OPTIMAS-DW, MetaCrop and VANTED: A Case Study for Data Integration, Curation and Visualisation in Life Sciences

Christian Colmsee[1], Tobias Czauderna[1], Eva Grafahrend-Belau[1], Anja Hartmann[1], Matthias Lange[1], Martin Mascher[1], Stephan Weise[1], Uwe Scholz[1] and Falk Schreiber[1,2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, Corrensstr. 3, 06466 Stadt Seeland, Germany
{colmsee,czaudern,grafahr,hartmann,lange,mascher, weise,scholz,schreibe}@ipk-gatersleben.de

[2]Martin Luther University Halle-Wittenberg, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany

Abstract: Since the data volume in life sciences has been growing exponentially in recent years, it is indispensable to develop databases and tools for efficient data integration, curation and visualisation. Focusing on data handling in crop plant research, this paper presents an approach, which combines (i) a data warehouse (OPTIMAS-DW) for integrating experimental data, (ii) an information system (MetaCrop) for manually curated biochemical pathways, and (iii) a visualisation software (VANTED) for integrated data visualisation. The functionality and usability of the concept will be illustrated by a use case.

1 Motivation

In recent years, a large increase of the amount of data in life sciences could be observed. To handle the vast amount of data arising from different –omics domains, it is necessary to find solutions for effective data integration and visualisation. In addition, aspects of data curation and data quality need to be considered. Figure 1 illustrates a general concept of data handling and visualisation in life sciences: Experimental data from different –omics domains need to be integrated into appropriate databases or data warehouse structures to ensure a persistent data storage as well as the linkage of these –omics domains. Furthermore, there is a strong need for dealing with fine-grained high-quality information from the literature, such as information about metabolic pathways, which has to be curated manually. The curation enables to maintain a public reference and it improves the quality of stored information in the focus of the own research fields. The experimental data can then be mapped onto the curated data to receive a better understanding of the relation between experimental data and for example biochemical networks. The consistent use of ontology

terms enables an efficient linkage of the data. Finally, the integrated and mapped data should be visualised.

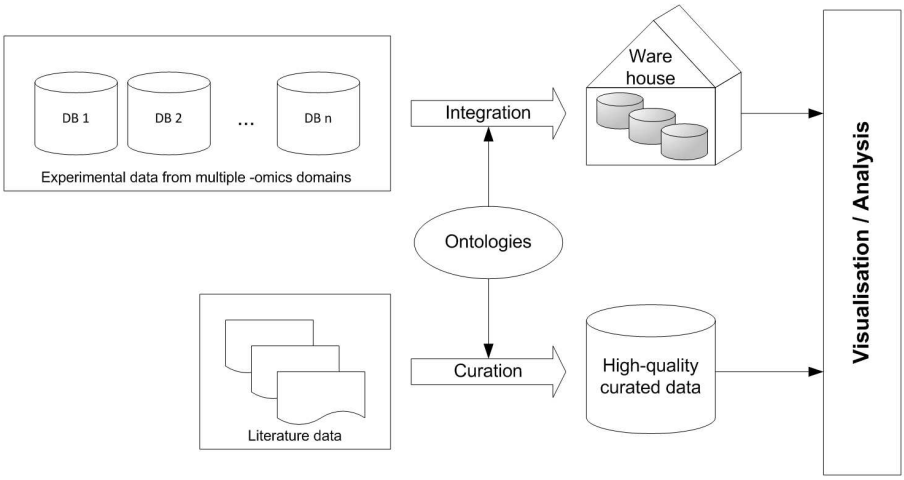

Figure 1: Experimental data and data from literature have to be stored in databases. Data integration and curation as well as the annotation by ontologies are necessary steps to increase the utility of the data. The data can then be used for visualisation and analysis.

There are several databases and tools existing for data integration and data visualisation in life sciences, such as UniVIO, a database that integrates and visualises hormone and transcriptome data in rice [KAK+13] or the KEGG database, which includes, for example, metabolic pathway data from different species [KGS+12]. Tools such as Cytoscape [SMO+03] allow the network-based visualisation and analysis of biological data. In this paper we describe our approach by presenting a case study for integrating and visualising life science data from different –omics domains with a focus on crop plants.

