Using Data Warehouse Technology in Crop Plant Bioinformatics

Christian Kuenne, Ivo Grosse, Inge Matthies, Uwe Scholz, Tatjana Sretenovic-Rajicic, Nils Stein, Andreas Stephanik, Burkhard Steuernagel and Stephan Weise*

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3, 06466 Gatersleben, Germany

Summary

Plant-specific data is managed in heterogeneous formats and is dispersed geographically. Based on this data, efficient analyses require a materialised integration, often realised with data warehouse technology today. We describe the requirements, problems and solution strategies for domain-crossing integration as the fundament for analysing plant biological data based on three current case studies. First, we introduce a system for retrieval of markers and mapping positions based on clustering of ESTs. The second case study illustrates the steps for diversity studies after genotyping a collection of about 3,000 ryegrass accessions (*Lolium* spp.), whereas in the third example data of approximately 250 barley cultivars (*Hordeum vulgare*) were used for associating haplotype- and SNP-patterns with malting parameters. For all case studies, we integrate data from different domains - sequence and marker data as well as IPK Genebank data including passport and phenotypic information. Specific problems associated with plant biological data and possible solution strategies are shown.

1 Introduction

Today, data amount is increasing enormously in plant biology. The use of modern high-throughput methods is more and more ruling out the traditional way of researching [1]. The scientific focus is moving away from the single-data-domain and problem-oriented approach towards work crossing the borders of data domains. Bioinformatic tools help to analyse data at a large scale. Often, extensive data sets gained from biological experiments cannot be handled individually, especially in the area of genomic data.

While the main focus of the scientific community is on human and several animals, such as fruit fly and mouse, plants are often under-represented. Several projects are working on a certain plant species and maintain relatively independent data sources, e. g. for barley or pea, designed for special application areas by using different technologies [2].
Very often, data about plant genotypes is quite rare and needs to be supplemented by data of related genotypes. We think, for integration and analysis of plant data, the issues of the heterogeneous and sparse data of plants should be addressed.

In this study, we use data from different domains (phenotypic data, marker data, sequence data and taxonomic data), which can be linked using abstractions of the plants as central objects. Therefore, so-called passport data is used. Passport data serves as identificator of genotypes. It comprises parameters such as accession number, habitat or scientific name. A possible linkage

*To whom correspondence should be addressed. Email: weise@ipk-gatersleben.de

of different data domains via passport data is shown in Fig. 1. Bioinformatical applications can work on data from different domains or on a single domain, respectively. The required data is available as Foxpro and MS Access databases, MS Excel and HTML files (phenotypic data), Oracle databases and flat files (marker data, sequence data and passport data). In order to establish a data domain spanning analysis platform, data need to be integrated.

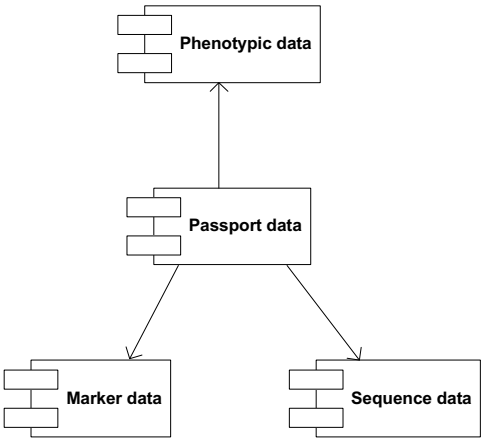

Figure 1: Linkage of data from different domains

In this paper we will describe the integration of data in crop plant bioinformatics. Typical problems occurring in this area of research will be discussed. An architecture for integrating and analysing data is shown. Finally, three case studies from crop plant biology are described.

2 Approach

2.1 Background

Integration is a service, which combines the contents of multiple, heterogeneous data sources [3]. Data sources in biology are often heterogeneous in different ways, e. g. their application, structure, content and software platform they are based on. The combination of such sources, often from the World Wide Web, is the typical task of data integration in bioinformatics.
There are two common methods for data integration in bioinformatics:

Virtual or logical integration: This is the standard approach for integrating data using the internet. Here, a query is forwarded to different data sources; the results will be combined in a report. The integration takes place at runtime; data is not stored locally. Examples for this approach are federated databases or mediators.

Materialised or physical integration: Data sources are scanned for new data at regular intervals. Data is integrated and stored physically. Queries are made against this data. This is the standard procedure for data warehouses.

Many databases in biology were created spontaneously and altered during time according to different user demands. Often, their structure is not sufficient and mainly they are based on

flat-files. In many cases, data access is only possible via web interfaces. Additionally, transformations are necessary, which cannot always be processed automatically and claim manual curation. Online access, particularly to web-based data sources, usually implicates high response time, whereas simple flat files cannot be accessed online at all. Thus, virtual integration does not suit for our purposes. In our opinion, the most promising approach for use in plant bioinformatics is the utilisation of materialised integration and data warehousing methods.

Data warehouses are collections of data, which are subject-oriented, integrated, time-variant and non-volatile [4]. In companies, they are used to support management decisions. Relevant data is extracted from different sources, transformed and finally loaded into the data warehouse. These three steps form the so-called Extract-Transform-Load (ETL) process. The data warehouse itself is intended for long-term storage. For analyses, data can be stored in data-marts, which are partial copies of the warehouse data for, e. g. structural business units or certain applications. Therefore, data is often stored in constructs called star schema, which consists of dimensions grouped around indicators (facts). Together, they form a data-cube.

