Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

D3.1
Recommended ontology set for use in phenotype description and epigenetic variability

Due date of deliverable: (M24, INRA): 31.8.2013 (M24)
Actual submission date: 2.9.2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: INRA

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor
INRA URGI **IPG PAS**

Introduction
Deliverable reference number:D3.1 Ontologies are sets of strictly defined and semantically related terms used to describe concepts of a domain or a reality. The complexity of ontologies is very variable and goes from very simple controlled vocabulary to very complex systems based on directed acyclic graphs (DAG) with semantically typed edges. It is therefore very important to choose the simplest possible model for the purpose of the ontology. Indeed annotating data with ontologies can serve for data quality control, for helping in data exchange and interoperability between information systems possibly up to open linked data or for knowledge discovery through advanced data mining. We have focused on the first level of ontology purpose and complexity: ontologies for data quality. This means ensuring that everybody calls a phenotype the same way and not plant height in one laboratory, height in a second one and top to first root distance in a third one. Two reasons lead us to this choice for simplicity: a complex ontology can become too big to be easily used for data annotation and most important ontology building relies on experts input, mainly biologist input, and requires a lot of time and effort to reach a consensus and therefore need dedicated persons. We must also distinguish Traits and Phenotypes. A Trait is a definition of an observable characteristic, like "Growth habit", "shootless embryo" or Yield. A Phenotype is the association of a Trait and a value, for instance "yield" = 1.5 tons by hect, "Growth habit"= spreading. Note that certain Traits, like "shootless embryo", are self-sufficient for data annotation and don't need values. Several categories are necessary for plant phenotyping ontologies. We need to describe the plant, list the traits and, since phenotypes are plants response to environment, we also need environment ontologies.

Methods
This ontology building has relied on experts inputs gathered through existing partnership or several seminars and workshops among which: • PhenotypeRCN Annual Meeting 2011 (previous Transplant but related to the thematic) • Plant Ontologies for Agronomic Traits Workshop 8-9th December 2011 (Transplant meeting) • Crop Plant Trait Ontology Workshop, Sept 13th-15th, 2012 at Oregon State University The network built through those events allowed to get input from ontologist and biologist to build the current deliverable. Those experts include people like Laurel Cooper and Pankaj Jaiswal, in charge of the Plant Ontology and Gramene Trait Ontology coordination, Elizabeth Arnaud in charge the Crop Ontology. We also gathered input from biologists working in the Hordeum-related project POLAPGEN-BD (www.polapgen.eu, through Pawel Krajewski, IPG PAS, coordinating that project and also transPLANT WP3), Jacques Legouis (INRA) involved in wheat national and international projects and partners of the Ephesis INRA project who have been involved in ontologies building reflexions since 2008. This bottom up approach should help the spreading and acceptance of those recommendations.

Results (if applicable, interactions with other workpackages)
Plant description. The Plant Ontology, supported by the Plant Ontology Consortium, exists since 2004 and is actively maintained and curated by a team of dedicated persons. This ontology is cross species and has been built historically for maize, tomato, rice and Arabidopsis. An elaborated system has been set up to ease submission of new or

modified terms to the ontology, therefore allowing the international community to be part of the project. This ontology allows to anatomically describe a plant and therefore correctly annotate the observed element of a plant. Furthermore, the Plant Ontology proposes a development stage section.

Therefore the recommendation for plant description is to use the Plant Ontology.

