A Distributed Information System for Managing Phenotyping Mass Data

Florian Schmidt[1], Benjamin Bruns[2], Thomas Bode[1], Hanno Scharr[2], Armin B. Cremers[1]

[1]Universität Bonn
Institut für Informatik III
Römerstraße 164
53117 Bonn
{schmidt2, tb, abc}@iai.uni-bonn.de

[2]Forschungszentrum Jülich GmbH
Institut für Bio- und Geowissenschaften
IBG-2: Pflanzenwissenschaften
52425 Jülich
{b.bruns;h.scharr}@fz-juelich.de

Abstract: On-going automation in plant phenotyping has led to an increasing amount of measurement data, which is often managed by specialized, rarely interconnected systems with custom hard- and software. Experiment and analysis scenarios across different systems and the setup of new systems quickly get expensive and tedious. Therefore, we propose a distributed information system, Phenomis, for managing phenotyping experiments based on Data Spaces. Its service-oriented architecture can be adapted to a wide range of plant phenotyping experiments and appliances, helping to overcome the "phenotyping bottleneck", the mismatch of automated phenotyping capability over analysis capacity.

1 Introduction

Plant phenotyping has been identified as an important field of research for progress in plant breeding and basic plant science [TS09]. Along with ongoing automation many technologies have been developed to increase the throughput of plant screening measurements. This commonly results in specialized, rarely interconnected systems with highly diverse datasets and custom analysis tools [Ja09], from now on referred to as a *measurement system* (MS). For a deep insight in plant performance and dynamics of plant functions information about the full history of plants is needed. *Plant histories* combine all available data about experiments, the life cycle, treatments, measurements and the environment of a plant. Since multiple MS can be used to capture different plant traits, relevant data is usually heterogeneous and distributed across several systems.

We propose a distributed information system, *Phenomis*, for managing phenotyping experiments across several MS, to eliminate the need of manual data integration. Users

can access plant histories at the so-called *Scientist Workplace* (ScientistWP). Compared with other published phenotyping platforms, like PHENOPSIS [Gr05] or GlyPh [Pe12], Phenomis is not limited to certain plant species, specific experimental layouts (i.e. MS) or a fixed list of traits. It aims to provide data integration for a complete phenotyping facility, i.e. the Jülich Plant Phenotyping Center (http://www.fz-juelich.de/ibg/ibg-2/).

2 Distributed Information System for Plant Phenotyping

Phenomis needs to be extensible by new attributes, measurements and treatments, and even new MS at runtime. Since MS are considered to operate autonomously, the effects of changes or system failures should limited. These requirements led to a service-oriented architecture (SOA). Its main components are shown in Fig. 1. Although this approach introduces some complexity by the need for discovery and governance services [Jo07], it allows the (dynamic) integration of MS with arbitrary persistence models, as implementation details and (to an extent) data models are encapsulated.

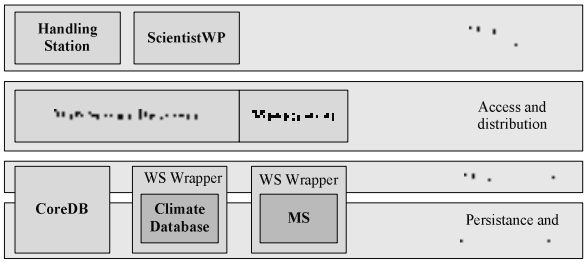

Fig. 1: System architecture of Phenomis

The central plant database (*CoreDB*) is used for spatio-temporal tracking of plants, as well as for storing treatments and measurements in a relational database (PostgreSQL). Its data model is split into a fixed part of relations between e.g. experiments, plants and locations, and an extensible part for managing treatments and measurements. Access is only provided by the CoreDB Web Service, exposing a slim generalized data model for decoupling and defining mappings across components. *Handling stations* are used to automatically collect data in case of manual treatments or measurements (watering, piquing, weighing, etc.) on plants. They are specially designed for human operators and provide efficient and robust user interfaces, utilizing barcode scanners and touch-screens. The collected data is transmitted to the CoreDB.

Measurement Stations are integrated into the system by Web Service wrappers. To limit development effort, all wrappers expose the same external interface. A central *Mapping Service* allows integrated queries across the CoreDB and several MS. It is designed to deal with the temporary unavailability of all registered components.

A specialized, freely available *Climate Database* (called BayEOS-Server, http://www.bayceer.uni-bayreuth.de/edv/de/programme/gru/html.php) is used to manage environment data, which is collected and forwarded by autonomous loggers. A dedicated Web Service provides access to the data; so far five aggregation intervals are supported.

The *ScientistWP* provides access to the distributed data, therefore integrating the data stored in the CoreDB, all MS and the Climate Database. Scientists can visually browse through the experiments and plant histories and export the data for further analysis.