The above-described data warehouse concepts do not always meet the requirements of, e. g. genomic data or proteomic data [5], because life-science data is subject to permanent changes and the process of creating a data warehouse is time-consuming and costly. Other aspects of data warehouses as, e. g. time variances as useful in business, are relatively unimportant in biology [6]. Nevertheless, there is an increased use of data warehouse methods in biological research, but a specific adaptation to the respective objective is necessary.

2.2 Problems in crop plant bioinformatics

Two major kinds of problems can occur by data integration in crop plant bioinformatics:

1. There are the difficulties of integrating data from the same domain, which are, e. g. originating from different experiments, sources etc.
2. It can be exhausting to try to link data records between different domains of data.

Reasons for difficulties of integrating data from the same domain might be the following:

IT reasons: In the last two decades, there is an increasing number of algorithms and applications, which are used in the bioinformatics community. Most of them work (due to their evolution) with different input and output formats, interfaces, models and mainly file-based. Standardised formats, such as SBML [7], MAGE-ML [8] or the FAO/IPGRI Multi-Crop Passport Descriptors (MCPD) [9] are unfortunately not commonly used. Therefore, the integration of both tools and data becomes quite difficult.

Biological reasons: Due to complex biological functions in plants, which are activated at different times or locations, also depending on differential gene expression and regulation, there are many various effects on phenotypes. Furthermore, there are external effects, such as biotic/abiotic stresses or environmental conditions, which possess an high impact on plants and thus resulting data. If modelled correctly, these interactions would make the integration extremely complex. In reality, such information is often not available or poorly documented. Hence, additional work of domain experts is strongly needed, e. g. for manual curation.

Process of data acquisition: This process includes all steps of the generation of data. It can be divided into acquisition of raw data and acquisition of derived data. In both cases, supplemental information, so called metadata, is crucial for data integration. Raw data is worthless if biological materials and experimental procedures are not documented expediently and results cannot be assigned to certain biological issues. The same problems are occurring if further calculations or method utilisations on raw data are not documented. Data derived this way would not be reliable because statistical errors are unknown and other computational errors cannot be tracked.

Conceptual reasons: Some information systems only allow the authors to change a data record. Thus, a lot of erroneous data will not be corrected. When certain biological facts are not well investigated or even unknown, the use of suitable prediction methods becomes necessary and might lead to problems. Integration of such data leads to possible multiplication of errors, especially if such information is handled as truth or if it is generalised [10]. Several studies are shown in [11] where, in the case of derived information about structural or functional features of sequences, the error rates are more than 40%.
Frequently, researchers in crop plant biology do not use the same vocabularies, although controlled vocabularies (in the context of ontologies) are more and more used in biology, e. g. the Gene Ontology [12] or the Plant Ontology [13]. Thus, vocabularies are successively unified in this area. However, in plant genetic resources, controlled vocabularies are not used enough yet. Often, there are different ranges of values used for a single trait (although there are recommendations given for certain crops, e. g. barley [14]) or there are inconsistencies inside the data, which makes it difficult to integrate and to compare such data. Furthermore, experimentalists often do not use standardised methods. This leads to problems when integrating and comparing data conducted in comparable experiments, but in different laboratories [15, 16].
Thus, only domain experts can help to improve data quality (within one data domain) in order to integrate the data [17].

Problems occur in most cases while connecting data from different domains. Often, there is no explicit linkage between records. A possible solution is the use of similarity-ranking methods or equivalence methods.
In the first step, attributes for mapping onto surrogate IDs have to be identified in order to map data from two or more data domains. This can be done by experts manually, or semi-automatically by using data-dictionaries or controlled vocabularies [18]. If those key attributes were found, the second step is to map their values. In case of numerical values this is quite trivial, but becomes difficult in case of alphanumerical values. Regarding the latter case, we often have been faced with different spelling of cultivar names etc. For example, the name of the barley cultivar “Ingrid” comes along with different dictions as “Ingrid WT” or “Zweizeilige Gerste Ingrid” containing additional information. Thus, established approaches as edit distance [19] or the Soundex algorithm [20] do not provide sufficient results. For example, the Soundex algorithm computes for the different genotypes “Ingrid BC mlo5” and “Ingrid MLG” the same value (I526), whereas for the identical genotypes “Ingrid WT” and “Zweizeilige Gerste Ingrid” the two different values I526 and Z242 are computed. This matches with [21] who show that only about a third of the identified mappings is correct. A longest common substring algorithm [22] or a local alignment algorithm [23] provides the best results in our case. Nevertheless, results need to be checked manually in many cases.

We like to point out that there are many interesting and promising approaches for record matching existing. However, our intention is to propose a variable process flow, which enables us to integrate and analyse data in a flexible way without substantial adaptations each time. Therefore, we suggest to focus on a limited set of algorithms, which should be used inside of the database, e. g. as user defined functions (UDF) using open-source implementations.

2.3 Proposed architecture

For flexible integration and analysis of crop plant data, we propose a multilayered architecture, which is shown in Fig. 2.

