DBE2 – Management of experimental data for the VANTED system

Hendrik Mehlhorn[1*], Falk Schreiber[1,2]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstr. 3,
06466 Gatersleben, Germany

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

Summary

DBE2 is an information system for the management of biological experiment data from different data domains in a unified and simple way. It provides persistent data storage, worldwide accessibility of the data and the opportunity to load, save, modify, and annotate the data. It is seamlessly integrated in the VANTED system as an add-on, thereby extending the VANTED platform towards data management. DBE2 also utilizes controlled vocabulary from the Ontology Lookup Service to allow the management of terms such as substance names, species names, and measurement units, aiming at an eased data integration.

1 Introduction

High throughput phenotyping facilities and modern wet lab techniques such as GC/MS, multidimensional protein gels, and microarrays produce an continuously increasing amount of biological data sets. Each data set comprises a set of biological measurements as well as annotation data.

The data type of biological measurements is in many cases a simple decimal number. A decimal number may represent, for example, the concentration of a metabolite, the relative content of a messenger RNA, or the propotion of the expression levels of an enzyme under various conditions. Ordered sets of decimal numbers also represent one dimensional gradients, such as the concentration of a metabolite in a cell over time. Upcoming facilities enable the high throughput phenotyping of, for instance, plants, which yields a hugh amount of two dimensional images. Other techniques such as NMR or CT produce three dimensional volume data. The magnitude of biological measurement data necessitates data management systems in order to enable appropriate data analysis and data exchange.

Biological measurements arise in the context of certain experiment conditions and represent properties of specific biological entities (e.g. the concentration of a metabolite). This is reflected in the annotation data. Experiment conditions such as the availability of nutrients, water, and light, the time point of the measurement, the underlying genotype, and tissue constitute notable annotation data in praxis. Annotation data also comprises the names of measured biological entities as well as the unit of biological measurements. Standards such as PEDRo,

[*]To whom correspondence should be addressed. Email: mehlhorn@ipk-gatersleben.de

doi:10.2390/biecoll-jib-2011-162

MIAME, and ArMet [1, 2, 3] provide all necessary fields to reproduce the underlying biological experiment. These formats have been proposed for the standardization of annotation data, but the input demanded from the user is exhaustive which often prevents their usage. Fortunately bioinformatic analysis techniques such as flux balance analysis, correlation analysis, and network mapping, need just specific input parameters. The reduced specification demand of these techniques is an advantage which eases fast and appropriate data set explorations.

A complex biological problem often necessitates the examination of data sets from different data domains. Most existing databases do mainly only cover single data domains and are not uniformly addressable, which results in the need of data integration systems such as ONDEX or BridgeDB [4, 5].

The distribution of data sets also causes the problem of data annotation inconsistencies. For metabolite terms, protein names, or measurement units various formats or synonym relationships can be found. Ontologies are intended to structure the knowledge of an area of interest. Life science ontologies such as CHEBI (e.g. chemical compounds), NEWT (species taxonomy), and Gene Ontology (e.g. protein functions) are utilizable to overcome data annotation inconsistencies by strictly annotating the biological measurements using ontology terms [6].

The aim of the DBE2 information system (**D**atabase for **B**iological **E**xperiments **2**) is to manage biological data sets from different data domains in a unified and simple way. Only a small amount of annotation data is required, which is nevertheless appropriate for recent analysis techniques. The management and storage of data sets is consistent despite of the domain(s) the data arises from. The focus is on the integration of multimodal biological measurements of different data types such as zero dimensional decimal numbers, one dimensional gradients, two dimensional images, three dimensional volumes, and even biological networks being interesting in the context of the biological measurements.

The DBE2 information system is based on the DBE information system [7]. The DBE information system proved its usefulness for biologists by an easy usage, the permanent availability, and it's focus on the integration of data sets stored in the DBE database in a persistent and structured way.

