The eDAL Suite: Tools and Concepts for Primary Data Citation

M. Lange, C. Colmsee, S. Flemming, J. Chen, M. Klapperstück, U. Scholz

Research Group Bioinformatics and Information Technology,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Gatersleben, Germany
Contact: <lange>|<colmsee>@ipk-gatersleben.de

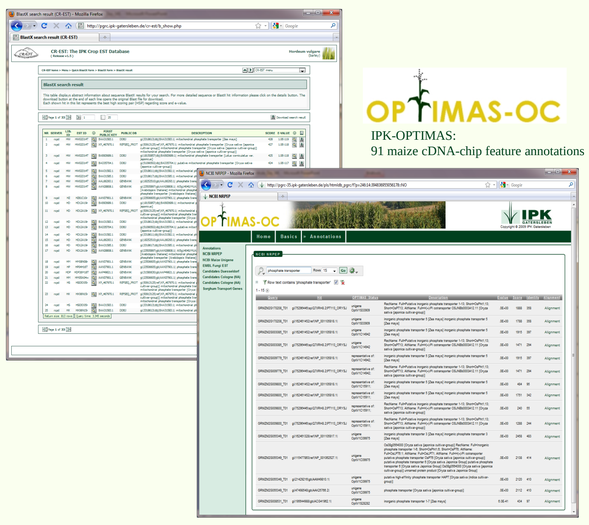
Motivation

Retrieval and citation of primary data is the important factor in the approaching age of "data science". Digital data are easily shared, and just as easily wiped or lost. The problem of keeping on-line data accessible and retrievable is especially difficult for SME like plant breeders plant biotech companies as well as research projects in this domain.
Intension of eDAL is the provisioning of an information retrieval and data citation infrastructure that meets the requirements of the "data science" age and implements a re-usable platform for data retrieval, data citation, and data publication. Like a shopping cart, the idea is to combine a search engine and a data cart, which retrieves, rank and collect query relevant data from crop plant data centers.

Distributed, Redundant, Alternative Data @ IPK: 916 data sets for „phosphate ransporter" in 4 IPK databases

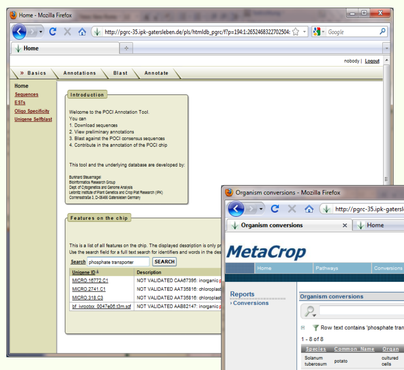
IPK Crop EST Database:
813 EST annotations

IPK-OPTIMAS:
91 maize cDNA-chip feature annotations

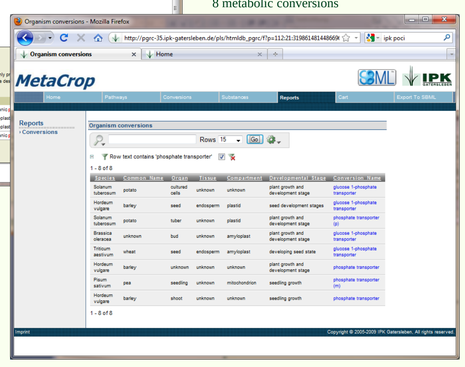


POCI

IPK-POCI:
4 Solanaceous cDNA chip features



IPK-MetaCrop:
8 metabolic conversions



Availability of Data

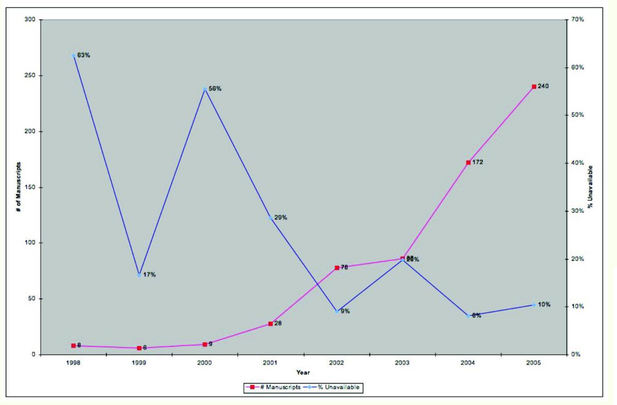


Chart: Number of manuscripts vs. % unavailable by year (1998–2005)