Trait ontologies

There are three levels of plant trait ontologies. The first one is the local controlled vocabulary which includes unit and measurement protocols and is most of the time phenotyping platform or project specific. The second level includes applications ontologies which are specialized on a domain or a species and also includes unit and protocol. At the third level, Reference Trait ontologies are cross species and don't include unit or methods.
The Crop Ontology (www.cropontology.org) is a repository of application ontologies, mainly species specific trait ontologies plus some domain specific ontologies like the Crop Research ontology. Some of those ontologies are progressively being linked to reference trait ontology, the Gramene trait ontology. This linking can be achieved through a simple cross reference, but it is better to have a rich semantic linking between ontology terms. For instance and following the Entity Quality Value model [Integrating phenotype ontologies across multiple species, Mungall et al. Genome Biology 2010, 11:R2], for the trait "leaf area", we would add a link to the Plant Ontology term "leaf" and a link to the Phenotypic quality ontology (PATO) term "area".
For the biologists we are working with, unit and methodologies are important, and the kinds of traits they are working with are more present in Crop Ontology than in reference trait ontology. Furthermore, the application ontology approach allow them to focus on their species or favourite domain, thus allowing to easily reach a consensus on terms definitions with a small international community which most of the time already shares the same concepts. On the other hand, the Gramene trait Ontologies seemed too big for everyday use and too big for quick curation.
The Crop Ontology has been used for the submission of several ontologies. Vitis ontology is a whole new ontology maintained by Eric Duchêne, INRA. Two contribution to existing ontologies are also in progress for Hordeum (POLAPGEN Consortium) and Triticum (J. Legouis for INRA). In each case, the submission process goes through the validation by the species specific ontology curator. This process seems to be sufficient for the current size of the ontologies. While there is a dedicated coordinator for the whole Crop Ontology, Elizabeth Arnaud, each ontology is maintained by a dedicated curator who is, and must be, an expert of the domain or the species. Therefore, this curation process might be slow in the future, but more because of the unavailability of the curators than because of difficulties to reach a consensus on terms.

Therefore the recommendation for Trait Ontologies is to use the Crop Ontology. Additions to this ontology should include EQV semantic linking using the Plant Ontology and PATO when possible.

Chemical Phenotyping

We also need references for Chemical analysis of plant samples. For this, the Chemical Ontology, Chebi is a known reference. But according to metabolic phenotyping experts, CheBi isn't sufficient, and two databases are of common use for referencing chemical composition of plant samples. Those are the Brenda Enzyme Information System (http://www.brenda-enzymes.org/) for enzymes and the Golm Metabolome Database (http://gmd.mpimp-golm.mpg.de/) for metabolites. They are commonly used by Metabolomics platforms like INRA's High Throughput Metabolic Phenotyping Platform (Y. Gibon, INRA Bordeaux). There drawback is that they are not formalised as ontologies.

Therefore the recommendation for Chemical Phenotyping is to rather to use Brenda and the Golm Metabolom Database, and possibly CheBi if necessary.

Environment Ontologies

Environment ontologies are currently less developed. There are two ontologies currently, EO and ENVO.

ENVO is a very general ontology maintained by known experts of bioontologies (S. Lewis, C. Mungall, M. Ashburner, B. Smith, …). ENVO contains terms for biomes, environmental features, and environmental material. The drawback of this general approach is that only a very small subset of the ontology might be useful for plant phenotyping. Furthermore, Phenotyping community will likely focus on dedicated plant environment ontology rather than a very generalist ontology like ENVO. The EO is maintained by Gramene and is plant specific. It is closest to the concern of plant phenotyping experiments but it still lacks a lot of terms.
It seems that currently no environment ontology really address the biologists needs. A consequent effort has to be made to reach the quality level of Trait and Plant ontologies. Dedicated projects are being set up to find the necessary resources. In particular, the PHOEBE project submitted within the 2013 ERA-CAPS Call addresses the need for environmental ontologies.

Therefore there is no clear recommendation for environmental ontologies for now. EO is an interesting draft, but EO evolutions or future ontologies are likely to be better suited.

Experimental Design and Investigation
Having a common ontology for describing Experimental design is a plus. This has been done in coordination with the D3.2 Format specifications for data exchange by flat file and web services. There are some promising ontologies for this, among which Ontology for Biomedical Investigations (OBI) and the Crop Research ontology included in the Crop Ontology. The Plant Ontology Driven Database: Australian Plant Phenomics Facility has also set up a very rich ontology for describing phenotyping experiments. Dedicated projects have also dedicated resources to try a unification of those ontologies.

Therefore the recommendation for Experimental Design and Investigation is to use OBI or the Crop Research ontology, but keeping in mind that major refactors of the ontologies might be on the way.