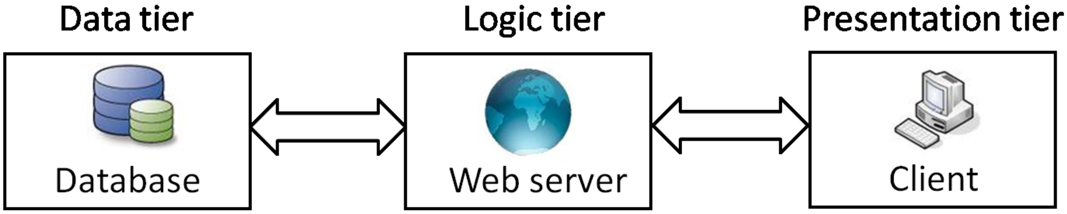


Figure 1: The *three-tier architecture* is being instanciated by the DBE2 information system. The *presentation tier* (the *DBE2 client*) calls the *logic tier* to download, upload, and edit biological data. The *logic tier* (the *DBE2 servlet*) employs the *data tier* (the *DBE2 database* and a file storage system) for the storage of experiment data and binary files. This happens using an underlying user management to manipulate or transfer data by defined queries.

There are several new developments in DBE2. The integration of further data domains and data types now enables a bigger area of application. Recent software engineering techniques helped restructuring the system, yielding a three-tier architecture (see Figure 1). The *DBE2 servlet* was introduced to implement all data accesses to the extended *DBE2 database* by queries and to facilitate the worldwide availability of the data. Big Files in biological data sets are stored in a *hierarchical storage management* by the *DBE2 servlet* to maintain the database performance.

The *DBE2 client* in the shape of a VANTED [8] add-on now provides a graphical user interface to the DBE2 information system. Each data set access is controlled by defined servlet queries including an user account management. For the convenient integration of further clients a library is being supported which implements all *DBE2 servlet* calls in a functional way.

In addition the utilization of controlled vocabulary from the *Ontology Lookup Service* [9] raises the quality of annotation data such as species names, substance names, and measurement units. An adaption of certain ontology structures enables the organization of the underlying data sets in a reasonable and intuitive way.

This paper is organized as follows. The (1) *Introdution* conveys the area and the background of the DBE2 information system. Section (2) *DBE2 schema* discusses the representation of data sets in the *DBE2 database* and the whole system. A servlet enables the continuous and worldwide access to data sets, which is being introduced in Section (3) *DBE2 servlet*. The *DBE2 client* is designed as an graphical user interface to the system which is presented in Section (4) *DBE2 client*. This paper closes with a (5) *Discussion* to resume and discuss the presented content.

2 DBE2 schema

The DBE2 information system is designed to handle biological measurements of diverse data types and from various data domains such as metabolomics, proteomics, and phenomics. This happens in the context of the adjacent annotation data. A tree data structure of four levels represents data sets in an adequate and flexible way (see Figure 2).

A data set of biological measurements with it's annotation data is called an *experiment*. It contains meta information such as the experiment name, the coordinator, and the start date of the project. An *experiment* branches into a set of *condition*s as the experiment context of the measured data. Each *condition* represents the species, genotype, and variety as well as the treatment of the examined biological being. Each *condition* branches into a set of *sample*s, which specifiy the measurings of the underlying measured data in space and time. Measured data is allowed in the shape of (i) simple decimal numbers, (ii) pictures, and (iii) volumes which correspond to data of zero, two, and three dimensions. It is also possible to represent biological (iv) networks as well as one dimensional (v) gradients (by a set of ordered (i) decimal numbers). An abstract example is shown in Figure 2.

The resulting hierarchical tree structure is being implemented in the shape of relational tables in the *DBE2 database* as the *data tier* for a persistent and structured data storage as well as by a XML document schema for data exchange tasks. This enables dealing with experiment data of different types from various domains in a unified way using a corporate data structure.

The *DBE2 database* schema is implemented in an Oracle database (version 11g) and is shown in Figure 3. The database schema consists of four conceptual modules, namely the (i) *User management* module, the (ii) *Experiment data* module, the (iii) *Basis data* module, and the (iv) *Supplementary material* module. The (i) *User management* module provides a basic user right handling. Users need to possess a DBE2 account to store experiments in the *DBE2 database*. User accounts get organized via user groups. For every stored experiment a user group is defined with users having the right to access it. The experiment data is being stored in the (ii)

