IAIC Standards and Ontologies (SO) Working Group (WG)
WG Chair: Pankaj Jaiswal, Oregon State University

Contributors:

Jelena Brkljacic (ABRC), Sean Walsh (EU Agron-omics[1] and TiMet), Sean May (NASC[2]), Eva Huala (TAIR[3] and Phenotype RCN[4]), Lukas Mueller (SGN[5]), Henning Hermjakob (IntAct[6]), Ruth Bastow (GARNET[7]), Ethalinda (Ethy) Cannon (Popcorn[8] and MaizeGDB[9]), Hans-Michael Muller (TextPresso[10]), Cathy Wu (BioCreative[11] and Protein Ontology[12]), Hong Cui (Flora of North America[13]), Pankaj Jaiswal (Plant Ontology[14]) and Gramene[15])

The Group started by identifying the major data set(s) or piece(s) of the Arabidopsis research information that we may work/deal with on a day to day basis. We also identified and concur with the resources and data set and the kind of informatics coordination required so that we can develop standards (if deficient) and tools on tagging and building connections with each other's relevant data like pieces of a puzzle. Some of the primary data sets that we think of are available and would be made available to the community via publications and online resources as follows. In order to integrate them in an AIP warehouse or create a network of data sets for useful integration and answering biological questions, these data sets and resources need standards and ontologies to describe data properties, exchange format and metadata.

List of data sets:

-germplasm
-phenotype (mutant and QTL)
-gene function
-polymorphism and GWAS scale of data
-resequencing genome
-interactomes
  -physical
  -co-expressed

[1] http://www.agron-omics.eu/
[2] http://arabidopsis.info/
[3] http://www.arabidopsis.org/index.jsp
[4] http://www.phenotypercn.org/
[5] http://solgenomics.net/
[6] www.**ebi**.ac.uk/**intact/**
[7] http://www.garnetcommunity.org.uk/
[8] http://popcorn.maizegdb.org/main/index.php
[9] http://www.maizegdb.org/
[10] www.**textpresso**.org/
[11] http://www.biocreative.org/
[12] http://pir.georgetown.edu/pro/
[13] http://fna.huh.harvard.edu/
[14] http://plantontology.org
[15] http://www.gramene.org

-others
-networks
-metabolic
-signaling
-regulatory
-proteomics
-transcriptomics
-metabolomics
-sequence
-sequence features
-environment/treatments
-assays/experiments/techniques
-literature
-IDs
-images and image analysis
-samples libraries
-genetic markers
-genetic maps
-taxonomy
-phylogenomics (genus, species and gene level)
-text mining
-Others (please add)

Working Group:
Introduced themselves by highlighting the activities they work with and quickly moved into discussing the following two topics:

#1 The role of 'germplasm/genotype' information that is critical for putting together the context of a study and making connections between phenotype, function, interactomes, proteome, genome, genes, genotype/polymorphism, markers, QTL, expression, etc.

Jelena (ABRC) and Sean May (NASC) mentioned that they have information on lots of germplasms for which they hold seeds in the stock center and some virtual (digital records only) stocks reported in the literature. Jelena mentioned ~4000 virtual stocks in ABRC and Sean covered that NASC holds about 800,000 accessions.

The Arabidopsis Biological Resource Center (**ABRC**)[16] is currently funded till 2017 with emphasis on collection, curation and maintenance. There is virtually no or very limited informatics support. TAIR supports the seed stock ordering process. Curation is primarily done by ABRC by following their internal guidelines.

[16] http://abrc.osu.edu/

Both ABRC and the Nottingham Arabidopsis Stock Centre (**NASC**)[17] have information such as phenotypes written in free text format. NASC made an attempt at a combinatorial approach of 'Entitity [inherits] Quality' where they used Plant Ontology[18] as an *entity* and *quality* from Phenotype and Attribute Ontology (PATO)[19] to describe phenotyped accessions. This was attempted on a small experimental scale and needs expansion with carefully crafted work flows, curation forms and by potential use of the Natural Language Processing (NLP) tools like Knowtator[20] ,TextPresso[21] and others.

Most of the time, stock/accession is connected to the AGI gene locus which also holds information on ecotype, geo references (collection site), owner, etc.

