checked for their existence and correct spelling against the database. The final step is the submission of data provenance information. Compulsory metadata for Population Submissions is listed in Supplementary Table 1.

During Trial Submission (Figure 5B), the first steps involve entering trial related information, before the traits studied in the trial are selected or defined in step 2. To avoid duplicated traits, the interface suggests traits based on an elastic search using two or more letters typed by the user. If traits are unavailable, the user
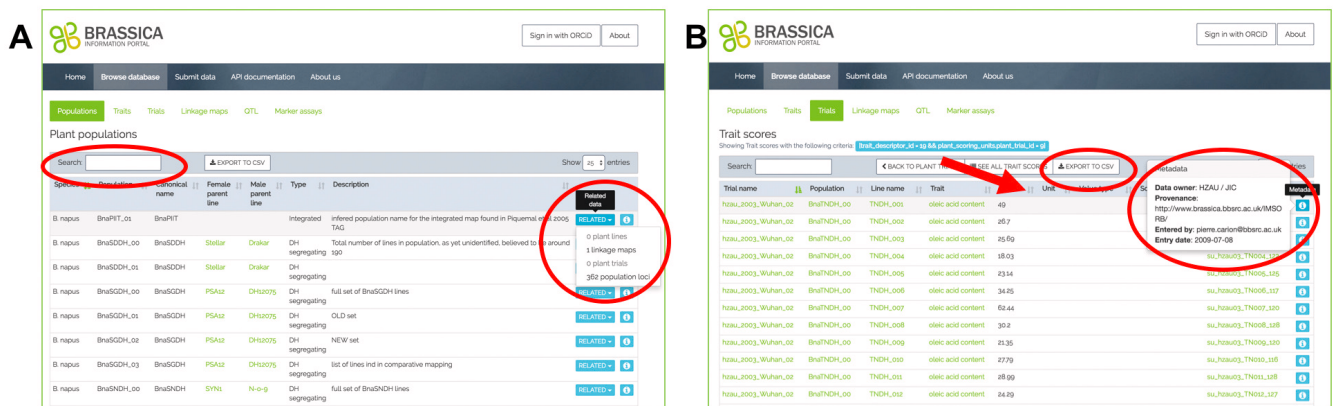

Figure 4. Multiple mechanisms to find, navigate and access data in BIP. A) refined querying and access to related data **B)** reordering of results, function for bulk download of data with accompanying meta-data, display of the meta-data.
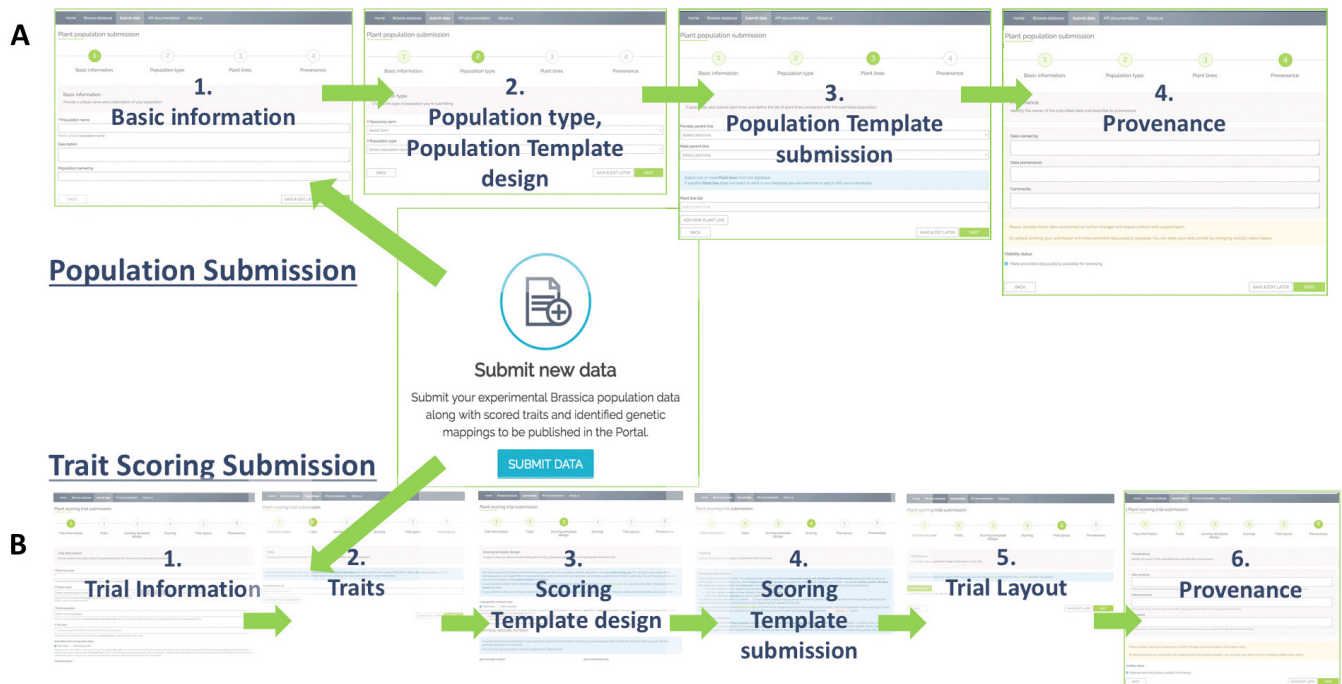
Figure 5. Wizards for Population Submission A) and Trait Scoring Submission B) comprise number of steps collecting required metadata, to generate customised templates for data collection and to finally submit complete data sets.