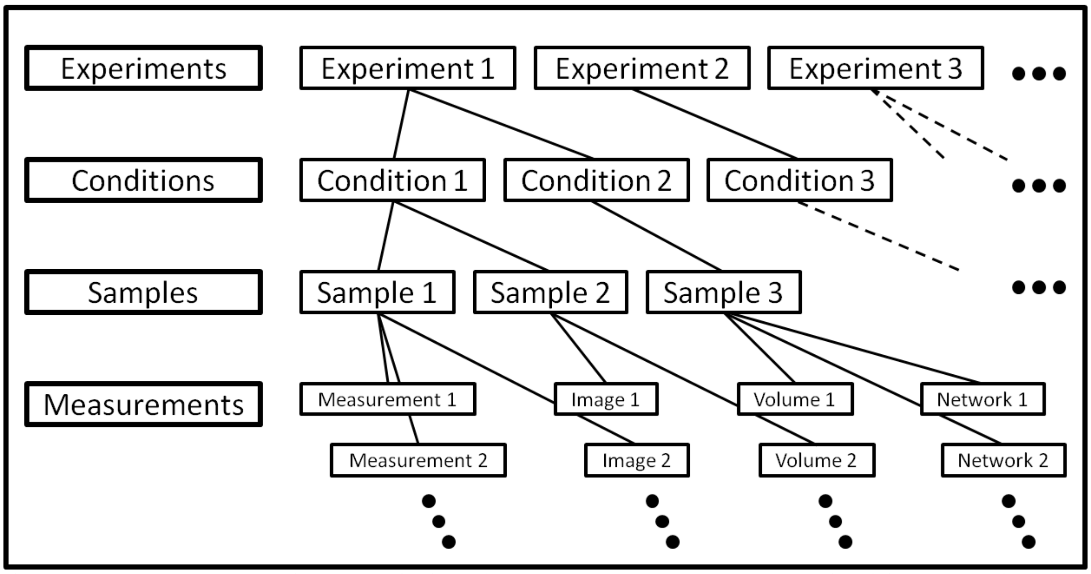


Figure 2: An *experiment* is a set of biological measurements together with it's annotation data and is represented by a tree structure in four levels. It branches into a set of *condition*s which represent the treatment and the genotype of the examined species. Each *condition* forks into a set of *sample*s, which provide information about the underlying measured data. Measured data is allowed to exhibit the shape of decimal numbers, images, volumes, and biological networks.

Experiment data module. This is done one to one according to the described hierarchical tree structure. A subset of the annotation data is being managed in the (iii) *Basis data* module. This module is designed to organize annotation data through controlled vocabulary of various areas such as species names, unit terms, and substance names. In addition the (iv) *Supplementary material* module enables to associate arbitrary supplementing files to any level of the *Experiment data* module as supplementary material. An important supplementing file could be a SDS gel image as the source of measured protein expression levels for instance.

Users of DBE are invited to supply annotation data to biological measurements according to the introduced hierarchical tree structure. Thus the support of information about the experiment in general, the species, the examined tissue, the time points of the biological measurements, and the name of the measured entities is necessary. This is sufficient for tasks such as the statistical comparison of data sets from varios differential treated breeding lines.

3 DBE2 servlet

Every request of the *DBE2 client* to the *DBE2 database* is implemented by the *DBE2 servlet*. The *DBE2 servlet* defines a set of queries, which builts an application programming interface (API). The encapsulation of *DBE2 database* transactions via a servlet as the *logic tier* assures worldwide data access and safe database manipulations. The *DBE2 servlet* is implemented as an Java HTTP servlet, which enables a dynamic treatment of all queries on the web server.

The *DBE2 servlet* implements the observance of the user right management. Every request associated to confidential user data sets contains parameters to authorize it. The handling of unexpected cases and errors includes exception reports and *DBE2 database* rollbacks. In this way the *DBE2 servlet* assures a consistent and safe data storage.