We discussed that besides in-house annotation efforts by NASC and ABRC, the research community must be encouraged to submit the germplasm seed and the meta-data associated with the germplasm. Leaving aside the past efforts we think an experimental scale effort is required to develop a COMMON portal or user interface for the following. *We don't want the community to get trained in different user interfaces.*

- searching different stock centers (global access of all national and international) using the APIs
- submitting the stocks (seeds and meta-data) and integrating the data in the respective source center closest to the researcher's home.
- use standard descriptors for phenotype (may need development of phenotype ontology), geo reference; cross-referencing each other's stocks to manage links to parents and descendants.
- connecting to genome, genes, alleles, etc.

* May need some NLP measures to do a first pass categorization and annotations.

* Sean May raised a question about the stock centers **other than ABRC and NASC.** Would they be interested in working with us on common UI and standards? **This needs to be discussed.**

Sean Walsh mentioned that they are recording phenotypes and traits in EU Agron-omics project using the Knowtator tool[22] to capture information from publications using ontologies to generate knowledge networks that are used to analyze HT data. Also have phenotype data from a developmental series (Granier Lab) and from a re-screening of SALK lines (Micol Lab). Integrating this data with other phenotype data sources would be possible.

Lukas gave an example of SGN (tomato). They have TGRC[23] similar to ABRC as their stock. The tomato stocks (~20,000) are getting annotated with loci, phenotype info (using GRIN[24] trait descriptors and Phenotype ontology developed by SGN*), associated images, track submitters/owners.

* The SGN phenotype ontology has cross-references to the generic/reference Trait Ontology (TO) maintained by Gramene and the Plant Ontology (PO).

[17] http://www.nasc.us/
[18] http://plantontology.org/
[19] http://obofoundry.org/wiki/index.php/PATO:About
[20] http://knowtator.sourceforge.net/
[21] www.textpresso.org/
[22] http://knowtator.sourceforge.net/
[23] http://tgrc.ucdavis.edu/
[24] http://www.ars-grin.gov/npgs/index.html

One suggestion was to either work with the TO developers to make sure there is Arabidopsis representation or make a similar effort like SGN's on developing an Arabidopsis phenotype ontology as a quick start.

There is also a need to build semantic model of genome features to functional entities (slide #21) to properly connect phenotype with genotype at the specific functional entity level. The protein functional entities (including splice isoforms, post-translational modifications, and complexes) can be represented in the ontological structures of the Protein Ontology (PRO)[25] and connect with the Gene Ontology (GO)[26] and Plant Ontology (PO)[27] for functional and phenotypic annotations.

#2 Common IDs and mapping to data objects

This issue came up in the Use Cases WG. May need more discussion on how to tackle it and if the community agrees with such standards. There are several IDs assigned to similar and/diverse datasets such as Arabidopsis Genome Initiative AGI IDs, UniProt, Unigenes, Array probes, ESTs, mutants, gene models, RefSeq, etc.

Other potential objects that needs creation/maintenance in the form of data is a catalog of genes. As we understand it, the AGIs are assigned to gene loci in the Columbia genome. There are 1000s of new genomes getting sequenced and we are going to find that many gene loci from these genomes do not exist in Columbia reference genome. There is no place holder for such records.

Note: In Gramene (mentioned by Pankaj) the solution was to associate genes to species and not a genotype's (e.g. Columbia-0) genome. Genes represented in the Columbia genome would thus be an allelic form, same as in other ecotype genomes. If the locus is missing in Columbia (it's a 'null allele') and if present uniquely in another ecotype (it's a 'new allele'). This means for every allele (from any genotype) there needs to be a reference gene tied to the species. However, clarification is needed for how we describe the allele. Genome Diversity projects call alleles at the SNP level and here we are dealing with alleles at the gene level.

- Despite the above new gene ID requirements we need standards on mapping entities, their relationships on derivation and assigning IDs
- Develop a universal ID look-up service portal. Make extensive use of APIs and RESTful webservices to drive the service.
- Ethy Cannon mentioned about the Popcorn portal she has developed for tracking all the data sources for corn. They're inviting maize resources to contribute the connections to various data objects in MaizeGDB vs those external to it. Kind of a look-up service for related data objects. It uses hard coded workflow for capturing the connectivity, but can be improvised for more dynamic workflow and use of APIs.