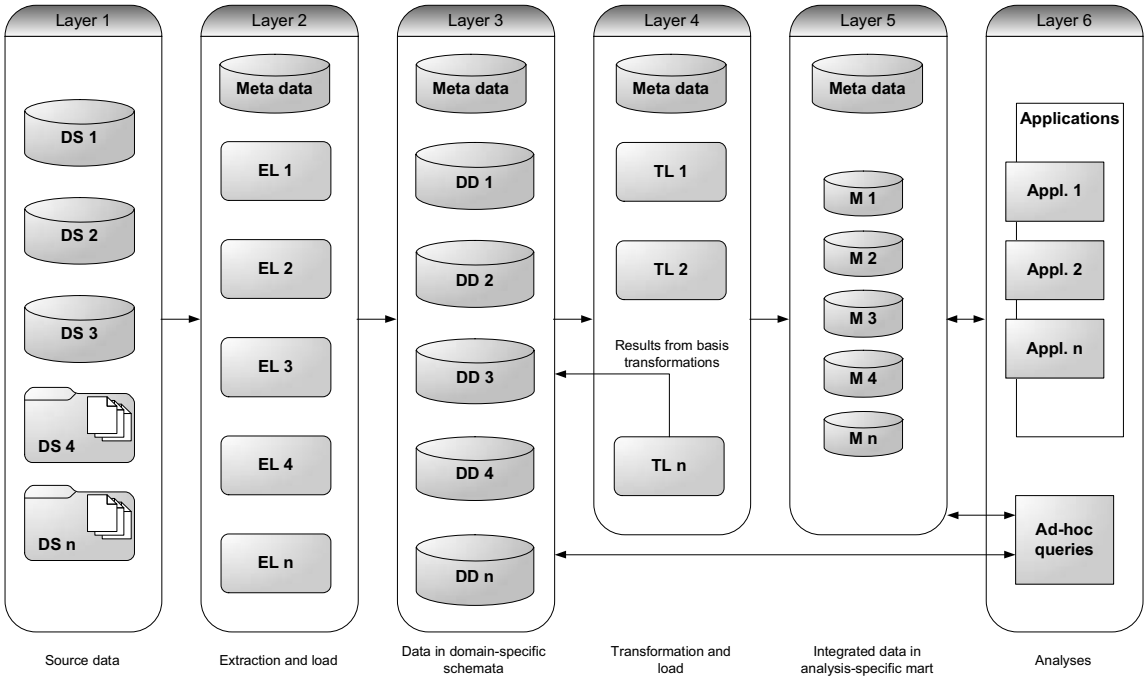

Figure 2: Overview of the proposed architecture. Applications of this architecture are illustrated in Fig. 3–5.

Layer 1: Operational data The first layer consists of source data, which can be available via operative database systems as well as flat-files or XML-files. This data is usually accessed read-only.

Layer 2: ETL The second layer comprises the extraction, transformation and loading steps that are necessary to import the desired data. Therefore, established technologies such as the Oracle Warehouse Builder [24] or, especially in the area of bioinformatics data sources, the Sequence Retrieval System (SRS) [25] can be used. We propose to focus on extraction and loading steps in this layer. However, basic transformations can also be done in this layer, but our intention is to store as much original data as possible in the data pool, which will be described in the next paragraph. In layer 4 further transformation can be processed in order to be flexible regarding various transformation methods or upcoming approaches. This is a tribute to the situation of the data in crop plant bioinformatics, which is often available in

proprietary formats only. Analysis-specific data marts can be accomplished later by additional transformation and loading steps.

Layer 3: Data pool The centre of our architecture is a kind of operational data store. In data warehouse terminology, an operational data store (ODS) is a database schema with the purpose to integrate non-aggregated / non-summarised and current data from multiple, often heterogeneous, sources [4]. Furthermore, the data can be augmented by derived (secondary) data.
Plant research is often driven by its objectives and scientific questions, which are valid for a limited time. Therefore, in our opinion it does not make sense to develop so-called "star" or "snowflake" schemata or even a general data warehouse schema, which is fixed for certain analyses and is filled with new data at regular intervals. Hence, it might be sufficient, to maintain a basis database (or data pool), which can become the basis of later analyses (which can be of different nature; not only OLAP frequently used in business-oriented DWH). In contrast to an operational data store, we propose to use a data pool, which can keep data for a longer period of time. For this basis database, the use of entity-attribute-value (EAV) schemata adapted to the specific data domains is suggested. The EAV approach allows to store heterogeneous data in a generic way. The underlying schema of the integrated data requires no alteration when data sources are extended or further data sources are integrated. A couple of years ago, this approach became popular in life-sciences in the context of clinical information systems (e. g. see [26, 27]). Originally, attribute-value pairs were used as LISP association lists in the area of artificial intelligence [28]. In order to avoid storing all data in, e. g. a LOB field, we suggest to use different generic schemata, one for each data domain. This is advantageous by taking into account specific features of data domains.

Layer 4: Transformation and loading The fourth layer consists of additional transformation and loading steps which are necessary for filling the analysis-specific data marts described in layer 5 with data from the data pool of layer 3, and provides methods for preparing (e. g. normalisation) data in a second later transformation step. Three issues are important:

(1) development of the mart schema and identification of linkages between records from different domains, (2) consideration of data quality (see also layer 2) and (3) preparation of data for later analyses, e. g. discretisation, outlier treatment, null / missing values, normalisation, aggregation.

The results of basis transformations could also be written back to Layer 3 (Fig. 2). In that case, transformed data should be stored additionally and original data should not be replaced. Hence, further transformation steps could be applied to the original data in the future.

As mentioned above, passport data represent the central object plant which links different domains of data. When connecting phenotypic data with sequence, marker or expression data it works well. However, practical experience has shown that physically often sequence data is in the centre. This is due to the fact that sequence data is often the main focus of research (Many sequence databases exist with links to other databases as annotations). The Crop EST Database [29] is a central resource at IPK.

Layer 5: Analysis-specific data marts Our intension is not to develop an all-embracing data warehouse as basis for all analyses. In our opinion, this is not suitable for crop plant bioinformatics. Instead, this layer consists of several data marts, which meet specific requirements for certain analyses.

