transPLANT milestone report

MS8 (work package 3): Standards in use internally and externally by transPLANT Consortium Members

This report synthetizes the work done in WP3 for implementation of standards developed in the project related to phenotypic data and metadata.

In 2013 the project delivered a set of recommendations concerning the content, annotation and format of data sets containing information from phenotypic experiments. The recommendations for content were described in the Minimum Information about Plant Phenotypic Experiment (MIAPPE) document, the recommended ontologies in Deliverable 3.1, and the recommended format was described as ISA-TAB for phenotyping configuration; the current information on these deliverables is available at a dedicated webpage, and is also available through the Biosharing portal as http://biosharing.org/bsg-000543 . The steps that followed in 2013/2014 concerned implementation of the standards in practice at partner's institutions. Implementation took place mainly at four institutions: IPG PAS, INRA, IPK and GMI.

IPG PAS

BII-Manager, a java application, validates ISA-Tab formatted data sets and stores information to database backend. BII Web application is the database front-end accessible via an Internet browser. Both applications are part of the ISA Software Suite published at http://www.isa-tools.org/tools.html. We are using this software to develop an instance of a database of phenotypic data compatible with MIAPPE and with the ISA-TAB file format. The applications were first installed on a desktop computer to test them locally. Currently installation on the *cropnet* server at IPG PAS is in progress. In addition to this, standards are promoted for data exchange and collection in national projects. In the project GENSEC devoted to development of molecular markers of disease resistance in rye ISA-TAB format is used to collect experimental data on resistance traits observed in the field and metabolomic traits observed by LC-MS protocols. In the project POLAPGEN (www.polapgen.pl) annotation of quantitative traits observed in the field, greenhouse and laboratory experiments was done according to the schema proposed by the Crop Ontology group and published as a table of trait description at http://www.cropontology.org. Integration of this table with the set of trait descriptions submitted to Crop Ontology as "ICARDA Trait Dictionary" is in progress.

INRA

Ephesis, or Environment and Phenotype Information System, is an INRA URGI platform project. It has been initiated by the Departement Genetic et amelioration des Plantes (DGAP), which has ensured its long term stability by assigning this mission to a permanent engineer. It provides a national database storing Genotype by Environment experiment results, developed within the URGI information system, GnpIS. Ephesis web interface offers the user with the possibility to create multi-trial dataset suitable for various analyses (meta-analysis, GWAS, etc.). For instance, all the experimental data for a wheat accession in year 2000 can be selected. The purpose of the action now is to provide the possibility of exporting search results as a single ISA-TAB data set (archive). In the import the "investigation" represents the whole search results. There is one "study" by trial. The study contains

only the subset of data corresponding to the user query. This export capability is now being evaluated and tested to gather more feedback on the format.

IPK

In order to publish research data from phenotype experiments, IPK make use of an in-house Laboratory Information Management System (LIMS). It documents experiments and unifies the data processing (Arend et al., Data Management Experiences and Best Practices from the Perspective of a Plant Research Institute, Data Integration in Life Sciences (DILS), 2014). Main components of the system are experiment planning, sample descriptions, measurement parameters and platform configurations that are linked to the data collected and to the measurement store. In collaboration with the German Phenotyping Network (DPPN) experiences and recommendations from transPLANT project where used to develop methods for homogeneous data exchange and coordinated with German DPPN nodes "German Research Center for Environmental Health" (HMGU) Munich and "Research Center Jülich GmbH". The decision was to apply ISA-TAB standard to publish metadata for phenotype experiments and to link raw and analyzed data. As first step, it was necessary to compile an ISA-TAB reference experiment of multiple data domains. We combined data sets from metabolite experiment, image analysis and manual, phenotypic measurements. All semantic and technical documentations, measured parameters, protocols and references to ontologies were converted into ISA-TAB manually. The result was validated by the ISA-TAB consortium and will be released in the third quarter 2014 as data publication with the DOI 10.5447/IPK/2014/4. Another focus was the establishment of a repository of research data that include ISA-TAB formatted data as well. As the basis of this the e!DAL infrastructure for data sharing and publication (Arend, Lange et al. e!DAL - A framework to store, share and publish research data BMC Bioinformatics 2014.) was used. All data published so far are listed in DataCite: http://alturl.com/p5co5. Like IPK, HMGU is contracted as DataCite data center. Thus HMGU will operate another repository of phenotype experiments. In addition, e!DAL was published as open source system (http://edal.ipk-gatersleben.de) and could be used as technology platform to build further research data repositories.

