

Application of database technology to manage, preserve and analyse plant genomics and phenomics data

[Extended Abstract]

Matthias Lange

Leibniz Institute of Plant Genetics and
Crop Plant Research (IPK), Germany

lange@ipk-gatersleben.de

EXTENDED ABSTRACT

The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) is committed to the conservation and valorization of plant genetic resources. Its research agenda comprises upstream and downstream analyses in the fields of genetics, physiology and cell biology aiming at a broad understanding of plants at molecular, cellular and organismic levels. The “big data” challenge has long been reached the IPK and the life science community in general. It forces the need for capable infrastructures to document, share, publish, integrate and explore research data. In this talk, we give an overview to IPK’s research projects in the fields of Lab Data Management, Data Citation and Information Retrieval.

Data documentation (LIMS)

Handling research data is an important task in IPK’s research strategy and particularly a central component along the value-added chain towards scientific publications, patents and biotechnological innovations. In consequence, there exists the conviction that it is essential to implement an intuitive and seamless data storage and documentation infrastructure, which can be easily embedded into existing workflows and will be highly accepted by scientists.

In practice it is very challenging to meet this aim. Bioinformaticians implemented individual, isolated and heterogeneous systems to manage experimental data, laboratory workflows and biological samples. Those considerations were the driving force for establishing a central Laboratory Information Management System (LIMS). In the talk, we will present experiences from last 6 year LIMS data management, pitfalls and successfully implemented lab processes.

Data Publication (e!DAL)

Besides publication of scientific findings, it is important to keep the data investment and ensure its future processing. This implies a guarantee for a long-term preservation and preventing of data

loss. Condensed and enriched with metadata, primary data would be a more valuable resource than the “re-extraction” from articles. In this context it becomes essential, to change the handling and the acceptance of primary data within the scientific community. Data and publications should be honored with a high attention and reputation for data publishers.

Here, we present e!DAL [2] (<http://edal.ipk-gatersleben.de>) as a lightweight software framework for the publication and sharing of research data. Its main features are version tracking, management of metadata, information retrieval, registration of persistent identifier (DOI), embedded HTTP(S) server for public data access, access as network file system, and a scalable storage backend. e!DAL is available as API for a local non-shared storage and remote API to feature distributed applications. IPK is an approved data center in the international DataCite consortium and apply e!DAL as data submission and registration system.

e!DAL is the software infrastructure for the Plant Genomics and Phenomics Research Data Repository (PGP) [3], a repository to comprehensively publish plant research data: This covers in particular cross-domain datasets that are not being published in central repositories because of its volume or unsupported data scope, like image collections from plant phenotyping and microscopy, unfinished genomes, genotyping data, visualizations of morphological plant models, data from mass spectrometry as well as software and documents.

PGP is registered as research data repository at BioSharing.org, re3data.org and OpenAIRE as valid EU Horizon 2020 open data archive. Above features, the programmatic interface and the support of standard metadata formats, enable PGP to fulfil the FAIR data principles—findable, accessible, interoperable, reusable.

Information retrieval (LAILAPS)

Due to advances in high-throughput technologies, the amount of data available over life science web resources is growing rapidly. It is becoming an increasingly difficult and time consuming task for scientists to derive information from those resources and to keep up-to-date even within their own field of research. For example, correct identification of causative genes for an important agronomic trait can be very valuable for effective marker assisted breeding. However, even well-defined QTL often span genomic regions that can contain hundreds of positional candidate genes. Evaluation of potential functional candidates from such long lists is often time-consuming and requires the integration of information from many different sources. In this context,

information retrieval (IR) is evolving to a key technology. Its increasing popularity for data exploration is because there is no need for a user to have knowledge about complex query languages, underlying data structures or data formats.

Here we will present how to use the LAILAPS [3] integrative search engine for plant genomics data (<http://lailaps.ipk-gatersleben.de>), which is developed in the frame of EU transPLANT consortium. LAILAPS supports the integrative search over the distributed genome annotation (traits, gene functions, agronomic factors). For this, 50 million records of most popular used genome annotation repositories, like UniProt, BioModels, OBO ontologies and PDB, are indexed. Moreover, 80 million gene annotations of plant genomic resources are linked by reverse identifier mapping. In order to select most relevant candidate genes for queried traits, LAILAPS use context based relevance ranking. The order of the search hits is computed by an artificial intelligence driven relevance ranking, which has been trained by domain experts and evaluated for QTL candidate gene prediction.

1. REFERENCES

- [1] Daniel Arend, Christian Colmsee, Helmut Knüpffer, Markus Oppermann, Uwe Scholz, Danuta Schüler, Stephan Weise, Matthias Lange. Data management experiences and best practices from the perspective of a plant research institute. In: Galhardas H, Rahm E (Eds.): Data integration in the life sciences: 10th international conference, DILS 2014
- [2] Daniel Arend, Matthias Lange, Jinbo Chen, Christian Colmsee, Steffen Flemming, Denny Hecht, Uwe Scholz. e!DAL - a framework to store, share and publish research data. BMC Bioinformatics. 2014 Jun 24;15(1):214
- [3] Daniel Arend, Astrid Junker, Uwe Scholz, Danuta Schüler, Juliane Wylie, Matthias Lange. PGP repository: a plant phenomics and genomics data publication infrastructure. Database. 2016
- [4] Maria Esch, Jinbo Chen, Christian Colmsee, Matthias Klapperstück, Eva Grafahrend-Belau, Uwe Scholz, Matthias Lange. LAILAPS - The Plant Science Search Engine. Plant and Cell Physiology. 2014 Dec 24;55(1):pcu185

About the Author

Dr. Lange works for more ten years in the field of bioinformatics and data management in life sciences. His primary research topics are dedicated to information retrieval, search engine technology, and research data management of different data domains, e.g. sequence, marker, metabolic and especially phenotypic information. His special focus is on standards and infrastructures for data sharing and publication. Here, he contributed to the MIAPPE and ISATAB standard for plant phenotyping data. Furthermore, Dr. Lange coordinates the central lab information systems in the IPK and is responsible for IPK datacenter activities in the frame of the DataCite consortium. As core service activity he is deputy administrator of IPK's ORACLE enterprise datamanagement backend. Furthermore Dr. Lange supervises work packages in German Plant Phenotyping Network (DPPN), the German Network for Bioinformatics Infrastructure (de.NBI) and contributes in EU transPlant research project to build up a trans-national data infrastructure for plant genomics data.

ORCID <http://orcid.org/0000-0002-4316-078X>