The increasing number of digital data should lead into a long-term storage system, within the data is available at all time. But Anderson et al. illustrates the problematic of the loss of digital data within biomedical publications. The reported up to 30% unavailability of supplementary data in 9 years.

Data Citation Rate

The importance of sharing data within publications is shown in Piwowar et al. for microarray data. Publications comprising shared data have a 69% higher citation rate than publications without shared data.

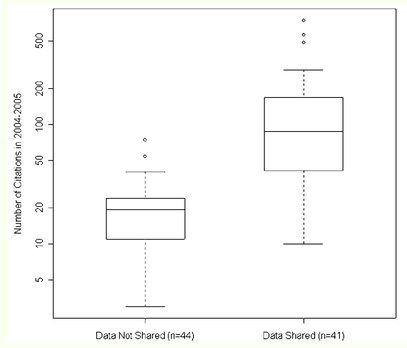
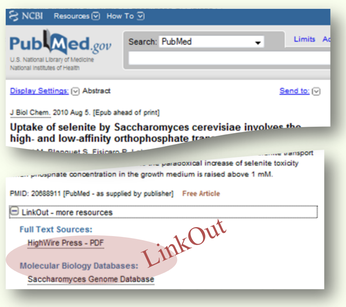


Chart: Distribution of 2004–2005 citation counts of 85 trials by data availability.

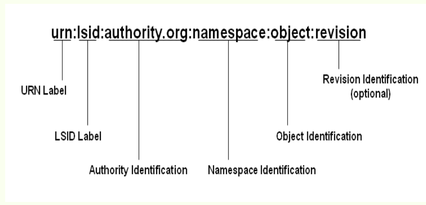
State of the Art

Data Citation

Properitary ID's:
e.g. Accessions, PubMed-LinkOut



Standard ID's:
e.g. Life Science Data Identifier (LSID)



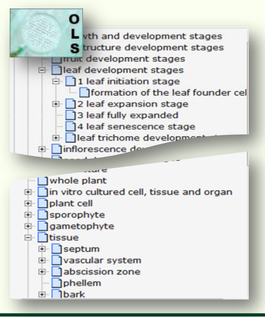
Meta Data

Technical meta data:
e.g. ISO 15836 (DCMES)

Dublin Core Metadata Initiative
The Dublin Core Metadata Element Set

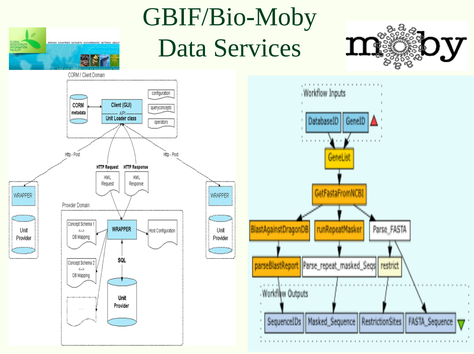
- Title
- Subject
- Description
- Creator
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