can define and add them manually (the Plant and Trait Ontology vocabulary is provided to assist the user in this task). A high quality of trait definition and trait measurement-interoperability between experiments is ensured by interposing a human curator between submitter and database.

For the submission of trait measurements, the user designs a .csv template in step 3 (see Supplementary Figure 1A for an example template), and then submits the completed .csv file in step 4. Depending on the type of data submitted (raw vs. analysed data), the template can differ in the fields to be filled out. Compulsory metadata for Trial Submissions is listed in Supplementary Table 1. Finally, the user can upload an image presenting the experimental layout (for instance, the greenhouse plant pot setup on benches) in the case of raw data submission, and fill out provenance metadata in step 6. Whether data are submitted using the wizard or the API, the same metadata is compulsory.

Programmatic submission is executed using POST requests through the BIP API. The query exposes the part of the database data the user wants to submit data to. Content to be submitted needs to be assigned a corresponding field in the database and is presented as files in .json format. Programmatic extraction is also API based. GET requests placed to the database can be more comprehensive or specific than the interface-based strategy. These are then output in .json format, ready to feed into analytic workflows run in, for example, python. Full API documentation is available on the portal's website, including all table fields beyond those that are compulsory.

A Ruby submission client facilitates the automated submission of new experimental populations and trial data. It parses an input .csv file searching for the elements (input columns) to be submitted. These input columns need previously to have been mapped against the database fields by the registered user. For each line in the .csv, new entries are created in the database and this content is then submitted to the database using the BIP API. Using the Ruby client tool is not required - it is provided only as a convenient alternative to users who prefer scripting approach to data submission and analysis.

The registered user is entitled to upload data directly to the database, which may then become public within seconds of submission. However, to ensure confidentiality of data for an ongoing project, at user's discretion, a submission embargo can keep data private for a defined period. Once public and older than the revocation period, datasets could be assigned (on user's demand) with individual DOIs, which act as a reference in publications.

Data download can be performed via the web interface and the API. Using the interface, the user can download any visible table, by clicking the "Export to CSV" button on top of the tables. This downloads the complete table in .csv format. Via the API, the user can use GET requests to interact with the database tables directly. This sometimes enables the user to retrieve more information than is displayed at the interface. The queries can be crafted according to the user's needs for data. Example Ruby clients are available on Github for modification and retrieval of more complex data.

Conclusions

The Brassica Information Portal (BIP) has been designed as an open-access, open-source, phenotype data storage resource and will facilitate addition of other features in the future. The general motivation for the BIP was to develop technologies enabling efficient use of phenotyping data. The creation of the database and its standards lays the foundation for Brassica genotype-phenotype studies drawing phenotypic data from this repository.

