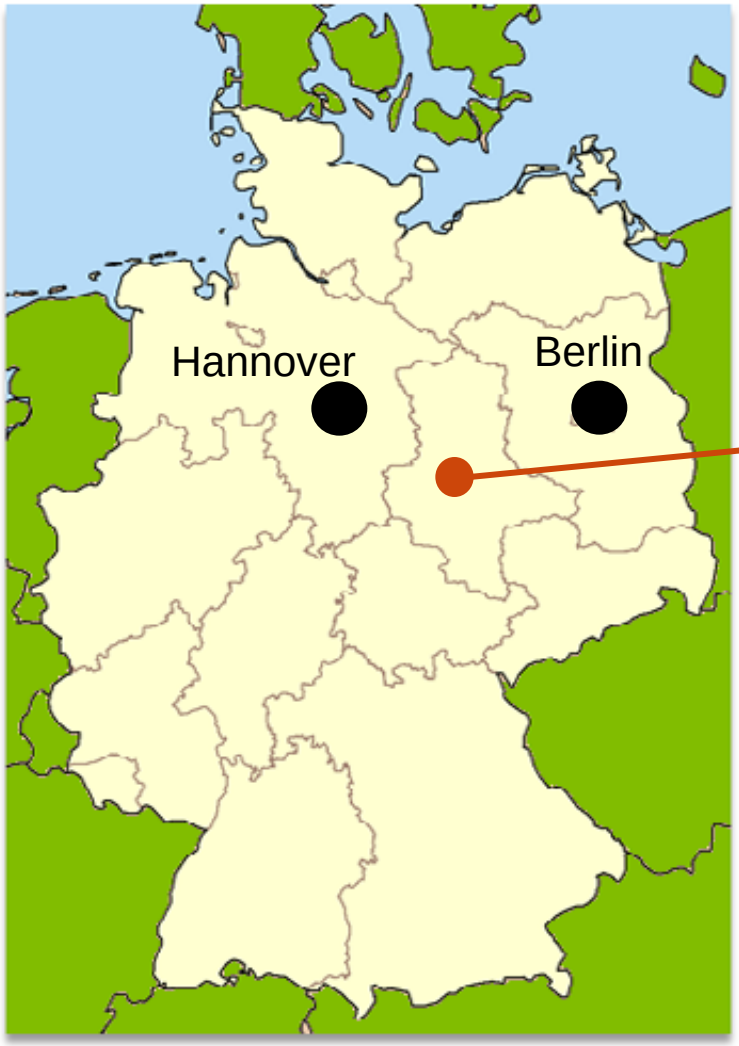
Management von Forschungsprimärdaten und DOI Registrierung

Dr. Matthias Lange (Bioinformatics & Information Technology)
June 19th, 2013

Outline

- Motivation: IPK data infrastructure
- LIMS: Integration of Lab Processes and Data
- DataCite: publish research data as citable resource

IPK - Leibniz Institut of Plant Genetics ans Crop Plant Research



- IPK is 70 years old
- „**Magdeburger Börde**“: soil with very high quality
- Source of the breeding industry in Germany