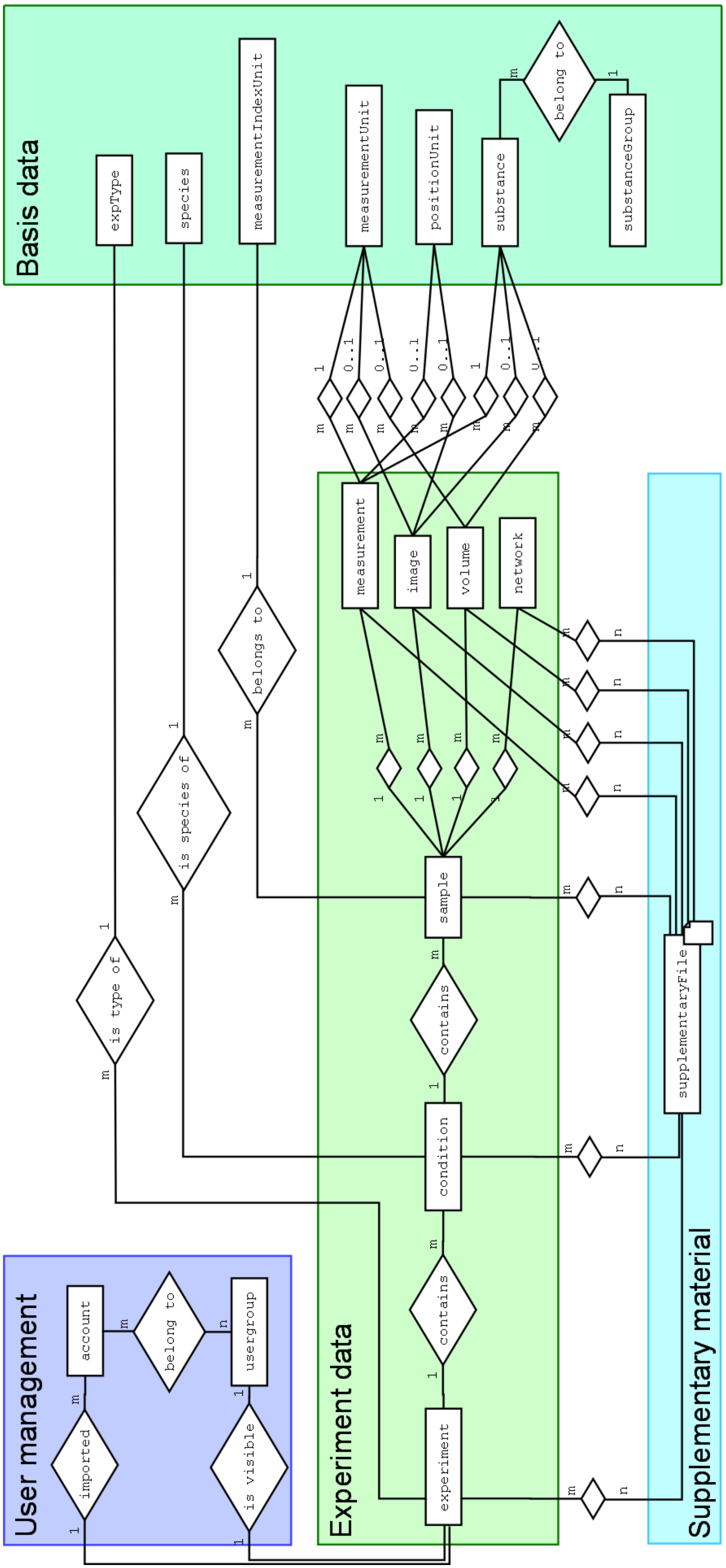


Figure 3: The *DBE2 database* schema as entity relationship diagram. The database schema comprises four modules: (i) *User management* module for user right handling, (ii) *Experiment data* module for experiment data storage, (iii) *Basis data* module for controlled vocabulary, and a (iv) *Supplementary material* module for the association of arbitrary files with entries in the *Experiment data* module.

The queries provided by the *DBE2 servlet* enable an functional access to the database. Every client functionality regarding the *DBE2 database* is being realized by a certain set of *DBE2 servlet* queries. This prevents the client developer from communicating directly with the database via SQL queries, yielding a rather clean and safe development style.

The transfer of data from the *DBE2 client* to the *DBE2 servlet* and back works through streams, which afford the transfer of arbitrary sized data. For database performance reasons, binary files are stored using a *hierarchical storage management* (HSM), which realizes a compromise between data transfer performance and storage costs. This is implemented by the unpublished BFiler package in a transactional and simple way. Thus BFILEs (an Oracle SQL datatype) are stored in the *DBE2 database* as references to the physical file in the HSM instead of BLOBs (another Oracle SQL datatype which represents the whole file data).

4 DBE2 client

The *DBE2 client* supports a graphical user interface (GUI) to the DBE2 information system and represents an instance of the *presentation tier*. It executes all user actions regarding the *DBE2 database* via a set of *DBE2 servlet* queries. The *DBE2 client* is designed as a VANTED add-on, which smoothly enables the usage of analysis and visualization techniques of the VANTED system [8]. It is also possible to implement further clients for the usage of the system. Therefore a java library called the *DBE2 servlet client* is being supported to communicate with the *DBE2 servlet* in a functional and easy way.