Layer 6: Analyses While biological research is so multifaceted, we do not focus on one analysis method or one class of methods. In contrast, due to the flexible nature of the architecture presented here (esp. the analysis-specific data marts), the analyses layer enables us to use data mining methods, e. g. for data driven analyses, as well as statistical methods.

The use of OnLine Analytical Processing (OLAP) [30] is also promising, e. g. in case of expression data, but in our opinion this is not always recommendable for standard use in crop plant bioinformatics.

The layers 2–5 should contain metadata, e. g. the sources the integrated data is originating from, or loading and transformation steps, which are applied to that data. Such information is essential in order to document and to make comprehensible the particular steps accomplished during the process of integration. Moreover, metadata are of high importance for making integrated data comparable.

3 Case studies

In the following section, we present three case studies applying the architecture described above, in order to clarify our proposal.

3.1 Sequence Mapping eXplorer (SMeX)

Development of molecular markers and their genetic and physical mapping is time-consuming and expensive. In this context, researchers need to know if there are any mapping populations available for either particular or homologous or otherwise related sequences. Therefore, a decision support system called Sequence Mapping eXplorer (SMeX) was developed, which is based on integrated operational barley sequence data from the IPK Crop EST Database [29] and marker data from the IPK Molecular Marker Database including the mapping data of approx. 1,000 barley markers [31]. Additionally, wheat ESTs from NCBI GenBank [32], cluster from TIGR wheat gene index [33] and wheat marker data from GrainGenes wEST database [34] were integrated into either a domain-specific schema of our data pool covering sequence data or a schema covering marker data using different extraction and load processes (Fig. 3). This starting data basis will grow by executing existing or setting up new extraction and load processes, dependent upon availability of new data in existing or in new operational data sources.

Further integration into the SMeX data mart now combines crop plant EST data, clustering data and multiple EST alignments, population data, molecular markers and their genetic mapping positions (Fig. 3). Here, materialised views which provide an application-specific view on the before mentioned data were used, while this data was stored locally in the data mart. Loading

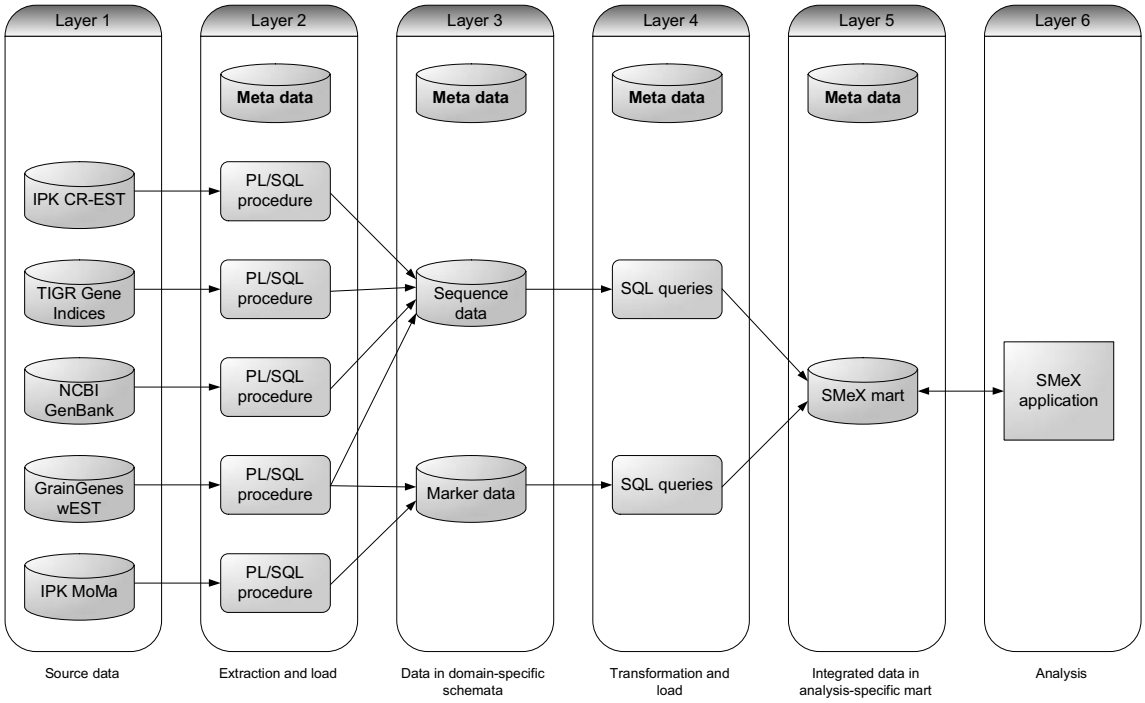

Figure 3: Implementation of our architecture for the SMeX application

of data was done with replication technologies on these materialised views supported by the database management system Oracle. No further transformation of data in terms of data curation was required. For this case study, the quality of data relies on quality assurance in both of the underlying schemata.

The SMeX system allows the retrieval of (1) markers and mapping positions for given ESTs and their homologues in the same and closely related species, (2) ESTs and their homologues for specific regions along the genetic map, (3) BLAST searches of user-specified sequences against all ESTs stored in the mart. An example is the search of related wheat markers for a given set of barley ESTs or vice versa by SMeX. The web-based application is integrated into the Plant Bioinformatics Portal [35].