[25] Natale et al Nucleic Acids Res. 2011 Jan;39(Database issue):D539-45. Epub 2010 Oct 8.
[26] www.geneontology.org
[27] www.plantontology.org

- As more number of Arabidopsis genomes are getting sequenced, pretty soon we will end up discovering lots of new and novel genes. Therefore, the question came up on 'How to resolve the gene nomenclature artifacts'. Currently TAIR curators moderate this process. As we move into more of a community adopted projects, we suggest formation of an International Arabidopsis Gene Nomenclature Committee. This may be a sub-committee of the MASC and would be responsible for gene nomenclature assignments and conflict resolution. Develop online gene registration and conflict checking forms and guidelines for mandatory submissions

#3 Interactome Datasets:

Henning (IntAct[28] at EBI): The IntAct database (http://www.ebi.ac.uk/intact) has worked with TAIR on taking the lead role on collection and curation of the primary gene-gene interaction data. The data are curated at the gene level. If the community thinks we need to curate and develop a resource where the actual interactions are tied to the genotype (the important factor including the environment), respective gene models, allelic forms, and the context (tissue, growth stage, treatment, etc)_ under which the interaction appears, we need to work with IntAct managers or develop a workflow and community annotation portals using the standardized annotations. The new submissions can be encouraged by developing standardized supplemental files as a prerequisite for journal publications. As of now from IntAct the resources available are as follows. Their data structure follows a widely accepted data standard.

- The IntAct editor is a web-based curation environment for molecular interactions. It is currently used not only by the 3 strong IntAct curation team, but also by 5 UniProt curators, and one/two curators each of InnateDB[29] (Toronto/Dublin), I2D (Toronto), Molecular Connections (Hyderabad). All data are tagged with the curation teams' identity and made freely available for download and interactive access.
- PSICQUIC is a distributed query system for molecular interactions, currently implemented by 21 different interaction data sources. If you enter
- IntAct recently did a large scale curation project for Arabidopsis[30]. Curation of Arabidopsis interaction data in IntAct by IntAct curators and domain experts, and then integration of the data into an Arabidopsis portal/aggregator like the MASCP Gator might be an interesting option for this data type.

#4 Proteomics: (from Henning at EBI)

The PRIDE database[31] provides a public repository for proteomics data. Web services and in particular a PRIDE DAS track are already available and might be useful in a distributed Arabidopsis data source. MORE TO BE DISCUSSED

[28] *www.**ebi**.ac.uk/**intact**/*
[29] http://www.innatedb.ca/
[30] Lee et al. Plant Cell. 2010 Apr;22(4):997-1005. Epub 2010 Apr 6.
[31] *www.**ebi**.ac.uk/**pride**/*

#5 API development

There are numerous project and online resources that have started and/or interested in opening up the access to their datasets via APIs. We find that this is an excellent service provided by such resources, however, due to lack of community standards on nomenclature of objects and annotations which these APIs are serving, the end-point source would have to write data transformation/translations scripts to capture and map the entities. Though many of the services provide nomenclature descriptions, it will be wise to start a pilot standardization and nomenclature project on small set of entities, like, genes, alleles, gene products, functional annotations suing ontologies, germplasm, gene-gene interactions.

#6 Natural language processing and text mining

The scientific literature represents the repository of knowledge. Currently database information is insufficient for knowledge discovery. Key knowledge gaps must be curated from the literature. Natural Language Processing (NLP) tools and text mining systems may be used in the biocuration workflow, from triage (document prioritization for curation) to information extraction (identification of entities and functional properties). In addition to broadly adopted NLP tools such as Knowtator[32] for text annotation and TextPresso[33] for information extraction, the BioCreative is a community-wide effort for evaluating text mining tools for biology. Recent Biocreative efforts focusing on evaluating and promoting the development of interactive systems[34] may result in text mining tools with enhanced utility and usability for biocuration and capability for active learning.

[32] http://knowtator.sourceforge.net/
[33] www.**textpresso**.org/
[34] Arighi et al BMC Bioinformatics. 2011 Oct 3;12 Suppl 8:S4.