The *DBE2 client* provides a user-friendly and easy way to upload, edit, and download *experiments* and to edit the annotation data of the experiment in the local XML format representation. In addition, the DBE2 information system administrator is allowed to manage terms in the *DBE2 database Basis data* module and to manage user rights by changing entries in the *DBE2 database User management* module. There is also a special public account which may be used by any user. This account enables users without an DBE2 account to download experiments which were explicitly approved for public access. See Figure 4 for an overview of all impotant *DBE2 client* use cases.

In the case of forbidden queries or unexpected cases the *DBE2 servlet* throws exceptions processed by the *DBE2 client*. In the case of known errors such as the injury of database table constraints, this happens either in way of offering an alternative proceeding or by descriptive messages. An example is that the name of *experiment*s has to be unique.

The usage of a controlled vocabulary for the standardization of annotation data supports easy mapping of data sets onto each other and onto biological networks. In case of the DBE2 information system the controlled vocabulary is being represented by the *Basis data* module in the *DBE2 database*. Whenever an user attempts to upload an *experiment* to the *DBE2 database*, the annotation data of the *experiment* has to be covered by the *Basis data* module. Uncovered annotation data has to be synchronized. Terms in the annotation data of the *experiment* which are missing in the *Basis data* module have to be added to expand the basis data pool or renamed to match the basis data pool. The *DBE2 client* offers the possibility to standardize annotation data by utilizing the *Ontology Lookup Service* (OLS). The OLS (*http://www.ebi.ac.uk/ontology-lookup/*) is a compendium of more than 70 life science ontologies retrievable through an unified interface. Users are able to search various ontologies for terms in the annotation data of experi-

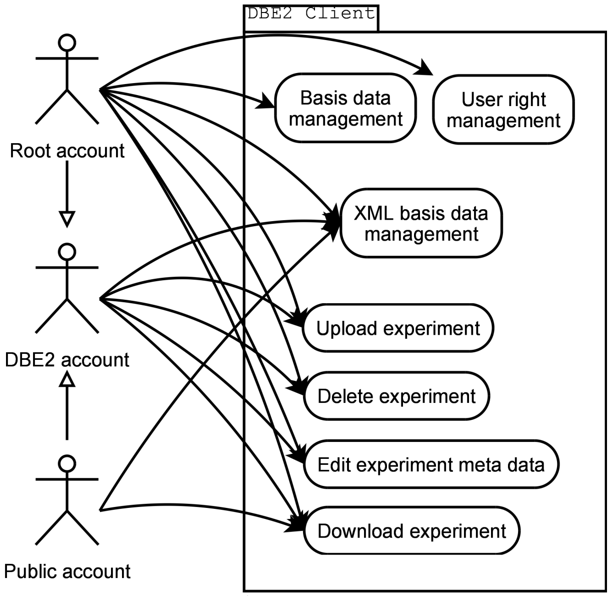


Figure 4: Use cases diagram of the *DBE2 client*. There are three types of accounts. The special (i) public account may used by everybody and is allowed to download particular experiments and to edit the basis data in the local XML representation of experiments. Users with a (ii) regular DBE2 account may upload, download, and edit experiments in addition. The DBE2 administrator uses the (iii) root account and is thus allowed to access all features including the management of *DBE2 database* basis data and the appointment of user rights.

