**transPLANT milestone report**

**MS16 (work package 6): Ensembl Plants, MIPS Plants DB and GnpIS integrated in transPLANT portal**

An integrated search has been developed to allow users to cross-query the primary databases represented in the transPLANT portal.

Search has been implemented using the Solr (v4.0) text search framework. A schema has been devised (in an iterative process) to describe information from the range of partner resources available, with most resource specific information going into a free-text description field. Additional fields were defined to capture the associated resource identifier, URL, and species information. One meta-data field records the partner database from where the associated information was obtained. Use of the schema ensures that matching data types can be identified as identical and co-filtered in a faceted search (see below).

The schema uses a simple inheritance pattern, such that more detailed entry types (defined below) inherit all the fields of the parent types. Sub-types can override the definition of fields, i.e. making an optional field mandatory. The schema is detailed in table 1, below.

**Table 1**. The transPLANT search data model

|  |
| --- |
| Hierarchy of data types:   * Database entry   + Sequence annotation   + Accession (variety, individual, genotype, germplasm, etc.)     - Phenotype annotation   + Experiment (trial)   + Marker   + Reaction   + Other (anything else)   Fields  (Optional fields are indicated with [braces].)  Database entry   * id - Unique id per item, per data-source. * text - Text to be indexed for searching. * url - A url for the entry in the partner database. * [entry type] - Type of entry. Should be one of the given entry types. * [species] - The associated species   Sequence annotation   * type - The sequence feature type, constrained to be a Sequence Ontology term, e.g. protein, transcript, EST, variation, or gene. * reference sequence id - Sequence on which the feature is being annotated. * start position - The start of the feature on the sequence (1-based). * end position - The end of the feature on the sequence (1-based).   Accession   * authority - The accessioning authority for the given id.   Phenotype annotation   * text - describing the phenotype.   Marker   * position - Marker position in cM. * [anchor] - id(s) of the sequence annotation(s) anchoring this marker on a given reference sequence.   Reaction   * participant - id(s) of the sequence annotation(s) for the participants of the reaction.   Other   * entry type - A mandatory type. |

Data is collected from partners in a simple, tabular format, and was loaded into the Solr search index using a sub-set of the schema described. The free text field is processed using tokenization and filtering rules optimized for English, allowing stemming and wild card search. Faceted fields are indexed separately to facilitate faceted searching. All fields, with the exception of URL, were copied to the description field index so that by default, searches for text in these fields would return the expected results.

The current data set (from all providers) can be indexed by the Solr server in under 10 minutes.

**Drupal integration:** The Solr search index is queried and the results presented within the Drupal powered transPLANT web-portal. Drupal has a modular plug-in system, and an existing module for linking Drupal with Solr was adapted for these purposes. The module depends on several APIs that are themselves provided by other modules, as described below. The module we developed functions to:

1) create a search page using the search API that uses the appropriate connection details of the Solr server,

2) register facets with the facet API, providing the facet blocks of the search page, and

3) process and display search results on the search page, adding paging, grouping and formatting for elements.

**On-going integration strategy**: Two different strategies can be employed to keep the search up to date. We can continue to accept static dumps of database information from partners, periodically updating the central index with this information. Alternatively, partners could establish their own Solr instances, using the common schema but indexing specific data. In this system the Drupal 'client' would issue a distributed search across several servers, integrating the results in the presentation layer. We have developed the framework for implementing both approaches, and an individual resource can be switched between the two modes of operation through a simple configuration step.

The presentation of the search results, which were originally displayed in a tabular form, has now been restructured, to make the top hits from more resources immediately visible. Navigation through the search results is provided mainly through the use of several search facets, covering the data source, data type, and species. Multiple facets can be added or removed dynamically, allowing the user to explore the available information before selecting one or more results. An examples of the faceting in use is shown in figure 1. Each result is accompanied by a URL, linking the user back to the appropriate source information.

The integrated search is now live on the web portal at <http://www.transplantdb.eu>. In addition to supporting search of the three resources originally envisaged, the integrated search also encompasses several additional transPLANT partner resources, namely IPK (CR-EST, GEBIS and MetaCrop) and PAS (PolapgenDB). A complete summary of the complete data currently integrated is shown in Table 1.

Table 2: Summary of partner resources indexed:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Partner** | **Database** | **Data types** | **No. data points** | **No. species** |
| EBI | Ensembl Plants | Gene-centric | 1,072,657 | 26 |
| MIPS | PlantsDB | Transcript-centric | 263,401 | 6 |
| IPK | CR-EST | ESTs | 218,927 | 6 |
| GEBIS | Passport data | 148,696 | 5,140 |
| MetaCrop | Biochemical reactions | 585 | 286 |
| PAS | PolapgenDB | Phenotype-centric | 93 | Hordeum vulgare |
| UGRI | GnpIS (Vitis) | Variations, markers and genes | 168,179 | Vitis vinifera |
| Siregal | germplasm | 16,266 | 3,278 |
| Ephesis | Trials, phenotypes, and accessions | 334 | 6 |

**Figure 1**.

A search for "carbamoyl synthase" currently returns 14,269 results. The 2 most relevant results form each source database are displayed to the user, with paging functionality above and below the results table. The user can select a resource of interest from the data source facet. After selecting Ensembl Plants, for example, the user can see the number of results per species update in the species facet, and can select a species of interest. Finally the user can review and select the different data types matching the search, given the current filters, can select one and, in this case, can click on the link to view the information in the Ensembl browser.