IPK - Leibniz Institut of Plant Genetics ans Crop Plant Research

Total staff: ~550

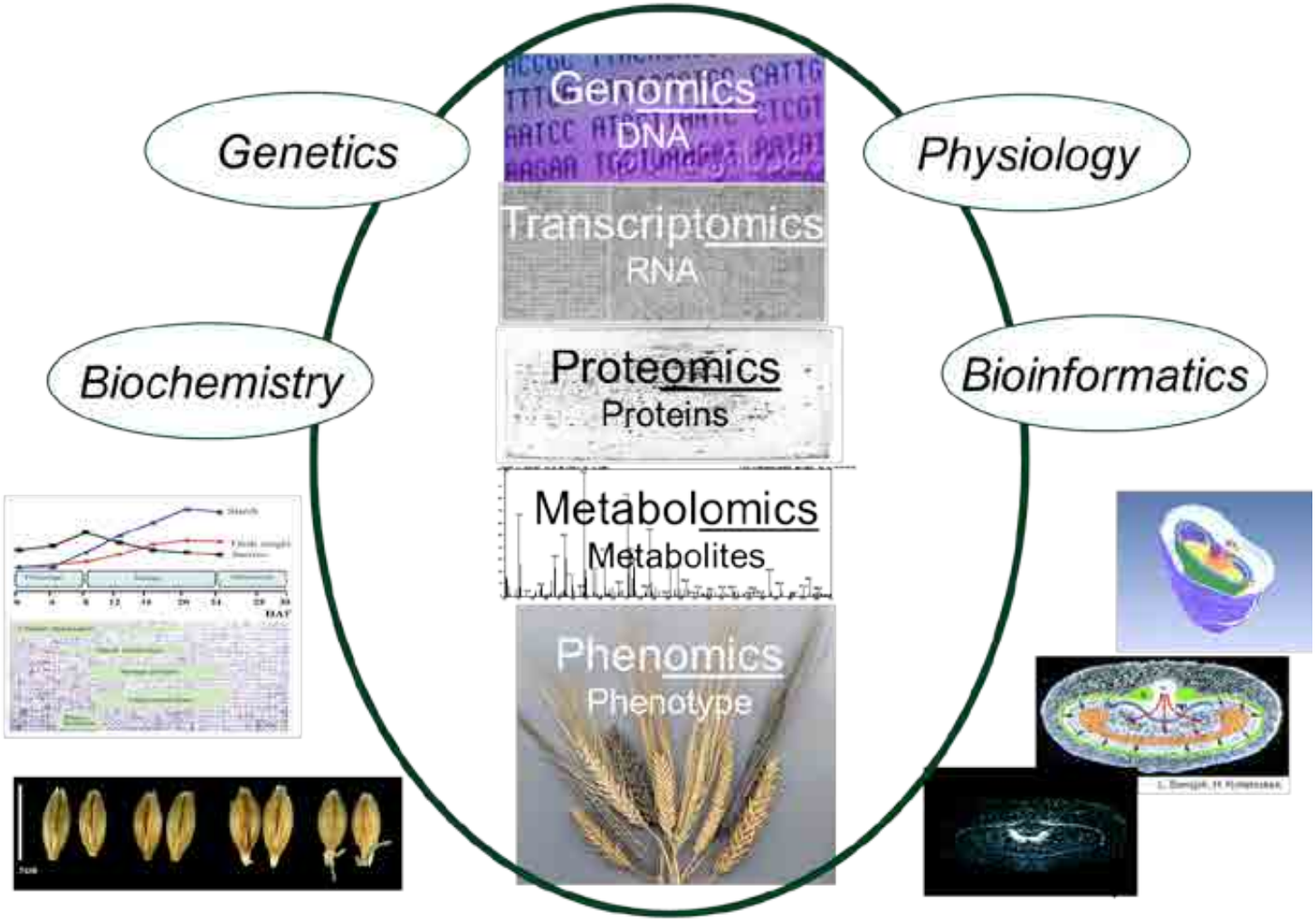
Scientists: ~200

30 Research Groups

Research Topics:

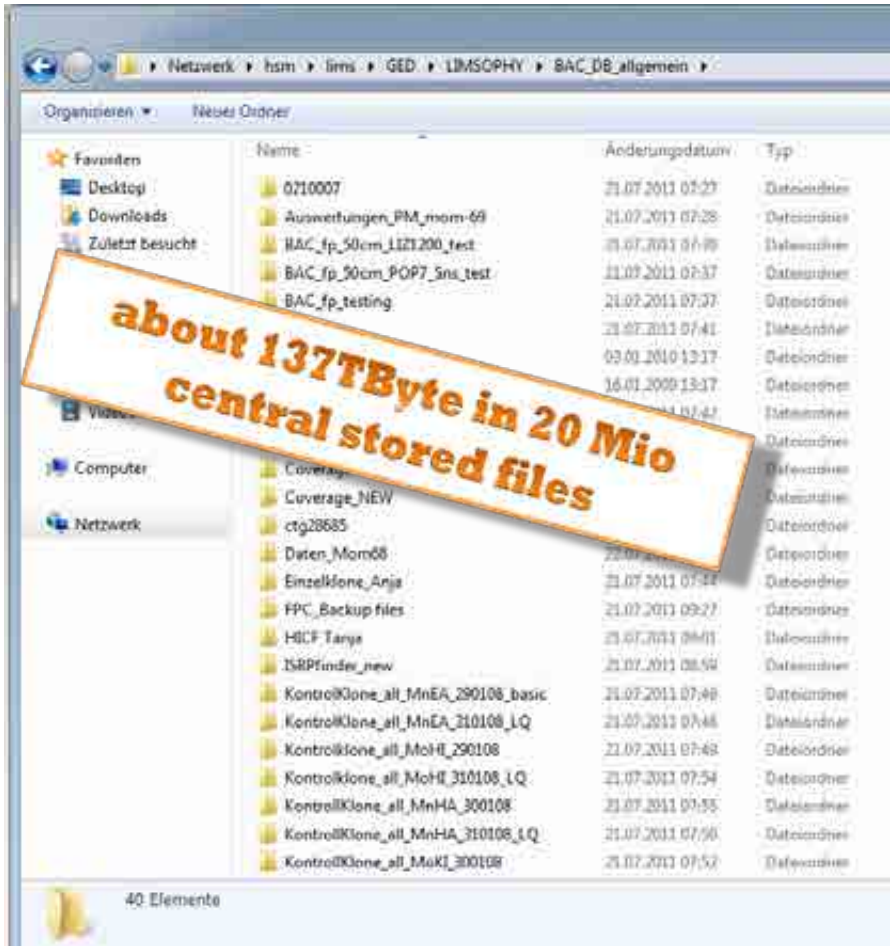
- **Diversity of crop plants**
- **Dynamics of plant genomes**
- **Integrative biology of plant performance**

IPK Data Domains

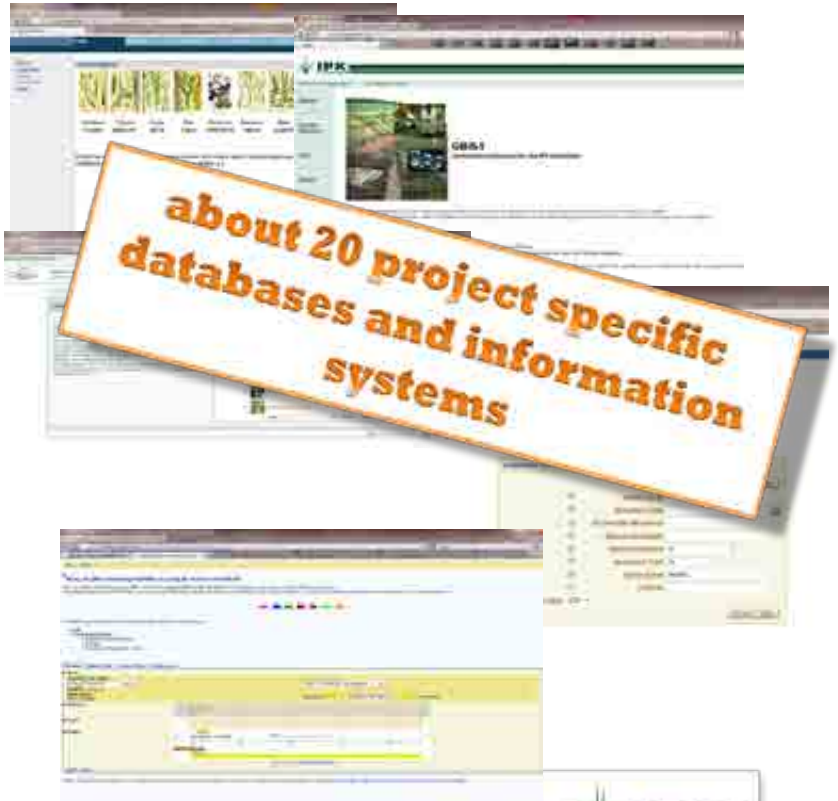


Status @ IPK: Data Access

- File system exploration/indexing (desktop search)

