Proposed **Roadmap** for a Multinational **Brassica Information System**

Phase I **Inventory**
Phase II **Collection**, triage and **prioiritisation**
Phase III Detailed **curation,** compliance and integration

on behalf of the ***Multinational Brassica Genome Project*** (MBGP)
Graham King (Chair, steering committee 2017+18); v 1.0 Nov 2017; v 2.0 June 2018

Background:

1. Based on discussions within MBGP over the past couple of years, there is an **ongoing need to collate disparate sources of data** and related meta-data, in order to maximise return on past and future R&D investments.
    - The **MBGP has been successful** as an international collective in sharing experimental resources and achieving collective goals.
        - include agreement on **standards** for chromosome and gene nomenclature
        - development of gene-chip, SNP-chip and related **platforms**
        - major **milestones** of reference and more recently emerging pan-genomes.
    - It has been agreed in principle by MBGP to establish a Brassica Information System. Here I propose a **roadmap** for how to achieve this. **Feedback** on approach and logistics is very **welcome**.

2. **Phase I** (Inventory) is key to our future collective efforts, particularly in providing a platform for new researchers and students entering the brassica universe.
    - Within the framework of a distributed curation effort and existing repositories, this process will be facilitated by agreeing and adopting a **wider set of standards** to describe entities and datasets.
    - This is very timely given emergence in the past few years of various FAIR data standards.
      FAIR = **F**indable, **A**ccessible, **I**nteroperable, **R**e-usable.
        - This includes initiatives such as MIAPPE (for plant phenotypic experiments), BraTO (brassica trait ontologies and trait dictionary). With the group at Earlham in UK led by Wiktor Jurkowski, we have been working directly with MIAPPE consortium to ensure that it works with brassica as a use-case.

3. A key issue is **persistence** of datasets.
    - A number of existing repositories are fit for purpose, particularly for generic data (eg Short Read Archive, Ensembl, Brassica Information Portal - BIP).
    - Others may need to be established, initially to hold datasets in a data '**warehouse'** to ensure long-term access. This may include datasets currently published as 'supplementary data' or at least inventory to such datasets that have a persistent DOI.

4. MBGP requires a **sub-group** to oversee the **dataset inventory**, and assign responsibility for different data classes (see Table below).
    - Within this group it would be helpful to have **individuals** volunteer as a **point of contact** to lead specific areas.
    - For reference, versioned Inventory Lists can be maintained at www.brassica.info

5. BRAD and BolBase managed by CAAS (Xiaowu Wang group et al.) contain valuable reference datasets and links, as does Genoscope (B. napus) managed in France.

6. EBI (European Bioinformatics Institute) confirmed in 2017 they will host and manage ENA/NCBI-registered annotated genomes at **Ensembl Plants** in a consistent format, processing through their Compara multi-species db and associated pipelines any reference *Brassica* genome that is agreed by the relevant community. They recognise MBGP as representing the Brassica research community.
    - There are several pre-processing steps we can coordinate to make this a validated process. GK will contact the relevant lead research groups to establish how this can be achieved.

- Within Australia both UWA (Dave Edwards, Jacqui Batley, Philipp Bayer) and SCU (Graham King, Abdul Baten) are nodes of EMBL-ABR which has strong links to EBI, including for training.

7. The UK BBSRC invested in establishing the **Brassica Information Portal** at Earlham Institute (Norwich). This has played an important role in establishing a pipeline for **user submission** and navigation of a range of Brassica datasets, with emphasis on trials, phenotypic traits etc.
8. **Funding** of course remains an **ongoing** issue for data curation/integration, so it is important that we have a resilient, interoperable and scalable approach.