Availability of a repository for integrative phenotype data should encourage the Brassica research community to store trait data in an open access repository, in a similar manner as for other data types, such as sequence, metabolomic and protein structure data, and share it with the research community. Ultimately, the Portal will facilitate understanding of Brassica diversity and trait variation for future research, breeding programmes and crop improvement.

With instant submission to the repository, it remains the primary responsibility of the submitter to guarantee the quality of the data. However, cross-linking with existing ontologies and spelling control holds errors in check prior to submission. Further, compulsory fields during submission ensure metadata minimal requirements are met for meaningful data interpretation in the future. Furthermore, each submission is granted a one week revocation period. During that time, the submitting author is asked to perform all important data quality checks, and in case of any discrepancies or inconsistencies are found, the author is able to revoke the publication, repair the submitted dataset, and publish it again. After a week has elapsed since the submission, the data are made read-only, and it can no longer be altered or removed. Moreover, it can then be accessed and quoted by other users as long as the embargo is lifted. The author may request a DOI number assigned to the submission, so it is possible to cite it in publications.

Future plans include the support of submission of imaging data to CyVerse, allowing BIP to serve as a single entry-point for multiple data types relevant to crop improvement studies. At the same time, visualisation tools will be developed to call and display the CyVerse-hosted images from within the BIP interface. This will include 1) images of single plant parts underlying definitions of specific traits; 2) time lapse images of single plants; and 3) time lapse images of field trials generated by high-throughput techniques, e.g. drone flights. Statistical analysis of phenotypic data will be performed with incorporated of R/Shiny applications run directly on the BIP server. For example, association studies could be performed with uploaded genotyping data matching selected plant lines with corresponding trait measurements available in the BIP. More computationally costly operations will be supported through development and incorporation of analytical workflows executed externally, e.g. using CyVerse infrastructure. Additionally, linkage markers recorded in the database will be cross-linked with their corresponding sequences within EnsemblPlants[22]. This will make it possible to visualise the location of the markers in context with the genome sequence and genes in close proximity.

Finally, BIP API will be further refined by implementing additional content checks, e.g. by cross-linking to Global and European cultivar registers (Plant variety database - European Commission, Community Plant Variety Office, CPVO Organisation for Economic Co-operation and Development (OECD) Variety List Query) and to comply with ELIXIR standards (currently under development) to facilitate common data exchange and communication with other bioinformatics resources.

Summary

We have introduced the Brassica Information Portal, a web repository for Brassica phenotype data. This article describes the current state of development and future plans. We describe the back end database schema and the front end web-interface. Also, we describe programmatic and wizard-based submission of Population and Trial data together with fields compulsory for submission to the repository. It was originally intended that this repository meet the needs of the Brassica Research Community in the UK for relevant experimental data. However, it is apparent from discussions within the Multinational Brassica Genome Project that it can now fulfill a broader community infrastructure requirements. We therefore hope that BIP encourages data sharing and re-use by helping to integrate and harmonise datasets generated worldwide.

Software availability

Brassica Information Portal can be accessed from: bip.earlham.ac.uk

Latest source code: https://github.com/TGAC/brassica

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.466050[23]

License: GNU 3.0

Author contributions

Competing interests

Grant information

Supplementary material

Supplementary Figure 1: Examples of .csv templates generated during (A) Trait Scoring Submission and (B) Population Submission. The latter can change substantially depending on the type of data submitted. In this case raw data is being submitted as the template asks for information on the trial design.

Click here to access the data.

Supplementary Table 1: List of fields that are compulsory during the submission of Populations and Trial (Trait Scoring) data.

Click here to access the data.

