MS26 WP 11 – data cart released

Matthias Lange/Uwe Scholz

22.08.2014

The transPLANT consortium will provide an information infrastructure for genomics resources. In particular databases and information systems are heterogeneous in interfaces, location, data models and content. In order to enable a seamless access to the genomics databases and to integrate them with other public data repositories methods from information retrieval (IR) are applied. The transplant-IR environment will access the existing database infrastructure, which is a network of worldwide distributed and interlinked information systems.

The focus of this milestone 26 is to provide a Data Cart. As concept and infrastructure for the collection, transformation and distribution of data references using global identifier, the Data Cart consists of three elements, (1) a search engine module, (2) a data sharing and citation infrastructure, (3) an resolver for database accessions

(1) search engine module to store references to query relevant database entries

Beside the visual exploration of search query results in the Web frontend, a relevance ordered list of references to transPLANT databases is provided. It comprise the found hit in the indexed corpus of traits, phenotypes and protein functions as well related annotated features or positions in genome databases (Figure 1). The main attributes of a search result export are all references ("annotation_resource_URL") for a relevant annotation ("annotation_detail_URL"), the relevance score, the annotation evidence and information to the found link path ("annotation_link").

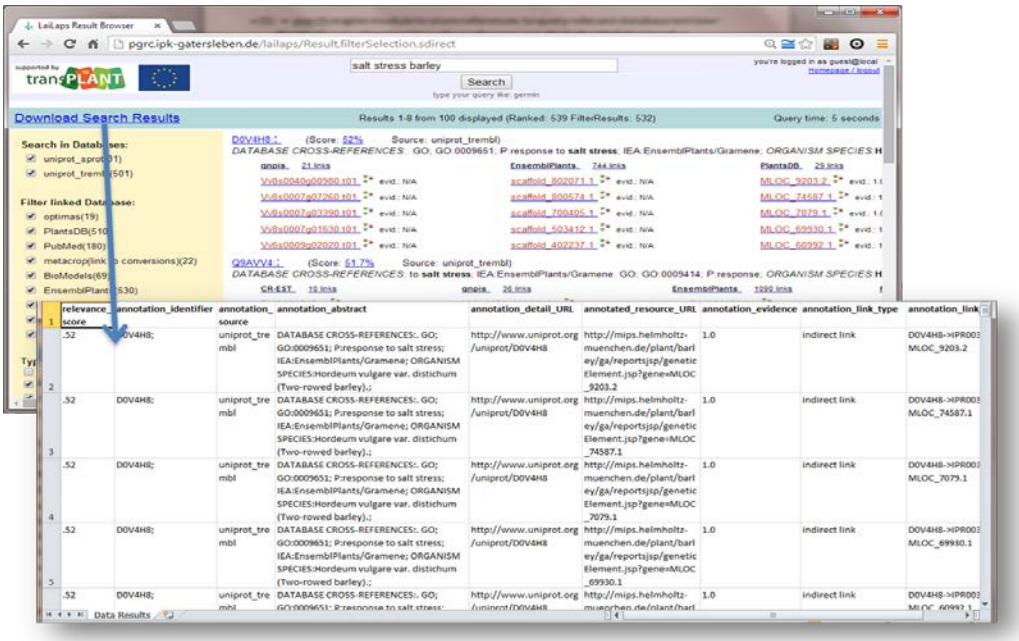

Figure 1 - Export of search results to data cart sheet

(2) data sharing and citation infrastructure

This component implements an infrastructure to store the referenced data records from search results, the result of downstream analysis as well as customer data. In consequence the e!DAL API has been developed as a lightweight software framework for publishing and sharing of research data (Arend, Lange et al. BMC Bioinformatics 2014). Its main features are version tracking, management of metadata, information retrieval, registration of persistent identifier, embedded HTTP(S) server for public data access, access as network system, and a scalable storage backend (see Figure 2). e!DAL is designed to embed into the data citation services of the international DataCite consortium (http://www.datacite.org) as data submission and registration system for DOIs. e!DAL is currently productive used to publish IPK primary data. A list of published data can be queried at: http://tinyurl.com/mo57qh5. The software can be downloaded from: http://edal.ipk-gatersleben.de