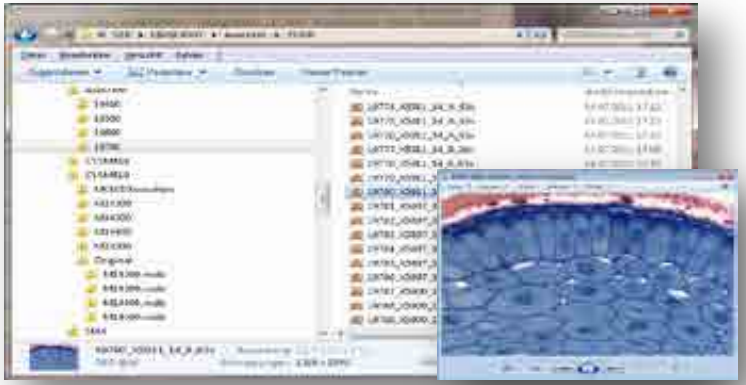


- Databases/Web applications

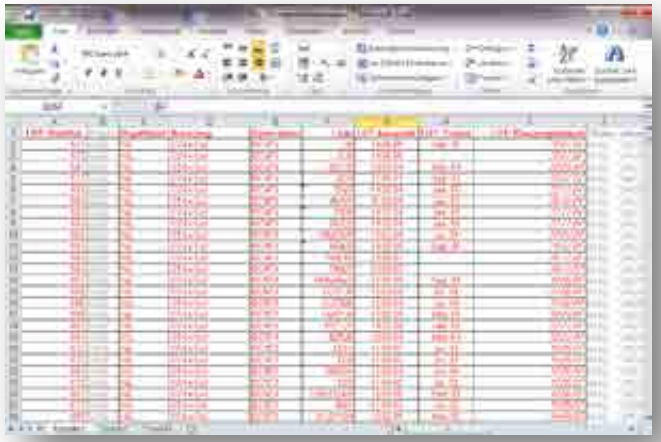


Status @ IPK: Documentation

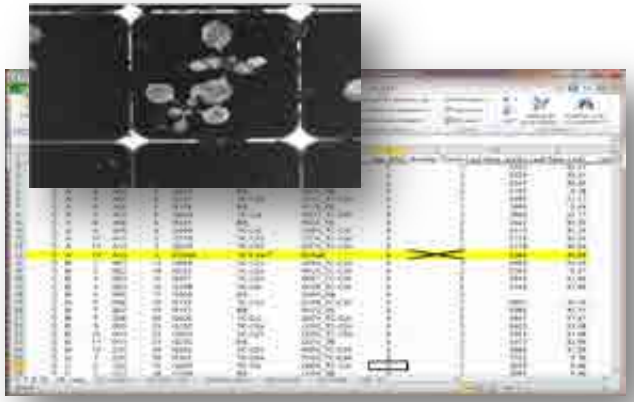
Sample storage - freezers, pockets and boxes are labeled with numbers



File management - coding device, condition, material into file names

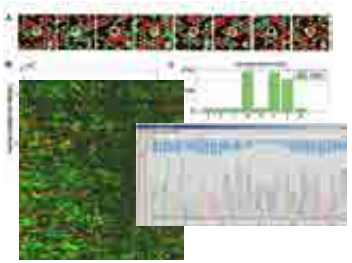


Seeds management - lines, crossings and sowing



Phenotyping result management - plant, box places, measurement

General “Publication Pipeline”



Sample ➡ Measurement ➡ Quantification ➡ Analysis ➡ Publication

Data Management:

- Meta data
- Protocols
- Analysis results
- Primary data
- …

Primary data *(Neuroth et al. 2010):*
The definiton of “primary data” is not clearly fixed. for some it is a raw data stream from a device, also called “Level 0” data. For others, it is pre-processed raw data, but without scientific analytic processing steps. Still others consider all data, which is used for scientific publications.