Epigenetic Ontologies
According to the Neuroscience Information Framework (NIF) Standard Ontology, the term "epigenetics" means "Changes in gene expression caused by mechanisms other than changes in the underlying DNA sequence. These changes may remain through cell divisions for the remainder of the cell's life and may also last for multiple generations. (Adapted from wikipedia.org/wiki/Epigenetics)". In agreement with this definition, epigenetics uses a vocabulary, which is to a great extent common with genetics and genomics. However, the recent fast developments in this area have led to creation of new specific terms. Some of them have analogues in the older vocabularies.
To find recommended ontologies for use in epigenetics, two textual data sets were used:
- Terms: a list of 142 genetic and epigenetic terms (keywords) selected from five selected papers (see references),
- Texts: a set of titles and abstracts of 250 recent publications found by the Web of Knowledge tool (Thomson Reuters) with the search condition "(epigenetics in Topic) and (plant in Topic)" treated as separate texts.
Recommended ontologies for epigenetics were found by two approaches:
- automatic annotation of Terms using the Ontomaton tool (http://www.isa-tools.org/) subsequently curated by a biologist,
- the approach described by Jonquet et al. (2010) using the ontology recommendation service provided by the Bioportal website (http://bioportal.bioontology.org/recommender). The method is based on automatic annotation of a text and scoring the ontologies providing annotations according to two parameters: "Score" indicating importance (coverage) of an ontology (equal to the sum of scores of all the annotations found in it, with direct annotations scored higher); and "Normalized score" (being the "Score" divided by the ontology size), indicating specificity and pointing out the ontologies specially suited for particular applications. The computations were done using scripts in Perl and Genstat (VSN Int.).
The automated annotation of 142 terms provided 462 annotations for 61 terms in 86 ontologies. The 81 terms for which no annotation was found are shown in Table 1. One should note that most of the epigenetic terms coined in the last years were not annotated. The curation of the rest of the results provided statistics summarised

in Tables 2 and 3. It appears that lack of definition in ontologies is a problem; the automatic search provided 2 annotations completely wrong, 13 by incorrect synonyms, and 97 partial ones. 89 annotations were classified as correct. The most frequently used ontologies are: NCIT, MESH and CRISP; GO and GRO are also useful, and have a relatively high mean weight. The top position of the Evidence Codes Ontology should be noted: it contains very good annotations of experimental procedures (protocols).
Results of automatic recommendation by the Recommender web service are summarised in Table 4 and can be interpreted as follows:
- According to the "Score", the top two ontologies are the same both for "Terms" and for "Texts": National Cancer Institute Thesaurus and Medical Subject Headings Thesaurus (MESH OWL version). Both contain very large sets of general terms for description of biomedical texts.
- According to the "Normalized Score", the top three ontologies for "Terms" are: Ontology for Genetic Interval, Gene Regulation Ontology, and SNP Ontology, with the Gene Regulation Ontology having the highest "Score" (and also the highest Score among those three for Texts). For "Texts", the top three ontologies are: BioPAX Ontology for Biological Pathways, IxnO Interaction Ontology, and PHARE Pharmacogenomic Relationship Ontology, with PHARE having the highest "Score".
To summarize, the recommendations found by both the automated and curated annotation and by the automatic recommendation are:
1. The large biomedical ontologies/vocabularies: Medical Subject Headings Thesaurus (MESH OWL version) and National Cancer Institute Thesaurus can be used for annotation of texts on epigenetic research to provide broad coverage of all general terms.
2. The specialized ontologies: Ontology for Genetic Interval, Gene Regulation Ontology, Gene Ontology, and Subcellular Anatomy Ontology can be used for annotation of more specific epigenetic terms, with addition of BioPAX Ontology for Biological Pathways, IxnO Interaction Ontology, and PHARE Pharmacogenomic Relationship Ontology for annotation of full texts. For plant science, the application of PHARE is questionable. It seems that the Gene Regulation Ontology (GRO, http://www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html) is a good target for addition of new terms appearing in epigenetic literature to improve its (relatively good at the moment) annotation coverage.