2 Methods

For implementing the concept described above, a data warehouse for experimental data management and integration (OPTIMAS-DW [CMC+12]), a database for manually curated data (MetaCrop [SCC+12]) as well as a tool for data visualisation and analysis (VANTED [RJH+12]) were developed. Figure 2 illustrates these developments in the context of the general concept of data handling in life sciences.

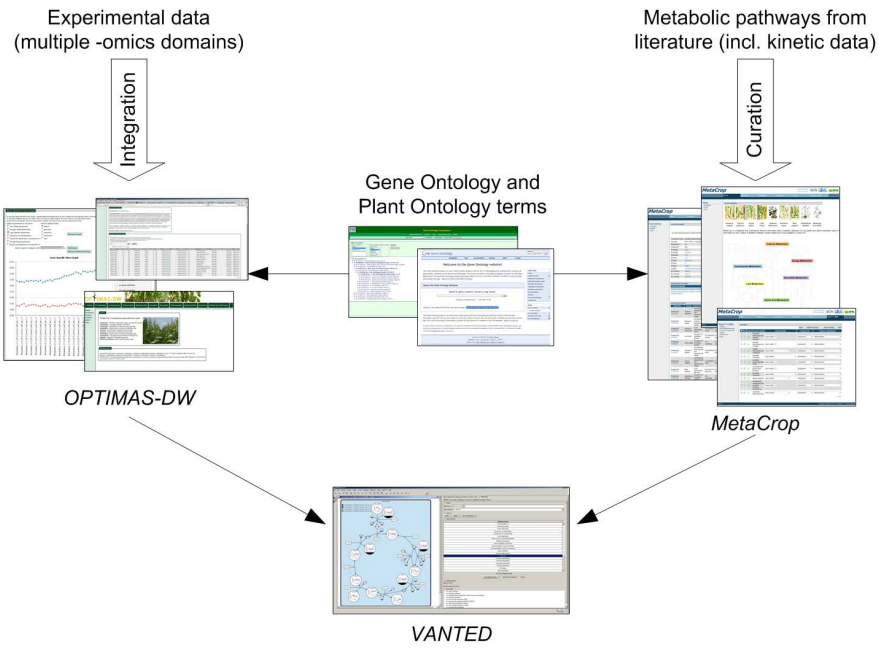

Figure 2: Life sciences data integration, curation and visualisation.

2.1 Data Integration

OPTIMAS-DW is a comprehensive data warehouse for maize containing transcriptomic, metabolomic, proteomic, ionomic and phenomic data. In order to enable integration and analysis across different data domains, the data have to be linked. Therefore, OPTIMAS-DW extensively uses metadata describing the samples, such as genotype, plant growth stage, treatment and plant anatomy. OPTIMAS-DW provides several BLAST [AGM+90] results describing the gene functions, e. g. a BLAST2GO [CGGG+05] was performed in order to enrich maize genes with Gene Ontology (GO) [HCI+04] terms. Furthermore, the plant anatomy of experiment samples was annotated with Plant Ontology (PO) [JAI+05] terms.

For a data warehouse like OPTIMAS-DW, data quality control is indispensable. Experimentalists use provided data templates to handle their experimental data. The templates include controlled vocabulary to ensure that different experiments are comparable. The filled templates are checked during data import, thus ensuring that only data are stored in the data warehouse that corresponds to the correct controlled vocabulary as well as to defined data types. The controlled vocabulary can be extended continuously according to the experimentalists' needs. Therefore, the templates are adapted as well as the database content.

Besides the storage of experimental data, different data analyses were performed and

stored inside the data warehouse as well. Both experimental data and analysis results are available via a web interface.

The system was designed with the aim of supporting an easy extension and adaption. The data warehouse can be easily extended by further data domains. Furthermore, it is possible to adapt the system for other plant organisms as well.

2.2 Data Curation

To provide a database for manually curated crop plant metabolic pathways, MetaCrop, a repository comprising high-quality data about crop plant metabolism including reaction kinetics of seven major crop plants and two model plants was developed. The curation process ensures high-quality data due to the extraction of important information from scientific literature by experts in their field. The conversion into a MetaCrop entry additionally includes the enrichment with ontology terms from GO and PO.

