# Ontologies for resolving semantic heterogeneity in information integration among plant phenomics databases

Anahita Nafissi[1], Benjamin Bruns[2], Fabio Fiorani[3]

**Abstract:** Increasing amounts of heterogeneous data are produced every year by plant researchers. For data management relational databases with application-specific schemas are mainly used in this field. However, due to absence of widely shared standards, data integration and exchange between independently developed and heterogeneous databases becomes very challenging. A critical point is to achieve semantic interoperability among these databases. The authors propose to use Semantic Web features for this integration task. Ontologies are the main core of the Semantic Web and are suitable to resolve semantic heterogeneity. In this work a semi-automated ontology based approach is defined for integrating heterogeneous data stored in distributed phenomics databases. The results of a real-world case study show that this approach creates reasonable semantic correspondences between domain-specific databases and publicly available ontologies and can significantly save time compared to classic (specification-driven) engineering approaches.

**Keywords:** Information integration, semantic heterogeneity, database-to-ontology, phenotyping

## 1 Introduction

Plant research is benefiting from an unprecedented methodological capability to generate large datasets addressing genotypic and phenotypic variation to selected environmental challenges. Non-invasive technologies and automated processes for data acquisition and analysis have been developed, with the aim to reduce the so-called "phenotyping bottleneck" ([FS13]). Plant phenotypic data, such as image repositories and computed plant traits, environmental data, and experimental (meta-)data, require the development of appropriate data management schemas and are generally stored in databases (e.g., the Phenopsis database [Fa11], the Golm plant database [Kö08], PhenOMIS [Sc13]). Biologists have characterized the responses of a wide range of plant species to their environment. As a result, hundreds of experiments have generated large phenotypic datasets. Generally, in the absence of widely accepted standards each laboratory has pragmatically developed local schemas and solutions resulting in distributed and heterogeneous database systems. Thus, a quantitative and comparative analysis across these datasets is currently cumbersome.

[1] IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, ana.nafissi@gmail.com
[2] IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, b.bruns@fz-juelich.de
[3] IBG-2: Plant Sciences, Forschungszentrum Jülich GmbH, 52425 Jülich, f.fiorani@fz-juelich.de

## 2 Methodology

The authors propose a semi-automated ontology-based approach for integrating heterogeneous plant phenomics databases. According to this approach semantic equivalences of database terms are determined with terms publicly available in ontologies applicable to this research field. The simplified architecture of this approach is given in Figure 1. For this an application-specific database DB and a global ontology O are given. The global ontology is a publicly available ontology (e.g. Plant Ontology [Fa11]) which describes the semantics of the domain of interest. The approach is composed of two processes: transformation and mapping. In the transformation process a local ontology LO for DB is generated. In the mapping process, the terms of LO are mapped to the terms of O.

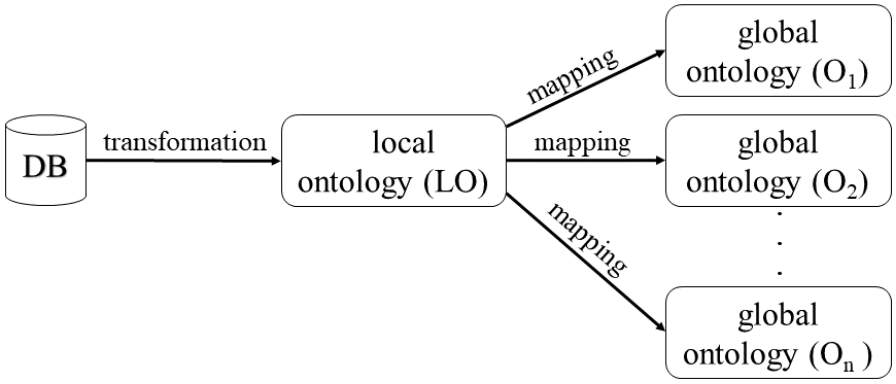


Fig. 1: The principle ontology-based approach for achieving data integration and exchange

The required steps in the transformation process are as follows. First, relevant tables and relevant attributes are manually determined. Then ontology concepts for all chosen table names and attributes are defined (in OWL2 syntax). After that selected properties are manually integrated into the ontology definition. Input files (‘concept candidates’) based on table and attribute names are proposed by a Java program parsing the schema of DB.

In the mapping process the results r= (LC, GC) are determined and contain semantically corresponding concepts LC from LO and GC from O. For achieving a high number of mapping results, the local ontology is mapped to several global ontologies $O_1$, $O_2$, …, $O_n$ which are indexed (currently over 200 ontologies with 5 million terms) by the free EMBL-EBI Ontology Lookup Service ([Ju15]).