3 Data Integration

The heterogeneity of the measurement data and the support of MS with arbitrary storage systems (RDBMS, file level storage, etc.) makes data integration a major challenge in Phenomis. Traditional data integration architectures propose the use of a single common schema (e.g. data warehousing) or a mediated schema with distinct wrappers for each data source. These approaches would be very hard to maintain: Every time a new MS is added, the common data model or the mediator would need to be extended or changed, requiring complex data transformations. To achieve better extensibility we started with a mediator-wrapper architecture and adopted methods of the Data Space approach, called Data Space Support System [FHM05]. The design goal was to separate technical and semantic integration and preserve the autonomy and heterogeneity of the data sources.

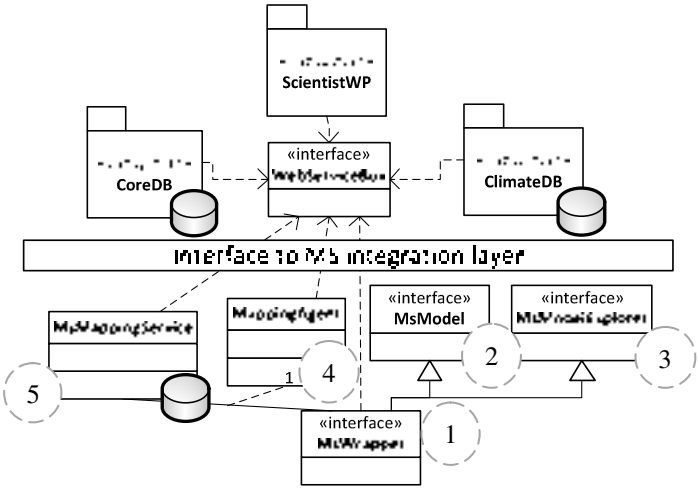

Fig. 2: Integration architecture for Measurement Stations

As the UML class diagram (Fig. 2) shows, a new MS can be quickly added without affecting other components. Just a lightweight wrapper (1) is needed that transforms MS data to the common data model (2). Additionally, all data is exposed in a flexible, table-like format by generic methods (3). The basic integration is completed by a mapping component (4) which is needed for every MS and relates common MS data and CoreDB concepts. Integrated queries across MS restricted to these relations are now supported by the central mapping service (5), as it stores and exposes all provided MS mappings.

A tighter, more complete integration with pay-as-you-go (data sources are integrated first technically, then semantic integration is added on demand [FHM05]) is achieved in two steps. First, a MS wrapper provides a machine-readable description of its supported

conceptual classes, attributes and all access methods. Second, the central mapping service can relate arbitrary concepts. Thus, more sophisticated integrated queries can be delegated to the ScientistWP, which can explore the generic methods using the MS descriptions. This integration approach outlined above has been successfully tested with the GROWSCREEN fluoro [Ja09] and the prototypic MS "PhenoSeeder". We could map 75% of the former MS' concepts to the common data model, and all of the latter.

4 Summary and Outlook

We have proposed a distributed information system, *Phenomis*, for the management of phenotyping mass data produced by MS with arbitrary storage models. These stations are integrated using a service-oriented architecture and concepts of a recent approach to data integration, Data Spaces. We have built a working prototype of Phenomis, consisting of a central plant database and two handling stations. It is in use at the JPPC and two MS have been successfully integrated in a pay-as-you-go style. So far it has been possible to discover and query these MS without knowing the underlying data models, just using plants managed by the CoreDB.

Phenomis is in development and as a research project (http://www.cropsense.uni-bonn.de/Forschung/teilprojekte/d1) part of the network for phenotyping science: CROP.SENSe.net, supported by German BMBF (0315531C). With the climate database and more MS added, we want to demonstrate Phenomis capability to integrate a broad range of phenotyping data. Integrated queries (using CoreDB concepts) across several MS are available with a first version of the ScientistWP completed. A flexible data export function is an integral part of the ScientistWP, eliminating the need of manual data integration. Combined with the formal interfaces for processing distributed queries, Phenomis provides a solid basis for new analysis pipelines to cope with the increasing amount of phenotyping data. Phenomis' extensibility and the exclusive use of established standards and free software (Web Services, Java, XML, PostgreSQL, etc.) make it easy to deploy and to adapt to other facilities. Portability is explicitly evaluated and tested in the PhenoCrops project, funded by EU EFRE (005-1105-0035).

References

[Ja09] Jansen, M. et al.: Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via GROWSCREEN FLUORO allows detection of stress tolerance in A. thaliana and other rosette plants, Functional Plant Biology 36(11), 2009, p. 902–914.

[Jo07] Josuttis, N.: SOA in practice: The art of distributed system design, O'Reilly, 2007.

[FHM05] Franklin, M.; Halevy, A.; Maier, D.: From Databases to Dataspaces: A New Abstraction for Information Management, ACM SIGMOD Record 34, 2005.

[Gr05] Granier, C. et al.: PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in Arabidopsis thaliana, New Pythologist, 2005.

[Pe12] Pereyra-Irujo, G. et al.: GlyPh: a low-cost platform for phenotyping plant growth and water use, Functional Plant Biology 39(11), 2012, p. 905-913.

[TS09] Tardieu, F.; Schurr, U.: White Paper on Plant Phenotyping, EPSO Workshop, 2009.