The web interface of MetaCrop allows a user to navigate through different levels of detail of a given metabolic pathway. For data analysis and visualisation, MetaCrop supports the following features: (i) The biochemical pathways are visualised using SBGN, the Systems Biology Graphical Notation [LNHM+09]. All pathways can be downloaded as SBGN-ML and GML file format. (ii) To support the creation of user specific metabolic models in SBML [HFS+03], the system provides an add-on for exporting selected data in SBML format. Both SBGN and SBML use the Systems Biology Ontology (SBO [LNCL07]). (iii) In addition, a set of SOAP-based web services allows the interaction with external tools.

2.3 Data Visualisation

To support data visualisation VANTED, a tool for the visualisation and analysis of networks containing experimental data, has been developed. Besides the creation of user specific pathway maps, VANTED enables users to load pathways stored in MetaCrop. Experimental data from OPTIMAS-DW can then be mapped onto these pathways, as described in [JRC+12]. The tool is available as a JAVA web start application as well as for download allowing users to use VANTED within their preferred operating system.

Figure 3 shows an example of data mapping using VANTED. Here, the leaf gradient experiment stored in OPTIMAS-DW was used. In this experiment a systematic analysis of a developmental gradient of the third maize leaf was accomplished to study organ development as well as physiological and biochemical functions [PBS+11]. The data were mapped onto the TCA cycle stored in MetaCrop. To perform the data mapping three steps are necessary: (i) The TCA cycle has to be loaded from MetaCrop into VANTED. (ii) The template containing the leaf gradient metabolite data has to be loaded into VANTED. (iii) The metabolite data from the experiment have to be mapped onto the TCA cycle. For each metabolite measured in the experiment the corresponding node in the network includes a

line chart. The metabolite concentration for each part of the third maize leaf from tip (1) to base (10) is visualised in detail. The SBGN notation is used for the graphical representation: the *process nodes* (squares) represent chemical reactions, which are catalysed by enzymes represented as *macromolecules* (rectangular containers with rounded corners). The *catalysis* is represented by an arc with an empty circle arrowhead. The metabolites are shown as *simple chemicals* (circular containers) and are either reactants or products of a reaction. When a metabolite occurs multiple times, it is decorated with a *clone marker* (black bar at the bottom of the circle), e. g. NAD+.