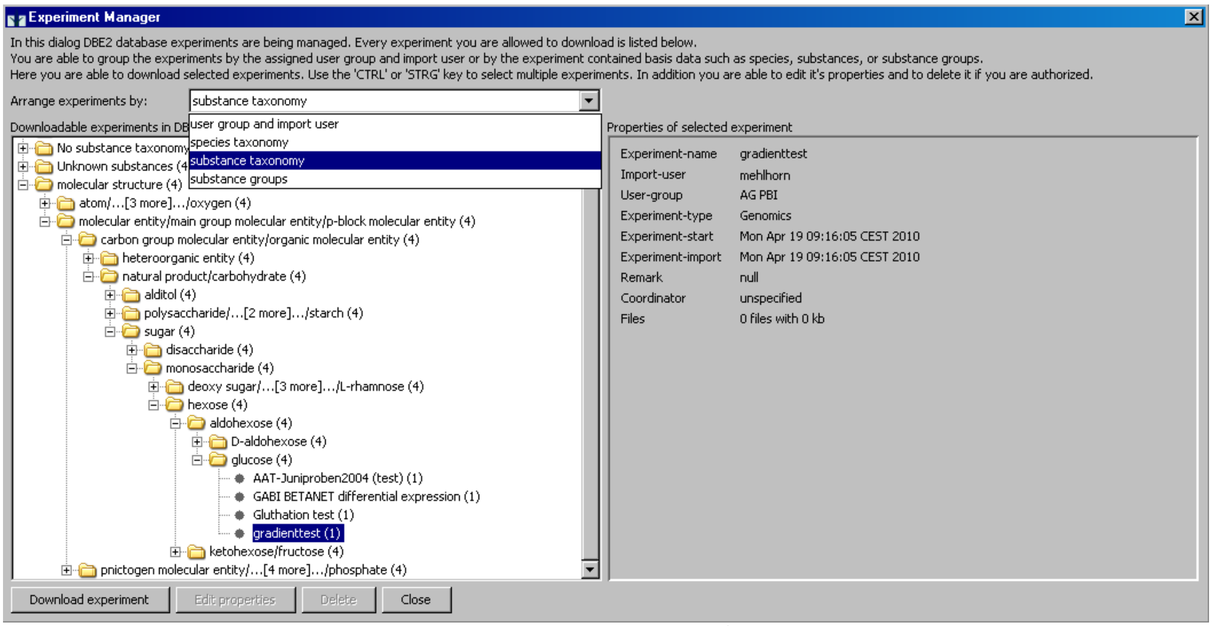
ments. Species names and the according taxonomy tree are accessible from the ontology *NEWT UniProt Taxonomy Database* (short name: NEWT) [10]. Names of chemical compounds and the corresponding compound taxonomy are accessible from the ontology *Chemical Entities of Biological Interest* (short name: CHEBI) [6]. With the help of these ontologies, a part of the *DBE2 client* named *experiment manager* (see Figure 5), is able to represent the set of accessible *experiment*s in a hierarchical way according to their annotation data.

Since the *DBE2 client* is integrated seamlessly into VANTED, the DBE2 information system user is instantly able to use the functionalities of the VANTED system. These include the mapping of *experiment*s on biological networks, the arrangement of *experiment*s according to KEGG pathway hierarchies, and the corresponding visualization via a graph.

5 Discussion

In this paper the functionalities and the *three-tier architecture* of the DBE2 information system were presented. The DBE2 information system is designed for the management of biological measurements of various domains and types in an unified and easy way. The *DBE2 database* represents the *data tier* and stores biological data in a structured and persistent way. The *DBE2 servlet* represents the *logic tier* and realizes all data access and data manipulation in a functional and safe way. The *DBE2 client* represents the *presentation tier* and provides an easy to use GUI to the DBE2 information system.

Ontology support from the OLS aids the standardization of annotation data. The according controlled vocabulary eases the integrated analysis of biological measurements from various data sets. Term taxonomies supported by life science ontologies enable a logically structured



Experiment Manager

In this dialog DBE2 database experiments are being managed. Every experiment you are allowed to download is listed below.
You are able to group the experiments by the assigned user group and import user or by the experiment contained basis data such as species, substances, or substance groups.
Here you are able to download selected experiments. Use the 'CTRL' or 'STRG' key to select multiple experiments. In addition you are able to edit it's properties and to delete it if you are authorized.

Arrange experiments by: substance taxonomy

Downloadable experiments in DB

```
⊞ No substance taxonomy
⊞ Unknown substances (4
⊟ molecular structure (4)
    ⊞ atom/…[3 more]…/oxygen (4)
    ⊟ molecular entity/main group molecular entity/p-block molecular entity (4)
        ⊟ carbon group molecular entity/organic molecular entity (4)
            ⊞ heteroorganic entity (4)
            ⊟ natural product/carbohydrate (4)
                ⊞ alditol (4)
                ⊞ polysaccharide/…[2 more]…/starch (4)
                ⊟ sugar (4)
                    ⊞ disaccharide (4)
                    ⊟ monosaccharide (4)
                        ⊞ deoxy sugar/…[3 more]…/L-rhamnose (4)
                        ⊟ hexose (4)
                            ⊟ aldohexose (4)
                                ⊞ D-aldohexose (4)
                                ⊟ glucose (4)
                                    ● AAT-Juniproben2004 (test) (1)
                                    ● GABI BETANET differential expression (1)
                                    ● Gluthation test (1)
                                    ● gradienttest (1)
                            ⊞ ketohexose/fructose (4)
        ⊞ pnictogen molecular entity/…[4 more]…/phosphate (4)
```