Requirements to Data Management

- Efficient storage of primary data (e.g. FASTQ files).
- Efficient storage of analysis results (e.g. VCF files)
- Managing project meta data
- Support data publication (e.g. DOI's for data sets)

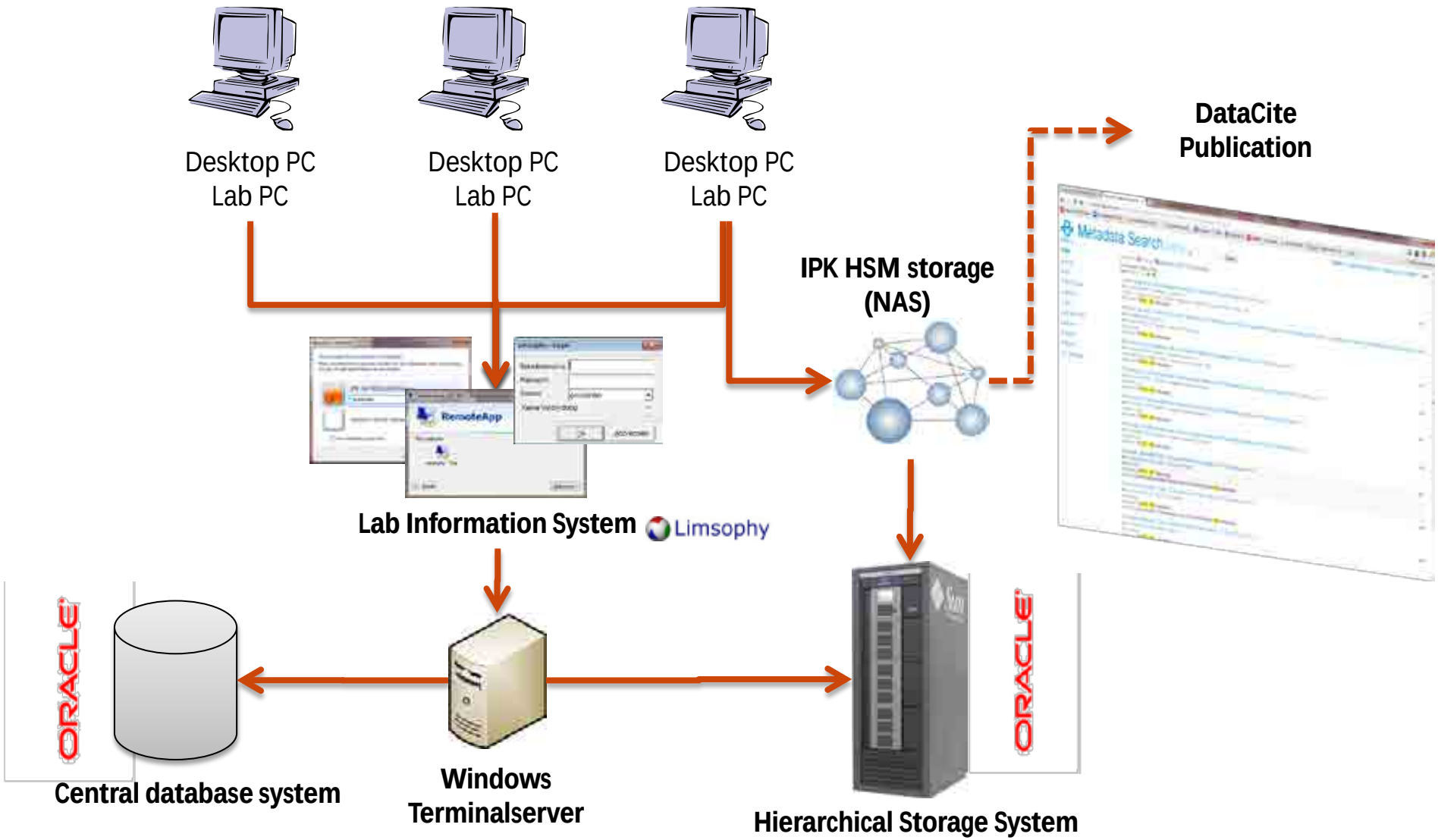
IPK Solutions:

Storage à **Hierarchical Storage Management: HSM**

Meta Data Management à **Labor Information Management System: LIMS**

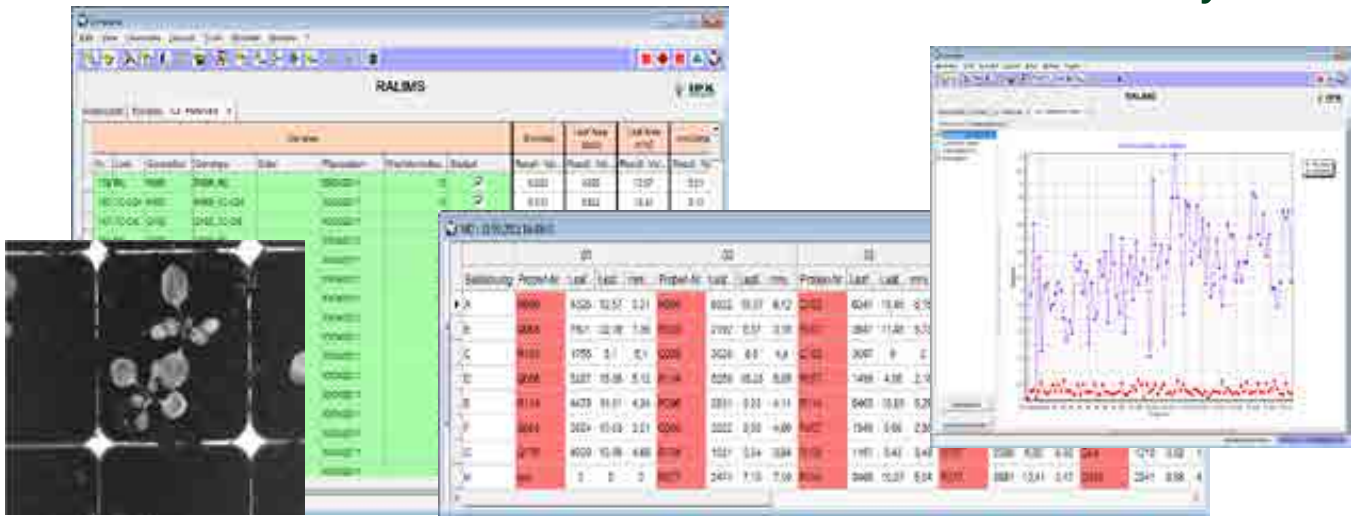
Publication à **electronical Data Archive Library: e!DAL**

IPK Data Managment Infrastructure

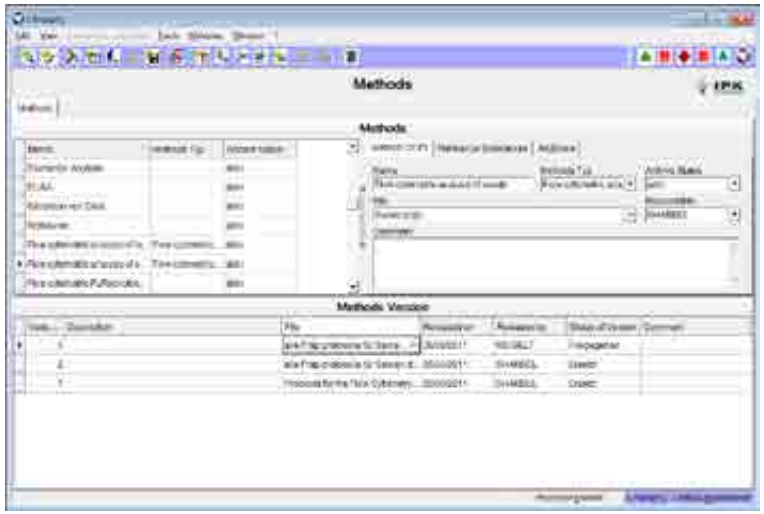


Lab Information Management:

- **Commercial LIMS System**
- **Substances, Devices, Methods, …**
- **Experimental Data**
  - Discrete values
  - Numbers
  - Boolean
  - Text
  - Binary files

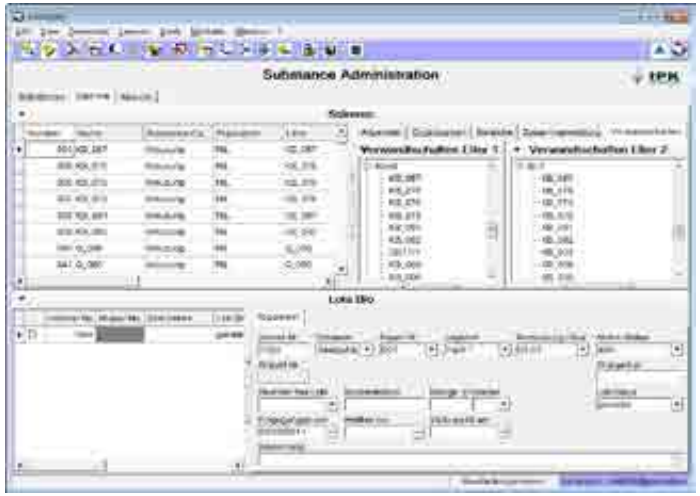


LIMSOPHY Modules

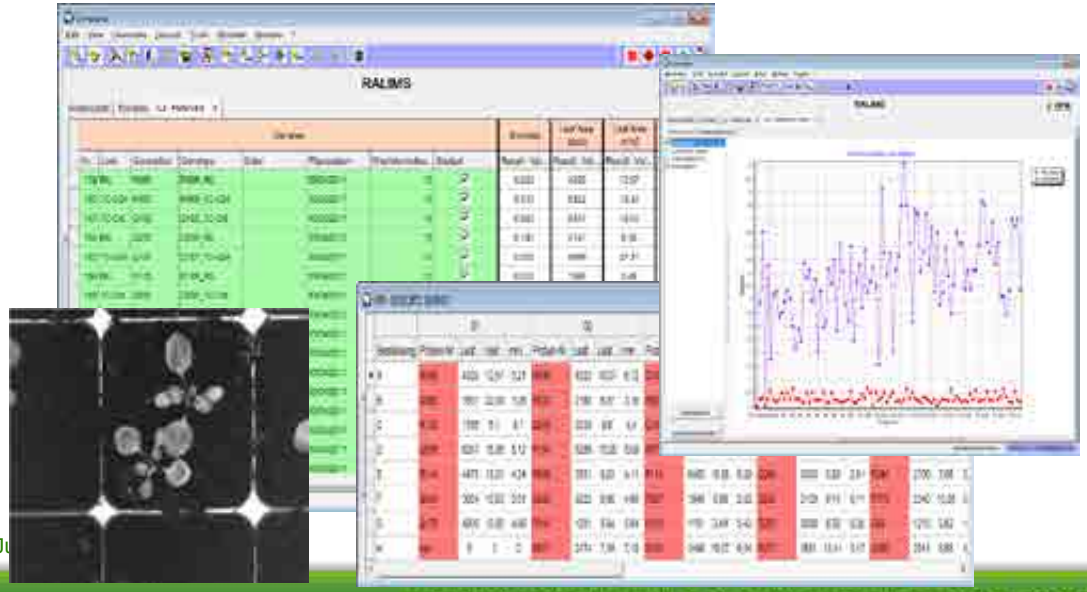


- **LIMSOPHY main menu**
  - function modules are grouped into management categories
  - each user has personal views, layout and permissions
  - manage and versioning of general descriptions of used methods, protocols and lab instructions
  - the documents are archived, secured, audited and finally assigned to related processes