References

Jonquet C, Musen MA, Shah NH (2010). Building a biomedical ontology recommender web service. *Journal of Biomedical Semantics 1(Suppl 1): S1.*

Splinter E. et al. (2011). The inactive X chromosome adopts a unique three-dimensional conformation thet is dependent on Xist RNA. *Genes & Development 25: 1371-1383.*

Tolhuis B. et al. (2012). Chromosome conformation capture on chip in single Drosophila melanogaster tissues. *Methods 58: 231-42.*

Zhao Z. et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics 38: 1341-47.*

Johannes F. et al. (2008). Epigenome dynamics: a quantitative genetics perspective. *Nature Reviews 9: 883-890.*

Zhang X. et al. (2007). The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. *Nature Structural and Molecular Biology.*

Integrating phenotype ontologies across multiple species, Mungall et al. *Genome Biology 2010, 11:R2*

Rosemary Shrestha, Elizabeth Arnaud, Ramil Mauleon, Martin Senger, Guy F. Davenport, David Hancock, Norman Morrison, Richard Bruskiewich, and Graham McLaren. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature *AoB PLANTS (2010) 2010: plq008 doi:10.1093/aobpla/plq008 first published online May 27, 2010*

Pankaj Jaiswal, Shulamit Avraham, Katica Ilic, Elizabeth A.Kellogg, Susan R. McCouch, Anuradha Pujar, Leonore Reiser, Seung Y.Rhee, Martin M. Sachs, Mary L. Schaeffer, Lincoln D. Stein, PeterF. Stevens, Leszek P. Vincent, Doreen H. Ware and Felipe Zapata. PlantOntology: A controlled vocabulary of plant structures andgrowth stages. *Comparative and Functional Genomics, 2005, Vol 6 (7-8), 388-397*

De Matos, P.; Alcantara, R.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Spiteri, I.; Turner, S. et al. (2009). Chemical Entities of Biological Interest: An update. *Nucleic Acids Research 38 (Database issue): D249–54. doi:10.1093/nar/gkp886. PMC 2808869. PMID 19854951.*

Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Larisa N Soldatova, Christian J Stoeckert, Jr., Jessica A Turner, Jie Zheng, and the OBI consortium. Modeling biomedical experimental processes with OBI. *J Biomed Semantics. 2010*

Table 1. The list of terms with no satisfactory annotation was found

allele-specific chromosome conformation capture-on-chip (4C)	epi-loci	nonescaping gene
alternative chromatin state	epimutation	non-heritable chromatin variation
array-based chromatin data	escape gene	pericentromeric heterochromatin
bisulfate sequencing	gene density	reduced genome
chromatin immunoprecipitation coupled with hybridization to	gene desert	repeat-rich
chromatin immunoprecipitation coupled with sequencing (ChIPS	gene positioning	repositioning of chromatin
chromatin marks	gene relocalization	RNA::DNA hybrid
chromatin modifications	gene-dense region	RNA FISH
chromatin state	gene-intrinsic property	silent chromatin state
chromodomain-containing protein	gene-level distribution	spatial organization of DNA
chromosome conformation capture with sequencing (4C-seq)	gene-poor region	spatial reorganization
cis-acting proteins	gene-rich region	spider plot
DamID-chip method	genome-wide profile	T7 amplification
daughter cell	H3K9me3	target gene
discrete domain of enrichment	heat-shock	three dimensional (3D) structure of chromosome
discrete interaction domain (DID)	heterochromatic gene silencing	three-dimensional topology of DNA
divergent primers	heterochromatin protein-1	tissue-specific factor
DNA:DNA hybrid	high-density whole-genome tiling	tissue-specific gene
domainogram	microarray (DamID-chip)	topology of the chromosome
epiallel	higher-order chromosome folding	trans-acting environment
epiallelic determinants of phenotypic variation (phQTL-epi)	high-resolution cryo-FISH	trans-acting protein
epiallelic form	inactive chromatin	transcription inhibition
epigenetic landscape	intragenerational context	transcriptional activity
epigenetic variation	inverse PCR	transgenerational context
epigenome	long-range DNA interaction	trimethylation of H3K27 (H3K27me3)
epigenotype	methylation of H4K20 (H4K20me1)	
	methylation pathway	
	methylation-specific PCR	
	methyllysine	
	monoallelically expressed gene loci	
	nascent RNA FISH	