References

1. Rakow G: **Species Origin and Economic Importance of *Brassica*.** Springer Berlin Heidelberg, Berlin, Heidelberg. 2004; **54**: 3–11.
   Publisher Full Text

2. Labana KS, Gupta ML: **Importance and Origin.** In *Breeding Oilseed Brassicas*. Springer Berlin Heidelberg, Berlin, Heidelberg. 1993; **19**: 1–7.
   Publisher Full Text

3. Brazma A, Parkinson H, Sarkans U, *et al.*: **ArrayExpress--a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res.* 2003; **31**(1): 68–71.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; **30**(1): 207–210.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Leinonen R, Akhtar R, Birney E, *et al.*: **The European Nucleotide Archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): D28–D31.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Leinonen R, Sugawara H, Shumway M: **The Sequence Read Archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): D19–D21.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Cheng F, Liu S, Wu J, *et al.*: **BRAD, the genetics and genomics database for Brassica plants.** *BMC Plant Biol.* 2011; **11**(1): 136.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Usadel B, Schwacke R, Nagel A, *et al.*: **GabiPD - The GABI Primary Database integrates plant proteomic data with gene-centric information.** *Front Plant Sci.* 2012; **3**: 154.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

9. Huala E, Dickerman AW, Garcia-Hernandez M, *et al.*: **The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant.** *Nucleic Acids Res.* 2001; **29**(1): 102–105.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Krishnakumar V, Hanlon MR, Contrino S, *et al.*: **Araport: the Arabidopsis information portal.** *Nucleic Acids Res.* 2015; **43**(Database issue): D1003–9.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. The 1001 Genomes Consortium: **1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*.** *Cell.* 2016; **166**(2): 481–491.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Love CG, Andongabo AE, Wang J, *et al.*: **InterStoreDB: a generic integration resource for genetic and genomic data.** *J Integr Plant Biol.* 2012; **54**(5): 345–355.
    PubMed Abstract | **Publisher Full Text**

13. Bancroft L, Barnes S, Li JY, *et al.*: **Establishment of an Integrated Marker System for Oilseed Rape Breeding (IMSORB).** 2006.
    Publisher Full Text

14. Trick M, *et al.*: **Brassicadb is a database of genetic and molecular information derived from key brassica species.** 2003.
    Reference Source

15. Qiu D, Morgan C, Shi J, *et al.*: **A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content.** *Theor Appl Genet.* **114**(1): 67–80.
    PubMed Abstract | **Publisher Full Text**

16. Shrestha R, Matteis L, Skofic M, *et al.*: **Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice.** *Front Physiol.* 2012; **3**: 326.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

17. The Plant Ontology Consortium: **The Plant Ontology™ Consortium and Plant Ontologies.** *Comp Funct Genomics.* 2002; **3**(2): 137–142.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res.* 1994; **22**(22): 4673–4680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Krajewski P, Chen D, C´wiek H, *et al.*: **Towards recommendations for metadata and data handling in plant phenotyping.** *J Exp Bot.* 2015; **66**(18): 5417–5427.
PubMed Abstract | **Publisher Full Text**

20. Ćwiek-Kupczyńska H, Altmann T, Arend D, *et al.*: **Measures for interoperability of phenotypic data: minimum information requirements and formatting.** *Plant Methods.* 2016; **12**(1): 44.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

21. Davey RP, Thanki AS, Bian X: **Tgac browser: visualisation solutions for big data in the genomic era.** 2016.
Reference Source

22. Tello-Ruiz MK, Stein J, Wei S, *et al.*: **Gramene 2016: comparative plant genomics and pathway resources.** *Nucleic Acids Res.* 2016; **44**(D1): D1133–40.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Gubała T, Szymczyszyn T, Nowakowski P, *et al.*: **TGAC/brassica: v1.0.0.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 1

✔ **Matthew M. Nelson** (iD)

Natural Capital and Plant Health, Royal Botanic Gardens Kew, Ardingly, UK

This manuscript describes a valuable resource for the Brassica research and breeding communities. It is logically and clearly written.

I felt that some of the methodology was inadequately cited. For example, on P4: the terms "PostgreSQL RDBMS technology", "Ruby on Rails", "Nginx" and "Elasticsearch" are used without any external references. Ideally scholarly or technical journal articles would be cited; if not available, then the publisher's name or project website.