GMI

At GMI the import as well as export of ISA-TAB archive files in the GWAS web-application was implemented (http://gwas.gmi.oeaw.ac.at). The configuration files for phenotyping investigations developed in WP3 worked well. However the data model in GWAS application is much simpler than what ISA-TAB archive is proposing. As a result there were some issues in proper mapping the ISA-TAB entities to the ones in GWAS DB:

- The investigation configuration (investigation.xml) file is the easiest one to map. It maps to the Study entity in the database.

- In the database a Study can have many observational units (i.e. genotypes/plants) and each of these observational units can have a phenotype value of a specific statistical type (count, mean, measure) that is grouped together to a phenotype entity. In other words, each Study instance can have multiple phenotypes. This is not easy to map to ISA-TAB terms. So it was decided to make it as simple as possible. Each Study in our DB maps to a single Investigation file with a single studySample file and a single assay file.

- In the studysample file (i.e. s_study.txt) we fill out the Characteristics[Organism] column (A.thalina) and the Characteristics[Infra-specific name] (accession id) .

- The a_assay.txt file basically has a 1:1 mapping to the studysample file and contains just hardcoded links to a t_trait_def.txt and a d_dervived_data.txt file.

- The t_trait_def.txt file contains the list of phenotypes and the d_derived_data.txt file contains the actual phenotype values in matrix form.

- For most of the fields and columns in ISA-TAB there are no corresponding values in the database (factors, treatment, organism part, etc) so they are ignored.

We tried to use as much functionality of the ISAcreator (https://github.com/ISA-tools/ISAcreator) library from ISA-TOOLS as possible in order to avoid re-inventing the wheel. However we ran into several serious issues. First of all ISAcreator is a full blown Swing application that has loads of dependencies. ISA-TOOLs don't provide any Java library for export and import of ISA-TAB archives. So in order to re-use the functionality, I have to include the entire ISAcreator.jar library as dependency to our Java Spring web-app. Unfortunately ISAcreator itself has some dependencies which conflict with our dependencies (JPA, hibernate, etc). I had to create a separate jar file that doesn't bundle the dependencies and exclude them from being automatically imported respectively (see issue https://github.com/ISA-tools/ISAcreator/issues/278).

Once this was done, I started to implement the export functionality because this one seemed to be easier. Unfortunately the ISAcreator behaves differently when exporting from the GUI compared to when exporting programmatically (see issue: https://github.com/ISA-tools/ISAcreator/issues/279). Furthermore there is no support for exporting d_derived_data.txt and t_trait_def.txt files. I had to write the code for that myself. Also there is no built-in method to create a zip archive.

After exporting was finished I started to implement the import functionality. Again there was no built-in method to parse the d_derived_data.txt file or the t_trait_def.txt file, so I had to build them myself. After parsing the ISA-TAB-archive I convert the ISA-TAB entities to our internal database entities and send the result back to the user and notify the user about any errors.

Summary

The efforts to implement ISA-TAB format in practice are generally successful. The documentation of the ISA-TAB format is fairly good. Some concepts are confusing and more complex than some data requirements but basically it covers any possible scenario. The ISAcreator GUI was tested and it works fine. However, the isa-tools libraries are quite buggy, not well documented, and they have a lot of dependencies that make it difficult to integrate into our own software. This is definitely an area that will need to be improved if this toolset is to be adopted on a wider basis, although it remains, in our opinion, the most promising candidate to support the data types are working with.