Table 2. Classification of annotations and assigned weights

Category	Number of annotations	Assigned weight
Single words only	87	1
Incorrect synonym	13	0
No definition	164	2
Ontology site offline	16	0
Correct	89	4
Partial	10	1
Totally wrong	2	0
Total number	381	

Table 3. Scores and counts for ontologies providing more than 3 annotations.

Ontology	Mean weight	Number of annotations

Evidence Codes Ontology	4.00	4
Experimental Factor Ontology	3.25	4
Medical Subject Headings	2.70	24
Gene Ontology	2.66	9
Gene Ontology Extension	2.60	15
National Cancer Institute Thesaurus	2.46	43
Sequence types and features	2.44	9
Ontology for Biomedical Investigations	2.20	5
Gene Regulation Ontology	2.18	11
CRISP Thesaurus, 2006	2.16	18
Cell Phenotype Ontology	2.00	5
Galen	2.00	4
Human Interaction Network Ontology	2.00	5
Interaction Network Ontology	2.00	6
Logical Observation Identifier Names and Codes	2.00	8
MESH Thesaurus (OWL version)	2.00	18
National Drug File	2.00	7
Read Codes, Clinical Terms Version 3 (CTV3)	2.00	6
SNOMED Clinical Terms	2.00	18
SNOMED International, 1998	2.00	10
Subcellular Anatomy Ontology (SAO)	2.00	4
Suggested Ontology for Pharmacogenomics	2.00	7
Neural-Immune Gene Ontology	1.85	7
IxnO	1.75	4
Foundational Model of Anatomy	1.70	10
PHARE	1.66	6
Systems Biology	1.60	5
SemanticScience Integrated Ontology	1.50	4
NIFSTD	1.44	9
Synapse Ontology	0.00	12

Table 4. Ontologies recommended for annotation of epigenetic terms and texts. The total number of ontologies found was 25 and 173 for “Terms” and “Texts”, respectively; in the table only the ontologies with the highest “Score” or “Normalized Score” are shown (for Terms – upper 25%; for Texts – upper 5%).

Ontology	Terms		Texts	
	Score	Normalized Score	Score	Normalized Score
BioPAX			9.00	**0.1324**
CRISP Thesaurus, 2006	59.00	0.0066		
ExO			5.00	0.0617
Fission Yeast Phenotype Ontology	43.00	0.0093		
Gene Regulation Ontology	50.00	**0.0988**	*11.29**	*0.0223**
Human developmental anatomy, timed version			27.60	0.0033
IxnO			7.00	**0.1321**
Logical Observation Identifier Names and Codes			30.06	0.0002
MESH Thesaurus (OWL version)	169.00	0.0006	47.33	0.0002
National Cancer Institute Thesaurus	185.00	0.0019	86.28	0.0009
NIFSTD	97.00	0.0011	25.36	0.0003
Ontology for Genetic Interval	26.00	**0.1262**	*9.39**	*0.0455**
PHARE			25.63	**0.1124**
Read Codes, Clinical Terms Version 3 (CTV3)			21.44	0.0001
Research Network and Patient Registry Inventory Ontology			4.00	0.0506
SNOMED Clinical Terms	81.00	0.0002	46.52	0.0001
SNP-Ontology	36.00	**0.0160**	*9.90**	*0.0044**
Synapse Ontology	71.00	0.0049		

Systems Chemical Biology/Chemogenomics			7.13	0.0686	
Taxonomic rank vocabulary			5.00	0.0847	
ThomCan: Upper-level Cancer Ontology			4.21	0.0751	
XEML Environment Ontology			10.00	0.0694	

* Values for ontologies not selected automatically for “Texts” but included for comparison.