Some minor edits required:
P3: The phrase 'assist and help' contains a tautology. Choose one of these terms, not both.
P3: Replace 'steming' with 'stemming'
P6: Presumed typo in the header "0.1 Content navigation"?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Plant breeding, crop genetics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 05 May 2017

✔ **Cyril Pommier** iD

Research Unit in Genomics-Info (URGI) UR1164, French National Institute for Agricultural Research (INRA), Versailles, France

The rational and the objective of the Brassica Information Portal is well explained. Its complementarity with existing data repositories including genomic data repositories and tools and in particular its role as an important repository for Brassica Phenotyping Data is clear. Its relation to other Phenotyping Brassica repositories could be further discussed as perspectives.

BIP makes good use of the current standards for plant phenotyping data, in particular MIAPPE and Cropontology. By being implicated in the current evolutions of MIAPPE, the BIP team demonstrate its will to ensure interoperability of this system with international standards. This is important for long term sustainability.

The data submission process is well described. It shows that significant efforts have been made to ensure data quality and to ease data managers task. Data consistency between population, ontologies and trait datasets should be ensured with the chosen mechanisms. Data citation is well ensured by a DOI generation for datasets.

BIP use modern and proven technological solutions, like Elasticsearch, Nginx web server, rich BIP Web Service API backing the whole user interface and allowing user or programatic interaction with the system.

For the technical description of the software tool to be fully sound, the following points should be improved.
- The role of BIP in the BrAPI should be further described. In the Author Contributions, BrAPI is cited but it doesn't appears elsewhere in the text.
- Some technical description could be improved, for instance "universal search" isn't very clear.

This paper shows very well the development process of the BIP information system. The current status of the system is clear. Its perspectives could be further discussed regarding FAIR compatibility, Open Data including data licences policies (CC-by). Alongside this point, BIP team curent opinion on distributed phenotyping information system, backed by the BrAPI, could be discussed.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: Collaboration in the Elixir Plant use case.

Referee Expertise: Distributed information systems, computer science, phenotyping data managment

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 03 May 2017

Christopher J. Rawlings (iD)

Department of Computational and Analytical Sciences, Rothamsted Research, Harpenden, UK

The background to the development of the Brassica Information Portal and its origins in the CropStore system are well described. This paper is an update on the work in progress to deliver an integrated resource linking phenotype to genotype data for Brassica species.

There are welcome new developments in the BIP including greater use of ontologies, in particular links with the MIAPPE initiative to capture phenotype in a systematic way and to adopt emerging metadata standards for crop phenotyping projects.

A summary of the contents demonstrates that the BIP already comprises a significant scale repository for Brassica pre-breeding data and where relevant this is linked to SNP data visible with a genome browser.

This paper is about the BIP as an information resource, rather than a publication of a new method. The paper provides a high level view of the technical implementation, reporting on adoption of the PostgreSQL relational database technology and the general use of Restful services as the basis for the data management operations in the BIP, including those that support improved interfaces for data loading and for browsing. A conceptual database schema is presented for the BIP database.

There are other areas of BIP that are adopting good practise in science data management, including the use of Orcid identifiers for user authentication and data search and retrieval using Elasticsearch technology.

The paper provides overviews of key data management tasks and gives more information in the

Supplementary material provided with the manuscript. There are also links to the Open Source repository Github where the software for BIP is available and to the example code for querying the data programmatically.

Overall, this is an informative paper about the current state of the Brassica Information Portal. It is not a primary research paper, but rather a report of work in progress supporting the Brassica research community with a useful resource and software that could be re-used for other crops where there is a requirement for the management and dissemination of crop genotype-phenotype datasets.

While there has been good progress so far towards developing the software so that the system supports open access to the software and data, it remains unclear to what extent the FAIR principles of data stewardship are being embedded in the BIP software. (e.g. Wilkinson et al[1]). In particular, it would be helpful for the authors to describe their plans for exposing discoverability metadata and how they view the role of BIP as a "FAIRport" for Brassica data.

References
1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B: The FAIR Guiding Principles for scientific data management and stewardship.*Sci Data*. 2016; **3**: 160018 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.