**MS25 WP 11 - software core released**

Matthias Lange/Uwe Scholz

01.02.2013

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 25 is to provide a query system for genomics meta-data, like functional annotation of genes or other genomics regions. First task was to review the partner's data bases, their interfaces and referenced 3-rd party databases. We compiled a list of resources, which were used by the transPLANT partners to annotate their genomic data. Those repositories, i.e. UniProt, Gene Ontology, PFAM, are the systems of intersection, and on the other hand well known data hubs for information search.

Next step was to combine search engine technology with an annotation reference network. The approach is to search for keywords in the mentioned public repositories and map relevant hits back to the partner databases. This concept has been implemented using existing LAILAPS search engine technology, which has been developed at the IPK Gatersleben. LAILAPS search combines a keyword based search, recommender system and an AI-based user specific relevance ranking.

The aim was to enhance LAILAPS towards integrated search engine for data networks. We collected references from those partners, who annotate their genome data using controlled vocabulary, ontologies or textual description of gene functions from public databases. The resulting compilation comprises a list of 8 major databases (Trait Ontology, Pfam, Gramene, Plant Ontology, SwissProt, TrEMBL, Gene Ontology, PDB). At the first step we indexed those databases in LAILAPS and already cross-linked IPK data bases GBIS, MetaCrop, Cr-EST as well as the genomics information system from transplant partners EBI (Ensembl database) and INRA (GNpIS database).

The implemented **software core** of the enhanced LAILAPS system (http://lailaps.ipk-gatersleben.de) now enable keyword search in integrated genomics resources and result in relevance ranked links to genomic data from transPLANT partner (see Figure 1).

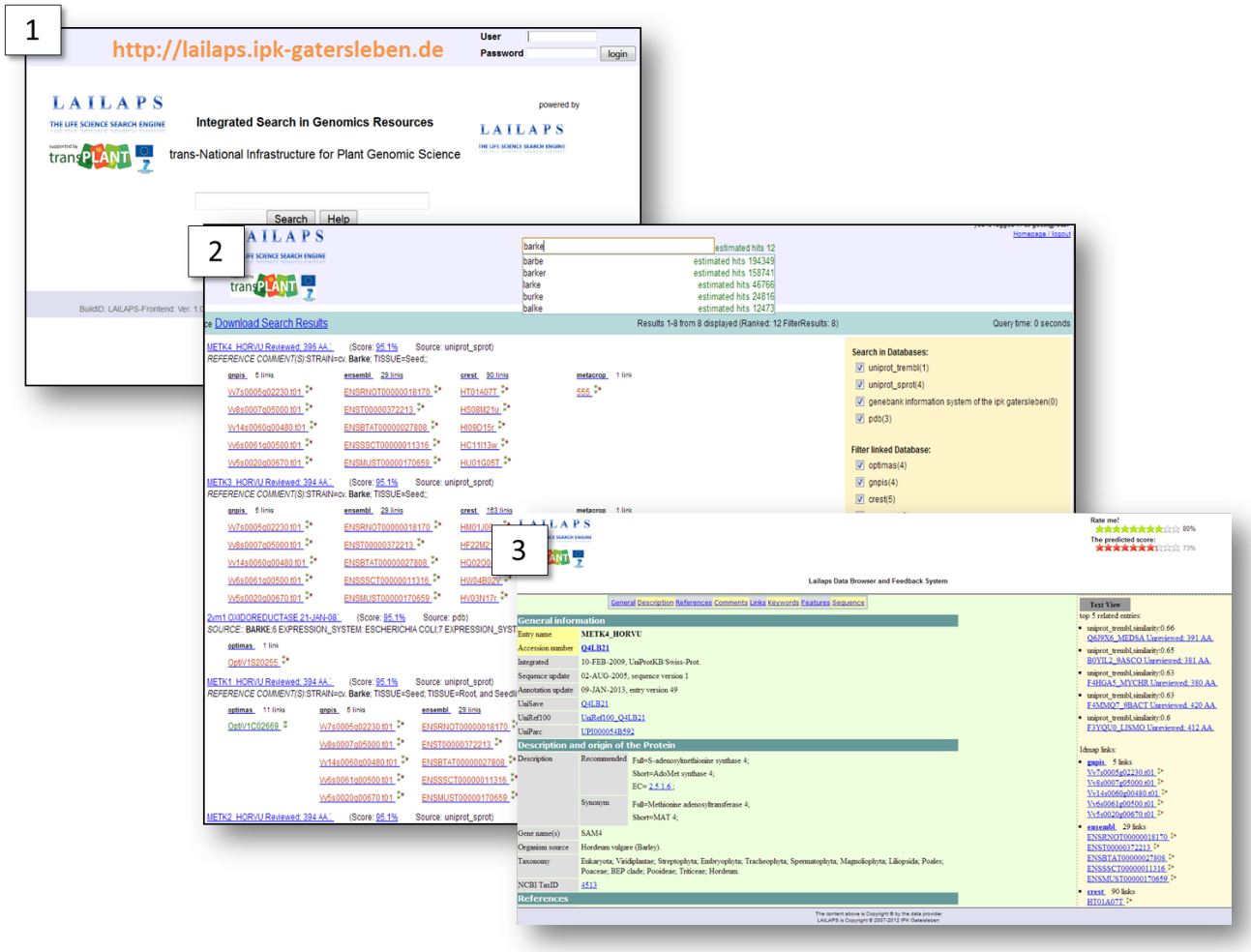


*Figure 1: The LAILAPS Search Engine for transPlant for integrated search in transPLANT genomics data network. Part (1) shows the entry point of the search engine. In screenshot (2) a result of a keyword search for "barke", a genotype of barley, is shown. The result contains relevance ranked hits in indexed genome annotation data hubs (UniProt, GeneOntology, PFAM etc.) and related linked genomic resources, i.e. Ensembl, GnpIS, CR-EST. In screenshot (3) the integrated data browser and feedback system, which act as input for the incremental training of the relevance predicting neural network.*