3.2 Diversity Studies Toolkit (DiSTo)

Lolium ssp. species are grown worldwide and are economically important because of their use as forage, meadow or lawn. During the Genebank fusion process between IPK Gatersleben and BAZ Braunschweig (2002–2006), the complete German collection of *Lolium* was moved to Gatersleben.

Since the collection comprises 3,187 accessions and has to be propagated by relatively high efforts, the necessity to take a closer look at its structure and diversity has emerged. Therefore, SNP markers for the analysis of the *Lolium* accessions have been developed (Fig. 4). The fingerprinting was done by using the pyrosequencing technology and applying the allele quantification method. Additionally, a phenotyping process was performed for all economically important traits. A part of the obtained data was stored in existing databases but especially for the *Lolium* SNP markers, allele frequencies and further experimental data the creation of a new database was necessary. This Pyrosequencing Database (PSQDB) [36], together with the IPK

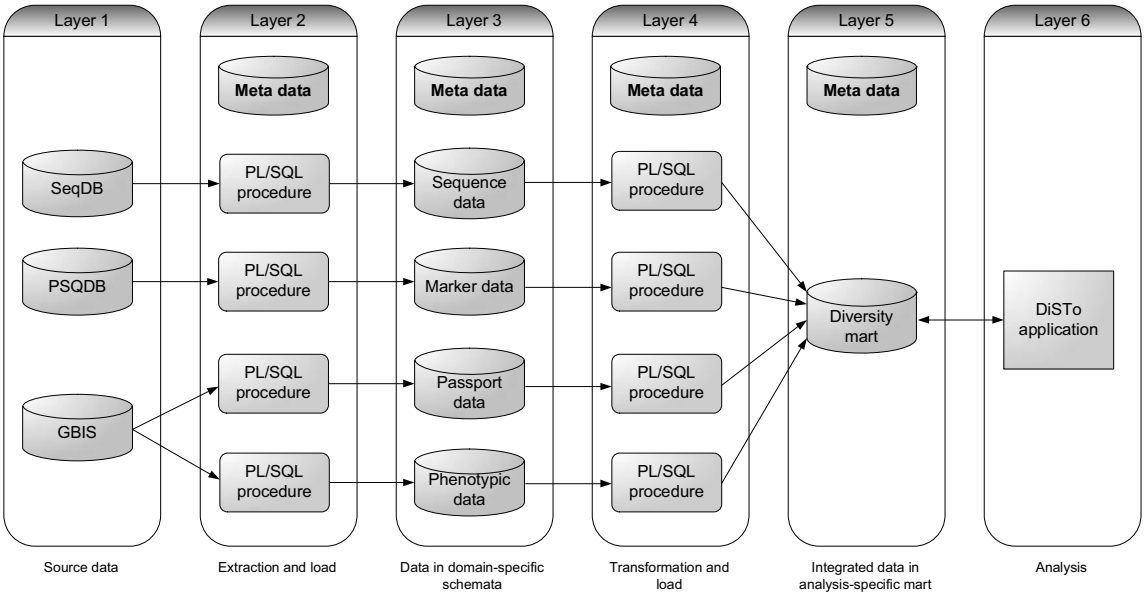

Figure 4: Implementation of our architecture for the DiSTo application

Sequence Database, passport data from the IPK Genebank Information System (GBIS) [37], and phenotypic data collected in proprietary files outside the IPK are the operational sources for comprehensive diversity studies of the whole *Lolium* collection.

The data integration was similar to the integration workflow covering the SMeX case study. First, the transformation and load processes were established for transferring data from the operational sources to the appropriate data pool schemata. Afterwards, necessary data for diversity studies was combined in an application-specific data mart. Materialised views were also used here and no data curation was applied at this application level.

Based on the "diversity" data mart, a tool for different data queries, comprehensive and multivariate data analysis, and results visualisation was developed. This Diversity Studies Toolkit (DiSTo) provides the opportunity that data obtained in the *Lolium* studies are more readily accessible to a broad range of potential end-users. Similar projects with different plant species collections will then get results much faster.

3.3 Barley Association Studies (BAS)

Malting quality is one of the most important traits in barley. In the GABI-MALT research network [38], subproject 4 is located at the IPK and focuses on (1) the identification of putative candidate genes involved in the malting process by using structural genomics approaches, (2) analysis of their allelic diversity, (3) association of haplotypes and SNP-patterns with malting parameters of barley. The allelic diversity of selected candidate genes is investigated by SNP- and INDEL-analysis within a representative genotype panel. The obtained results will increase our understanding of the genetic impact on metabolic pathways underlying malting quality.

One important part is the development of suitable SNP-markers out of selected candidate genes with impact on malting traits. Therefore, 48 genes encoding enzymes which are known to be important during germination and the malting process from the literature, e. g. [39, 40], and 16 interesting candidate genes derived from expression analysis were chosen to detect differ-

ent SNP- and haplotype patterns. Sequence analyses were performed based on PCR-amplified genomic fragments of 300 to 600 bp. SNP-fingerprinting of 1,000 barley accessions was performed by pyrosequencing. These accessions comprise modern European varieties, a representative part of the Barley Core Collection [41] and actual breeding material. Furthermore, INDEL-analysis in this large set of genotypes and development of CAPS markers was done. SNP data is stored in a specific schema in our data pool.