For determining the mapping results r, two mapping rules $R_1$ and $R_2$ are defined (see Table 1) based on exact string or substring comparisons on concept names. To rank the mapping results r, a score to each mapping rule R is assigned. The scores are chosen according to application and should generate reasonable mapping results (see Table 2). A score to each mapping result r is assigned so that the score of r is equal to the score of

the holding rule R (evaluating $R_2$ only when $R_1$ does not apply). A mapping result with a higher score is considered more probable than one with a lower score. Note that the rules should only be applied to meaningful (self-descriptive) English attribute and table names. Acronyms should been exchanged by their full names in a pre-processing step.

| Rule | Score | Definition |
|---|---|---|
| $R_1$ | 10 | if *LC=GC*, then *r=(LC,GC)* |
| $R_2$ | 1 | If *LC≠GC* and *LC* contains *GC* (or vice versa), then *r=(LC,GC)* |

Table 1: The mapping rules used for our PhenOMIS case study

## 3 Results

To evaluate the quality of the mapping results generated by the proposed approach standard quality parameters are calculated. For this |*RR*| is introduced as the number of mapping results returned automatically and checked for correctness by a human expert (i.e. true positives), |*TM*| as the number of correct mapping results determined manually (i.e. true positives plus false negatives) and |*TR*| as the total number of mapping results returned automatically (i.e. true positives plus false positives). Now *Recall*, *Precision* and *False Positive Rate* (FPR) can be defined as:

$$Recall = \frac{|RR|}{|TM|} \qquad Precision = \frac{|RR|}{|TR|} \qquad FPR = \frac{|TR| - |RR|}{|TR|}$$

As a case study the information system PhenOMIS ([Sc13]) has been chosen which contains plant phenotyping traits and environmental information. PhenOMIS consists of several subsystems including the CoreDB (plant tracking) and several structurally similar measurement station databases. According to the proposed transformation process the local ontologies CoreO (68 concepts) and MeStO (65 concepts) are generated for the CoreDB and the measurement station databases, respectively. Afterwards the mapping process (see Table 1) is applied to CoreO and MeStO.

| Local ontology | \|*TM*\| | \|*TR*\| | \|*RR*\| | *FPR* | Recall | Precision |
|---|---|---|---|---|---|---|
| CoreO | 58 | 30 | 26 | 0.13 | 0.45 | 0.87 |
| MeStO | 62 | 47 | 46 | 0.02 | 0.74 | 0.98 |

Table 2: Key figures for the evaluation of our PhenOMIS case study

The evaluation results are shown in Table 2. In both cases the precision values are significantly higher than the recall values. So with a high probability the generated results have been considered as correct. However, in particular in the case of CoreO, many concepts (about half) could not be mapped automatically. However, compared

with previous projects using a classic (spec. driven) ontology engineering approach the time required for the involvement of a human expert could be significantly reduced.

## 4 Conclusions

In this work a semi-automated procedure has been introduced for enabling data integration among heterogeneous plant phenomics databases. In the case of PhenOMIS results showed that this approach eases and shortens the process of defining formal mappings between local and global ontologies by generating reasonable semantic correspondences. The quality (notably the recall value) of the mapping process was here mainly determined by the semantic expressiveness of LO and the portion of modelled domain-specific knowledge (specialized CoreO vs. generic MeStO). Regardless of the concrete local ontology the generated mapping results should be evaluated by a human expert (plant biologist) to eliminate invalid results ("false positives").

By ontologizing PhenOMIS the basis for implementing semantic or federated query capabilities has been created. For this (freely) available software solutions (e.g. D2RQ: http://d2rq.org/) are available which provide semantic query capabilities for non-subject-predicate-object databases (e.g., relational databases) based on a formal mapping concept (ontology).

*Acknowledgment*: This work was partially performed within the German-Plant-Phenotyping Network project (DPPN) which is funded by the German Federal Ministry of Education and Research (project identification number: 031A053).

## References

[Fa11] Fabre, J. et al: Phenopsis DB: an information system for Arabidopsis thaliana phenotypic data in an environmental context. In: BMC Plant Biol. 11, pp. 77, 2011.

[FS13] Fiorani, F.; Schurr, U.: Future scenarios for plant phenotyping. In: Annual Review of Plant Biology 64(1), pp. 267-291, 2013.

[Ja05] Jaiswal, P. et al: Plant Ontology: a controlled vocabulary of plant structures and growth stages. In: Comparative and Functional Genomics 6(7-8), pp. 388-397, 2005.

[Ju15] Jupp, S. et al: A new Ontology Lookup Service at EMBL-EBI. In: Proceedings of SWAT4LS International Conference, Cambridge, 2015

[Kö08] Köhl, K. et al: A plant resource & experiment management system based on the Golm Plant DB as a basic tool for omics research. In: Plant Methods 4(1), pp. 11, 2008.

[Sc13] Schmidt, F. et al.: A distributed information system for managing phenotyping mass data. In: Referate der 33. GIL-Jahrestagung, pp. 303-306, Potsdam, 2013.