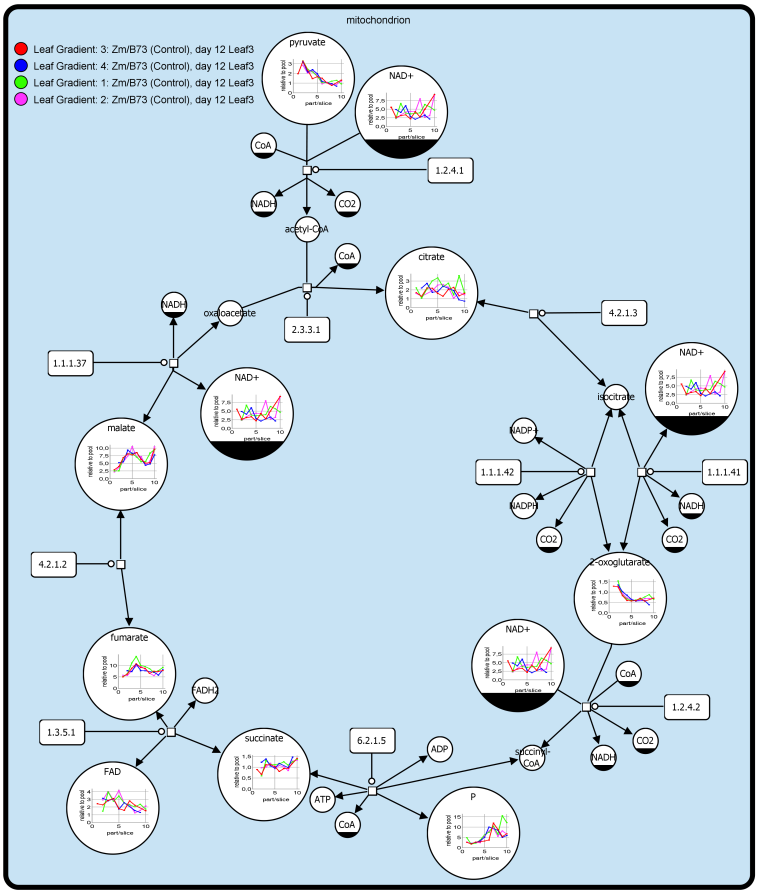

Figure 3: Metabolite data from the leaf gradient experiment was mapped onto the TCA cycle using VANTED [CMC+12].

In addition to the above mentioned use case, other data can be mapped onto the pathways as well. By providing for example a mapping between genes and enzyme codes, expression data can be mapped onto the pathway. To compare the data with phenotypic traits, the phenomics data can be represented as a node in VANTED.

3 Conclusion

In this paper an approach for data integration, curation and visualisation in life sciences, mainly focusing on crop plants, was presented. The usability of this approach was demonstrated by a case study incorporating three systems. OPTIMAS-DW enables to integrate data from different –omics domains, whereas MetaCrop allows to manage fine-grained and high-quality pathway data being manually curated from the literature. Finally, VANTED provides the possibility of a joint visualisation of this data. With this approach, life science researchers, especially in the field of systems biology, can be supported.

References

[AGM+90] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[CGGG+05] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, and M. Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

[CMC+12] C. Colmsee, M. Mascher, T. Czauderna, A. Hartmann, U. Schlüter, N. Zellerhoff, J. Schmitz, A. Bräutigam, T.R. Pick, P. Alter, et al. OPTIMAS-DW: A comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize. *BMC Plant Biology*, 12(1):e245, 2012.

[HCI+04] M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(D1):D258–D261, 2004.

[HFS+03] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[JAI+05] P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, et al. Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and Functional Genomics*, 6(7-8):388–397, 2005.

[JRC+12] A. Junker, H. Rohn, T. Czauderna, C. Klukas, A. Hartmann, and F. Schreiber. Creating interactive, web-based and data-enriched maps with the Systems Biology Graphical Notation. *Nature Protocols*, 7(3):579–593, 2012.

[KAK+13] T. Kudo, K. Akiyama, M. Kojima, N. Makita, T. Sakurai, and H. Sakakibara. UniVIO: a multiple omics database with hormonome and transcriptome data from rice. *Plant and Cell Physiology*, 54(2):e9, 2013.

[KGS+12] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.

[LNCL07] N. Le Novere, M. Courtot, and C. Laibe. Adding semantics in kinetics models of biochemical pathways. In *Proceedings of the 2nd International Symposium on experimental standard conditions of enzyme characterizations*, pages 137–153, 2007.

[LNHM+09] N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M.I. Aladjem, S.M. Wimalaratne, et al. The systems biology graphical notation. *Nature Biotechnology*, 27(8):735–741, 2009.

[PBS+11] T.R. Pick, A. Bräutigam, U. Schlüter, A.K. Denton, C. Colmsee, U. Scholz, H. Fahnenstich, R. Pieruschka, U. Rascher, U. Sonnewald, et al. Systems Analysis of a Maize Leaf Developmental Gradient Redefines the Current C_4 Model and Provides Candidates for Regulation. *Plant Cell*, 23(12):4208–4220, 2011.

[RJH+12] H. Rohn, A. Junker, A. Hartmann, E. Grafahrend-Belau, H. Treutler, M. Klapperstück, T. Czauderna, C. Klukas, F. Schreiber, et al. VANTED v2: a framework for systems biology applications. *BMC Systems Biology*, 6(1):e139, 2012.

[SCC+12] F. Schreiber, C. Colmsee, T. Czauderna, E. Grafahrend-Belau, A. Hartmann, A. Junker, B.H. Junker, M. Klapperstück, U. Scholz, and S. Weise. MetaCrop 2.0: managing and exploring information about crop plant metabolism. *Nucleic Acids Research*, 40(D1):D1173–D1177, 2012.

[SMO+03] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.