Apart of this, phenotypic data of approximately 120 malting and brewing parameters obtained from 250 barley cultivars, which are also characterised through molecular markers, were collected in the past 20 years at different locations in Germany. It is stored in publicly available sources such as Bundessortenamt (BSA) and statistical year books of the German Brewing Society (BGJB). This data is handled in another specific schema of our data pool. Additionally, data from actual field and micromalting experiments of 64 cultivars in 2004, 2005 and 2006 were delivered from subproject 3 of GABI-MALT by the Bayerische Landesanstalt für Landwirtschaft (LfL) in Freising. Up to now, approximately 80,000 phenotypic data points are available for association studies to a large extent. The data stock will be actualised and extended by adding more data of micromalting experiments from breeder strains and other available sources.

In order to get information about the genetic diversity of the used barley cultivars, their population structure was revealed by randomly distributed SSR markers over the whole barley genome. Therefore, the software Structure [42] was used to obtain a kinship matrix.

The developed architecture (see Fig. 5), allows association studies to a large extent by combining genotypic and phenotypic information and pedigree data with respect to population structure. Currently, association studies itself are carried out using the software TASSEL [43] assuming the General Linear Model (GLM) [44] or the Mixed Linear Model (MLM) [45], respectively. The genetic diversity was revealed between all barley varieties analysed here. Association studies show how genetic changes in certain candidate genes possess an impact on malting parameters [46]. This is of high importance not only for breeders, but also for the brewing industry.

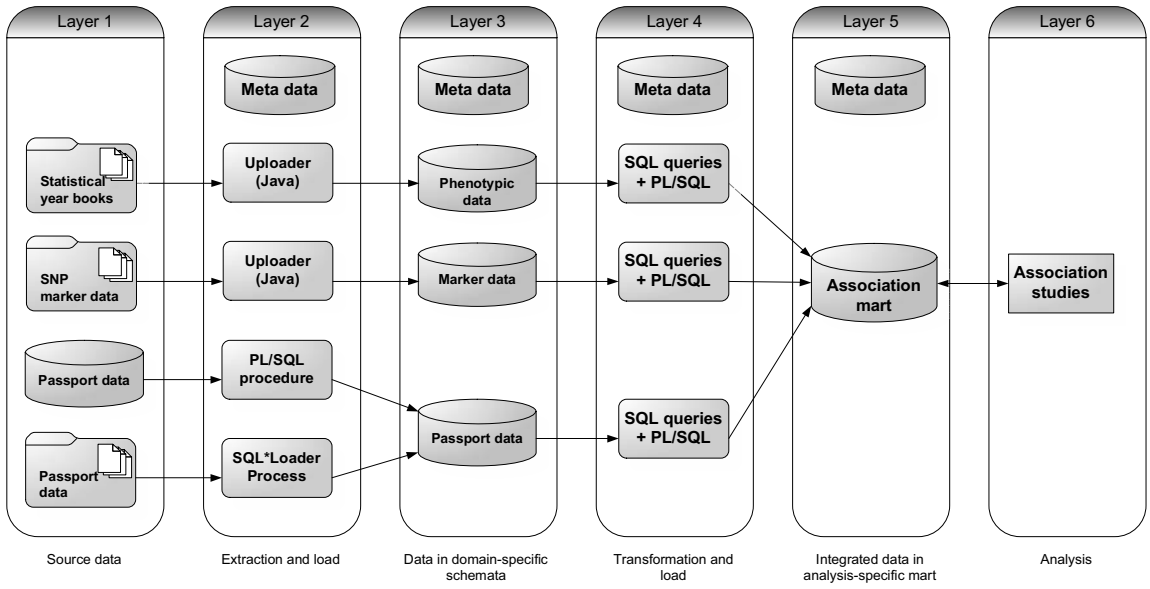

Figure 5: Implementation of our architecture for the barley association studies

4 Discussion

In crop plant bioinformatics data warehousing becomes more and more important. Therefore, data integration is a main task. In this paper we described challenges of crop plant bioinformatics concerning data integration and suggested possible solutions for linking records from different data domains. A generic process flow consisting of six layers for the flexible integration and analysis of crop plant data is proposed here. In contrast to other approaches, which often focus on one domain of data or one class of application only, or try to use a generalised data warehouse schema, this approach focuses on integration and data analysis across different domains.

The integration and analysis platform Gene-EYe [47] enables e. g. the flexible integration of data, but focuses on genomic data, esp. sequence data. Another approach, Columba [48], integrates and analyses protein structure data from a limited set of sources, whereas GeWare [49] uses OLAP for the analysis of microarray-based genexpression data and annotation data. Although this is a very promising approach, in our opinion OLAP is suitable for standardised data generated by comparable methods, e. g. genexpression and annotation data. Atlas [50] focuses on molecular data and enables the joined analysis of different data domains. The integrated data is stored in separate schemata, one for each domain, which are fixed and thus hamper flexible extensions in case of new requirements. BioWarehouse [51] is an approach for integration and analysis of pathway oriented data by using a fixed database schema, which makes extensions difficult. Finally, BioMart (formerly EnsMart) [52] is a framework for generating star schemata and user interfaces for searching and filtering data, where data should exist in relational tables, which is often problematic in plant bioinformatics.

Here, we develop an approach, which integrates data into a data pool with domain-specific schemata and then uses this data pool to fill application-specific marts, not a universal data warehouse schema. Three case studies applying our approach are presented here. The decision support system SMeX allows the retrieval of sequence mapping positions, whereas the Diversity Studies Toolkit DiSTo enables the user to examine the genetic diversity of plant collections. The third case study allows association studies of barley cultivars to a large extent in order to obtain genotype–phenotype correlations.

5 Acknowledgements

We thank Ralf Hofestädt, Matthias Lange, Thomas Rutkowski and Thomas Thiel for valuable discussions, and the German Federal Ministry for Education and Research for financial support.