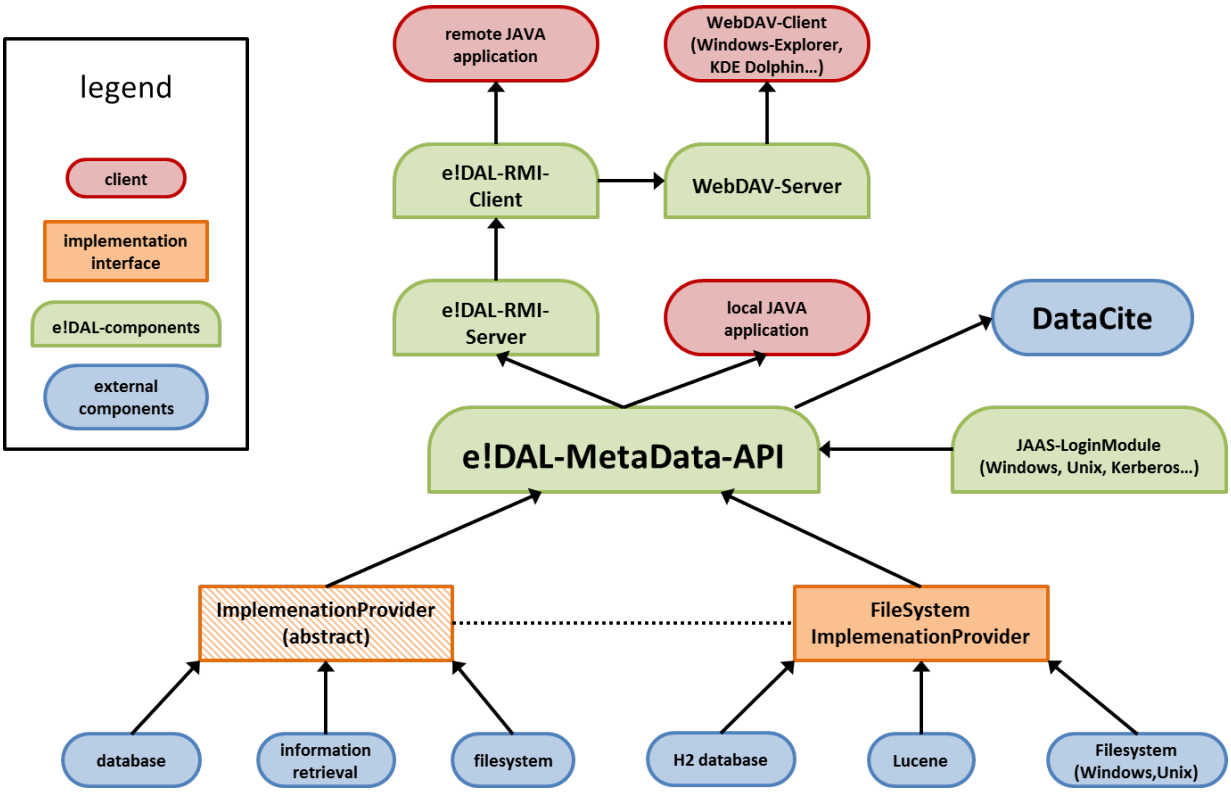

Figure 2 – e!DAL system architecture

(3) resolver for database accessions

The data content will be accessed in its original format and source by a data center specific URL. Doing so, for all databases that are referenced in the LAILAPS search engine we implemented a resolver. It translates identifier to its repository specific data-URL. For each linked database its URL pattern is stored and used to resolve database accessions to the internal link for a direct data access. This URLs are requested by partners and synchronized with the transPLANT list of genome resources and the public service “Identifiers.org” (http://www.identifiers.org) that is hosted at European Bioinformatics Institute.