Phase I Establish MBGP Inventory

Principles:

- A MBGP **data entity** will have a unique identifier, conforming to an agreed nomenclature standard
- A MBGP **dataset** may consist of one or more **DATA CLASS** (Table 1)
- MBGP should agree / adopt data standards where possible
  - Opportunity to assign a **DOI** to a dataset (via DivSeek/FAO) – included in MIAPPE etc
    - This seems a key step if we are to make a reliable inventory
  - Wherever possible assign to **ORCID** for persons and/or institutions involved
  - Establish look-up MBGP **registry** of distinct entities to reduce ambiguity
    - eg so that consistent unique identifiers for eg biosamples, maps, genome versions, SNP markers etc etc.
    - There will be different ways of achieving this:
      - Biomart may be suitable (https://www.ensembl.org/biomart);
      - *other suggestions welcome*

Proposed Workflow

- MBGP **sub-group for data management** established (does this include you?)
  - Determine means of managing DOIs (Divseek have agreement with FAO to allocate sets of these)
  - A group met at PAG in Jan 2018 – see http://www.brassica.info/info/reference/minutes/MBGP_Minutes_Jan2018.pdf

This group ***recommended***:

- Adopt approach taken by the Wheat Information System (**Wheat IS**), which uses distributed databases, support individual efforts, and is linked by a common search tool.
- A community-wide call for datasets to build **MBGP dataset Inventory**
- **Prioritise datasets to enable:**
  - Centralised indices for existing databases and flat files, likely building on the Apache Solr system used by Wheat IS
- Wider development and adoption of **standards** for brassica-related experimental entities (biosamples, trials, traits, etc)
- Addressing issues of synonyms/homonyms – scope for allocating **DOIs**. Moving towards establishing look-up **registries** and allocation of community-wide unique identifiers
- Accept the reality of **pan-genomes** and develop a means of managing **gene-name** referencing
- Distribute a community-wide **consultation questionnaire** via brassica.info mailing list

- Identify a MBGP **lead contact** identified for each Data Class
- Convene Data Class group to assist in developing Standards and Inventory
    - Propose and publish standard **nomenclature** on www.brassica.info
- Allocation of dataset **DOIs** in a **registry** and controlled vocabulary identifiers
- Build **registry** for each MBGP data **entity** so that identifiers are unique and searchable
- Allocate/download datasets into **data-warehouse**, accessible from www.brassica.info

DATA CLASS	RELEVANT STANDARDS	CURRENT REPOSITORIES	COMMENTS	MBGP LEAD CONTACT
Experiment/evaluation • Project • Trial • Occasion • People/institutions involved	MIAPPE " " " " ORCID	BIP for some		
BioSample Plant_populations/collections, consisting of: • Lines • Accessions • Scoring_units	BioSample (& BioProject) at NCBI https://www.ncbi.nlm.nih.gov/biosample/ at EBI https://www.ebi.ac.uk/biosamples/ and MBGP examples http://www.brassica.info/resources.php	BIP for some	○ Reference to registered cultivars ○ Management of pedigree data	
Genome sequence • ***MBGP validated reference*** genome, anchored and oriented ○ Meet nomenclature standards ○ Evaluated for completeness/accuracy • Pan-genome • Draft genome • Resequencing • Methylome • Exome • Repeats	MBGP gene model nomenclature http://www.brassica.info/resources.php	• Ensembl • (BrAD)/BrGDB • GenoScope	○ Genome data in an agreed format of GFF3 (i.e. include same categories of information for different brassica genomes – currently this is not the case) ○ Can tools for MBGP be placed in public domain (eg. incl. tools that convert GFF3 to standard MBGP nomenclature) ?	
Transcriptome • RNA-seq • Affymetrix GeneChip™ data? • Bead array	?? MIAME	SRA , ENA		
SNP/Marker • SNP-Chip data • GWAS, GBS etc – link to registered biosample (plant				

accession etc), trait_descriptor etc (see below) • other				
Phenotypic trait • Defined descriptors or reference to established/published methods • Trait scores (raw and/or mean) associated with specific biosample/experiment above)	MIAPPE, BraTO Plant Ontology & MBGP examples http://www.brassica.info/resources.php	BIP Others?		
Genetic Map • SNP-anchored linkage maps and associated populations ○ Different marker platforms • Integrated maps • Legacy maps • QTL	MBGP http://www.brassica.info/resources.php	some in BIP		
Other				