References

[1] D.S. Roos. Computational Biology: Bioinformatics – Trying to Swim in a Sea of Data. *Science*, 291(5507):1260–1261, 2001.

[2] E. Grafahrend-Belau, S. Weise, D. Koschützki, U. Scholz, B.H. Junker, and F. Schreiber. MetaCrop: a detailed database of crop plant metabolism. *Nucleic Acids Research Advance Access published on October 11, 2007, doi:10.1093/nar/gkm835*.

[3] G. Wiederhold. Foreword: Intelligent integration of information. In G. Wiederhold, editor, *Intelligent Integration of Information*. Kluwer Academic Publishers, Boston, 1996.

[4] W.H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 4th edition, 2005.

[5] J. Augen. Information technology to the rescue! *Nature Biotechnology*, 19(6s):BE39–BE40, 2001.

[6] C. Schönbach, P. Kowalski-Saunders, and V. Brusic. Data warehousing in molecular biology. *Briefings in Bioinformatics*, 1(2):190–198, 2000.

[7] 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, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novre, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[8] P. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, A. Shojatalab, M.and Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. Aronow, A. Robinson, D. Bassett, C. Stoeckert, and A. Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9):research0046.1–research0046.9, 2002.

[9] A. Alercia, S. Diulgheroff, and T. Metz. FAO/IPGRI Multi-Crop Passport Descriptors (MCPD). FAO (Food and Agriculture Organization of the United Nations) - IPGRI (International Plant Genetic Resources Institute), 2001.

[10] P. Bork and A. Bairoch. Go hunting in sequence databases but watch out for the traps. *Trends in Genetic*, 12(10):425–427, 1996.

[11] H. Müller, F. Naumann, and J.-C. Freytag. Data quality in genome databases. In *Proceedings of the Conference on Information Quality (IQ 03)*, Boston, 2003.

[12] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[13] K. Ilic, E.A. Kellogg, P. Jaiswal, F. Zapata, P.F. Stevens, L.P. Vincent, S. Avraham, L. Reiser, A. Pujar, M.M. Sachs, N.T. Whitman, S.R. McCouch, M.L. Schaeffer, D.H. Ware, L.D. Stein, and S.Y. Rhee. The Plant Structure Ontology, a Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. *Plant Physiology*, 143(2):587–599, 2007.

[14] IPGRI. Descriptors for Barley (*Hordeum vulgare* L.). International Plant Genetic Resources Institute, Rome, Italy, 1994.

[15] Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, 2(5):351–356, 2005.

[16] R.A. Irizarry, D. Warren, F. Spencer, I.F. Kim, S. Biswal, B.C. Frank, E. Gabrielson, J.G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S.C. Hilmer, E. Hoffman, A.E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S.Q. Ye, and W. Yu. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5):345–350, 2005.

[17] H. Müller, M. Weis, J. Bleiholder, and U. Leser. Erkennen und Bereinigen von Datenfehlern in naturwissenschaftlichen Daten. *Datenbank-Spektrum*, 15:26–35, 2005.

[18] M. Lange, A. Himmelbach, P. Schweizer, and U. Scholz. Data Linkage Graph: computation, querying and knowledge discovery of life science database networks. *Journal of Integrative Bioinformatics*, 4(3):e68, 2007.

[19] V. Levenshtein. Binary Codes of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[20] R.C. Russel. Index. US patent 1,261,167, 1918. pp. 1–4, [http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=1,261,167] As of 2007-11-10.

[21] A. Lait and B. Randell. An assessment of name matching algorithms. Technical Report No. 550, Department of Computing Science, University of Newcastle upon Tyne, 1996.

[22] P. Weiner. Linear pattern matching algorithms. In *Proceedings of the 14th IEEE Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.

[23] T. Smith and M. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[24] Oracle Corporation. Oracle Warehouse Builder. [http://www.oracle.com/technology/products/warehouse/index.html] As of 2007-11-10.

[25] T. Etzold, A. Ulyanov, and P. Argos. SRS: information retrieval system for molecular biology data banks. *Methods in Enzymology*, 266:114–128, 1996.

[26] C. Friedman, G. Hripcsak, S. Johnson, J. Cimino, and P. Clayton. A generalized relational schema for an integrated clinical patient database. In *Proc 14th Symp Comput Appl Med Care*, pages 335–339, 1990.

[27] P.M. Nadkarni and C. Brandt. Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database. *Journal of the American Medical Informatics Association*, 5(6):511–527, 1998.

[28] P.H. Winston. *Artificial Intelligence*. Addison-Wesley, Reading, Mass, 2nd edition, 1984.

[29] C. Künne, M. Lange, T. Funke, H. Miehe, T. Thiel, I. Grosse, and U. Scholz. CR-EST: a resource for crop ESTs. *Nucleic Acids Research*, 33(suppl_1):D619–D621, 2005.

[30] E.F. Codd, S.B. Codd, and C.T. Salley. Providing OLAP to User-Analysts: An IT Mandate. White paper, E.F. Codd & Associates, 1993.

[31] N. Stein, M. Prasad, U. Scholz, T. Thiel, H. Zhang, M. Wolf, R. Kota, R.K. Varshney, D. Perovic, I. Grosse, and A. Graner. A 1000 loci transcript map reveals insight into barley genome evolution. *Theoretical and Applied Genetics*, 114(5):823–839, 2007.

[32] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. GenBank. *Nucleic Acids Research*, 34(suppl_1):D16–D20, 2006.

[33] Y. Lee, J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung, and J. Quackenbush. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Research*, 33(suppl_1):D71–D74, 2005.