Semantic meta data:
e.g. EBI Ontology Lockup Service



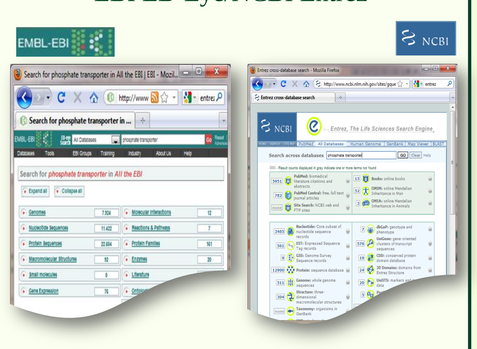
Data Access

GBIF/Bio-Moby Data Services



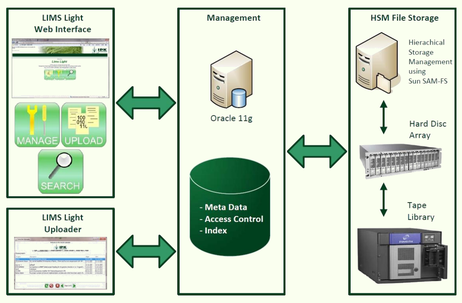
Search Engine

EBI EB-Eye/NCBI Entrez



Primary Data Management at the IPK

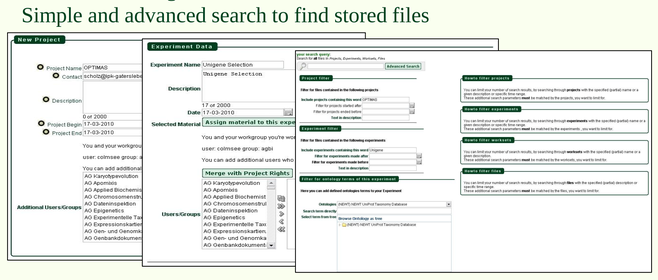
LimsLight: IPK Primary Data Mangament



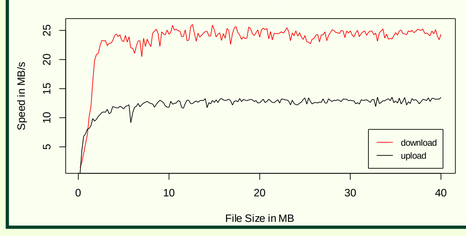
The **architecture** is divided in three parts:
1. The client components to upload, search and download the data,
2. The database to maintain meta information (up to a range of Gigabytes) ,
3. The raw data storage component to archive primary data (volume above Terabyte range)

The LIMS Light Web Interface:

- Web application based on ORACLE Application Express (**APEX**)
- Management of Projects, Experiments, Worksets and files
- Using controlled vocabulary (**Ontology Lookup Service**) and free text information to tag the files
- Simple and advanced search to find stored files



Performance Progression



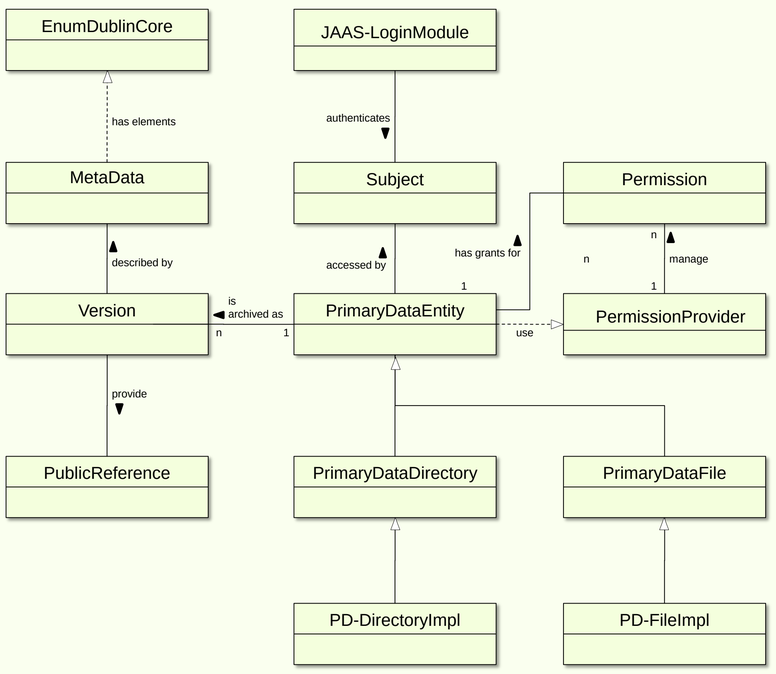
1. Downloading files with a speed of ~24 MB/s
2. Uploading files with a speed of ~13 MB/s
3. Meta data maintenance for high number of small files causes low upload and download rates

eDAL-API: Interface to Store, Manage and Annotate Primary Data

The eDAL-API takes care of the central aspects of primary data management:

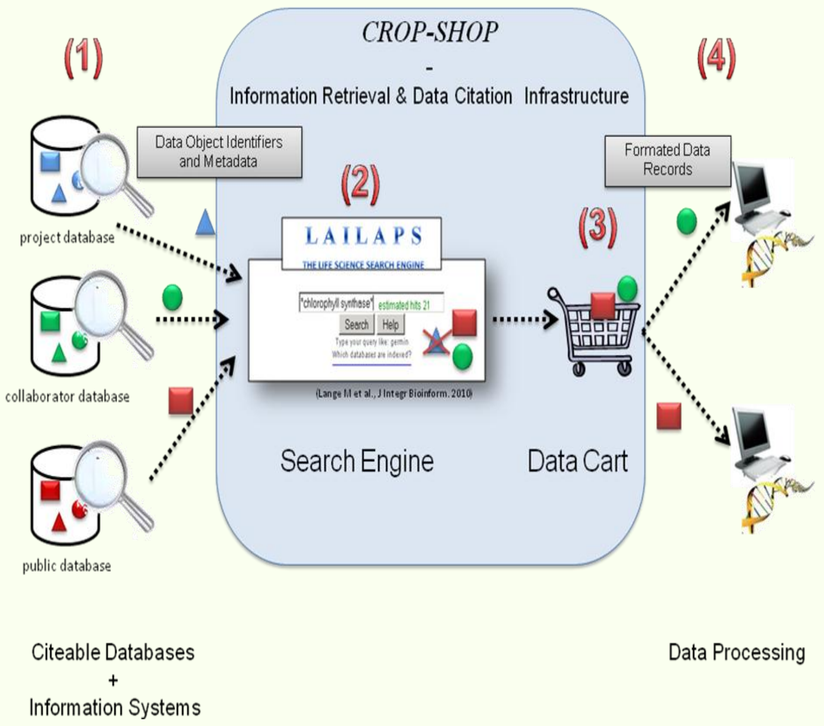
- **Storage** component to archive the primary data
- **Persistent identifier** to make primary data citable
- Dublin Core for defined **meta data** information
- JAAS **security** for data access control
- **Versioning** of primary data

With the eDAL-API the user can implement his own Primary Data Management system and could realize the information retrieval by using the life science search engine like **LAILAPS** (Lange et al. 2010).



Outlook: The CROP-SHOP approach (submitted as BMBF proposal)

Data centers and primary data archives (1) provide access to their databases and information systems by a uniform data citation system. The information retrieval component is featured by a search engine (2) and a data cart (3). The search engine retrieves data, which are ranked for their query and user specific relevance profiles. The data cart module collect global object identifiers resulting from a query. The identifiers are linked to administrative meta-data and will be delivered to the end-user for later data processing (4) in adequate data formats.



References

- M. LANGE et al. (2010) The LAILAPS Search Engine: Relevance Ranking in Life Science Databases. Journal of Integrative Bioinformatics, 7(2):e110, 2010.
- Anderson et al. (2006) On the persistence of supplementary resources in biomedical publications. BMC Bioinformatics 2006, 7:260. doi:10.1186/1471-2105-7-260.
- Piwowar et al. (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3):e308. doi:10.1371/journal.pone.0000308.

Nature Precedings - doi:10.1038/npre.2010.4935.1 : Posted 28 Sep 2010