Properties of selected experiment

Experiment-name	gradienttest
Import-user	mehlhorn
User-group	AG PBI
Experiment-type	Genomics
Experiment-start	Mon Apr 19 09:16:05 CEST 2010
Experiment-import	Mon Apr 19 09:16:05 CEST 2010
Remark	null
Coordinator	unspecified
Files	0 files with 0 kb

Figure 5: The *Experiment manager* dialog of the *DBE2 client*. The user is able to arrange the accessible *experiment*s in four ways. The (i) user group arrangement helps to survey user rights to *experiment*s. The (ii) substance taxonomy arrangement (shown), the (iii) species taxonomy arrangement, and the (iv) substance group arrangement enable an *experiment* overview according to the annotation data.

and intuitive survey over a big number of data sets.

Techniques from the seamlessly integrated VANTED system such as data mapping become even more powerful in the course of the consistent usage of ontology supported annotation data. For several use case examples of the VANTED system see [11, 12, 13].

The DBE2 information system is currently in use for three projects with users from the IPK as well as external users. There are 43 registered users and 73 experiments represented in the *DBE2 database*.

Future work concerns the support of additional features such as an extended ontology support for an enlarged data integration potential. In the course of a growing user community and usage of the DBE2 information system the identification and elimination of potential performance bottlenecks will be of great importance.

6 Availability and Requirements

- DBE2 information system: *DBE2 client* and *DBE2 servlet client*
    - DBE2 Web site: *http://www.vanted.org/addons/DBE2/index.html*
    - License: GNU General Public License
    - Programming language: Java version 1.5 or higher
    - Requirements: The *DBE2 client* requires the VANTED program

- VANTED program
  - VANTED web site: *http://www.vanted.org/*
  - License: GNU General Public License
  - Operating system(s): Platform independent
  - Programming language: Java version 1.5 or higher
  - Requirements: Screen resolution of 1024 * 768 or higher, mouse, minimum 512 MB RAM recommended

Acknowledgements

We would like to thank Christian Klukas, who developed the initial DBE information system and contributed to the development of DBE2, Matthias Lange, Steffen Flemming, and Thomas Münch for their developmental support. In addition we would like to thank Hardy Rolletscheck for his valuable feedback on the usability of the *DBE2 client*. We also thank Tobias Czauderna, Hendrik Rohn, Anja Hartmann, Eva Grafahrend-Belau, Astrid Junker, and Matthias Klapperstück for fruitful discussions and testing. We thank Klaus Hippe for assistance in establishing the OLS support. The development of the DBE2 information system was supported in the frame of the project GABI Betanet (BMBF FKZ 0315054B).

References

[1] K. Garwood, T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver, and N. W. Paton. Pedro: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics*, 5(1):68, 2004.

[2] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001.

[3] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, and D. B. Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*, 22(12):1601–1606, 2004.

[4] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, 2006.

[5] M. van Iersel, A. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B. Conklin, and C. Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1):5, 2010.

[6] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350, 2008.

[7] L. Borisjuk, M.R. Hajirezaei, C. Klukas, H. Rolletschek, and F. Schreiber. Integrating data from biological experiments into metabolic networks with the DBE information system. *In Silico Biology*, 5(2):93–102, 2005.

[8] B. Junker, C. Klukas, and F. Schreiber. Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1):109.1–13, 2006.

[9] R. Cote, P. Jones, R. Apweiler, and H. Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):97, 2006.

[10] I. Q. Phan, S. F. Pilbout, W. Fleischmann, and A. Bairoch. NEWT, a new taxonomy portal. *Nucleic Acids Research*, 31(13):3822–3823, 2003.

[11] T. Czauderna, C. Klukas, and F. Schreiber. Editing, validating and translating of SBGN maps. *Bioinformatics*, 26(18):2340–2341, 2010.

[12] E. Grafahrend-Belau, C. Klukas, B. H. Junker, and F. Schreiber. FBA-SimVis: interactive visualization of constraint-based metabolic models. *Bioinformatics*, 25(20):2755–2757, 2009.

[13] C. Klukas and F. Schreiber. Integration of -omics data and networks for biomedical research with VANTED. *Journal of Integrative Bioinformatics*, 7(2):112, 2010.