[34] G.R. Lazo, S. Chao, D.D. Hummel, H. Edwards, C.C. Crossman, N. Lui, D.E. Matthews, V.L. Carollo, D.L. Hane, F.M. You, G.E. Butler, R.E. Miller, T.J. Close, J.H. Peng, N.L.V. Lapitan, J.P. Gustafson, L.L. Qi, B. Echalier, B.S. Gill, M. Dilbirligi, H.S. Randhawa, K.S. Gill, R.A. Greene, M.E. Sorrells, E.D. Akhunov, J. Dvorak, A.M. Linkiewicz, J. Dubcovsky, K.G. Hossain, V. Kalavacharla, S.F. Kianian, A.A. Mahmoud, Miftahudin, X.-F. Ma, E.J. Conley, J.A. Anderson, M.S. Pathan, H.T. Nguyen, P.E. McGuire, C.O. Qualset, and O.D. Anderson. Development of an Expressed Sequence Tag (EST) Resource for Wheat (*Triticum aestivum* L.): EST Generation, Unigene Analysis, Probe Selection and Bioinformatics for a 16,000-Locus Bin-Delineated Map. *Genetics*, 168(2):585–593, 2004.

[35] A. Stephanik, H. Bachmann, T. Funke, C. Kuenne, E. Langer, T. Thiel, S. Weise, and I. Grosse. Das Plant Bioinformatics Portal. In *Ausgewählte Vorträge aus GPZ-Arbeitsgemeinschaften*, volume 70 of *Vorträge für Pflanzenzüchtung*, pages 81–83, Göttingen, 2006. Gesellschaft für Pflanzenzüchtung e. V. (GPZ).

[36] Pyrosequencing Database (PSQDB). Leibniz Institute of Plant Genetics and Crop Plant Research / Bioinformatics Centre Gatersleben-Halle. [http://bic-gh.de/portal/page/portal/PG_BICGH/P_BICGH/P_BICGH_RESOURCES/PSQDB] As of 2007-11-10.

[37] Genebank Information System (GBIS). Leibniz Institute of Plant Genetics and Crop Plant Research / Bioinformatics Centre Gatersleben-Halle. [http://gbis.ipk-gatersleben.de] As of 2007-11-10.

[38] I.E. Matthies, K. Foerster, and M.S. Röder. GABI-MALT: An integrated approach to the genetic and functional dissection of malting quality in barley Subproject 4: SNP-detection and haplotype analysis in candidate genes for malting. In *GABI-PROGRESS-REPORT*, pages 30–32. 2007.

[39] G.P. Fox, J.F. Panozzo, C.D. Li, R.C.M. Lance, P.A. Inkerman, and R.J. Henry. Molecular basis of barley quality. *Australian Journal of Agricultural Research*, 54(12):1081–1101, 2003.

[40] P.M. Hayes, A. Castro, L. Marquez-Cedillo, A. Corey, C. Henson, B.L. Jones, J. Kling, D. Mather, I. Matus, C. Rossi, and K. Sato. Genetic diversity for quantitatively inherited agronomic and malting quality traits. In R. von Bothmer, T.J.L. van Hintum, H. Knüpffer,

and K. Sato, editors, *Diversity in Barley (Hordeum vulgare)*, volume 7 of *Developments in Plant Genetics and Breeding*, chapter 10, pages 201–226. Elsevier, 2003.

[41] H. Knüpffer and T.J.L. van Hintum. The barley core collection - an international effort. In T. Hodgkin, A.H.D. Brown, T.J.L. van Hintum, and E.A.V. Morales, editors, *Core Collections of Plant Genetic Resources*, pages 171–178. Wiley, Chichester, 1995.

[42] D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7(4):574–578, 2007.

[43] E.S. Buckler, P. Bradbury, and D. Kroon. TASSEL: Trait Association, Evolution, and Linkage Analysis software package. Buckler Lab for Maize Genetics and Diversity, 2007. Version 2.0.1 [http://www.maizegenetics.net] As of 2007-11-10.

[44] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, 135(3):370–384, 1972.

[45] J. Yu, G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2006.

[46] I.E. Matthies and M.S. Röder. Haplotypendiversität von Kandidatengenen mit Einfluss auf die Malzqualität in Gerste. In *Bericht über die 57. Tagung "Pflanzenzüchtung und Genomanalyse" der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs, HBLFA Raumberg – Gumpenstein, 21.–23. November 2006*, pages 61–63, 2006.

[47] P. Rieger, S. Heymann, and H. Müller. Datenbankgestützte Wissensakquisition in den Lebenswissenschaften. *Datenbank-Spektrum*, 10:14–21, 2004.

[48] S. Trissl, L. Rother, H. Müller, T. Steinke, I. Koch, R. Preissner, C. Frömmel, and U. Leser. Columba: An Integrated Database of Proteins, Structures, and Annotations. *BMC Bioinformatics*, 6:e81, 2005.

[49] E. Rahm, T. Kirsten, and J. Lange. The GeWare data warehouse platform for the analysis of molecular-biological and clinical data. *Journal of Integrative Bioinformatics*, 4(1):e47, 2007.

[50] S.P. Shah, Y. Huang, T. Xu, M.M.S. Yuen, J. Ling, and B.F.F. Ouellette. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*, 6:e34, 2005.

[51] T.J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D.W.J. Stringer-Calvert, J.D. Tenenbaum, and P.D. Karp. Biowarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7:e170, 2006.

[52] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Research*, 14(1):160–169, 2004.