LIMSOPHY Modules

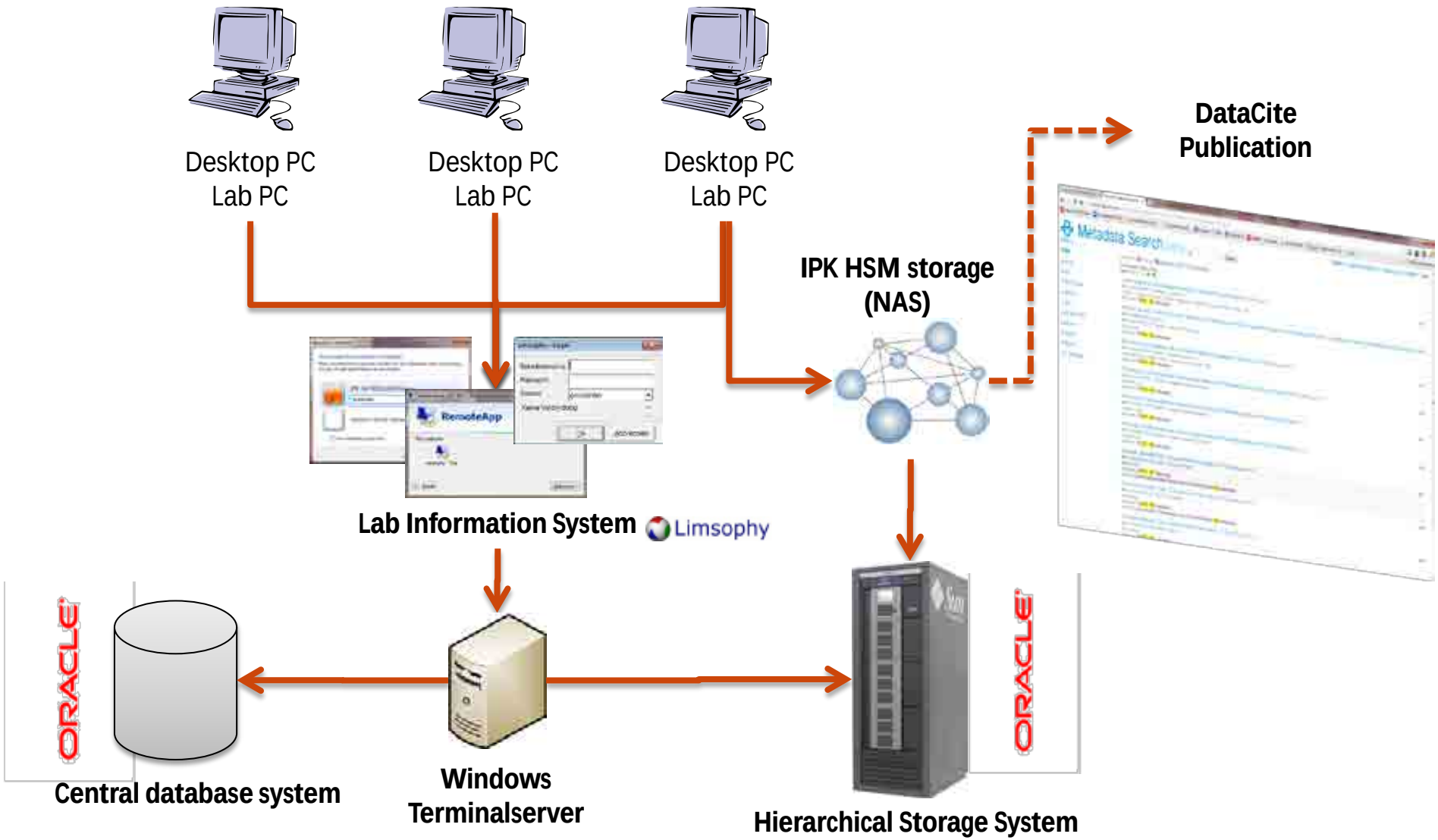


- **Substances, Chemicals and Plants**
  - management of substances like chemicals or biological material
  - this includes support for storage, stock and phylogeny management etc.

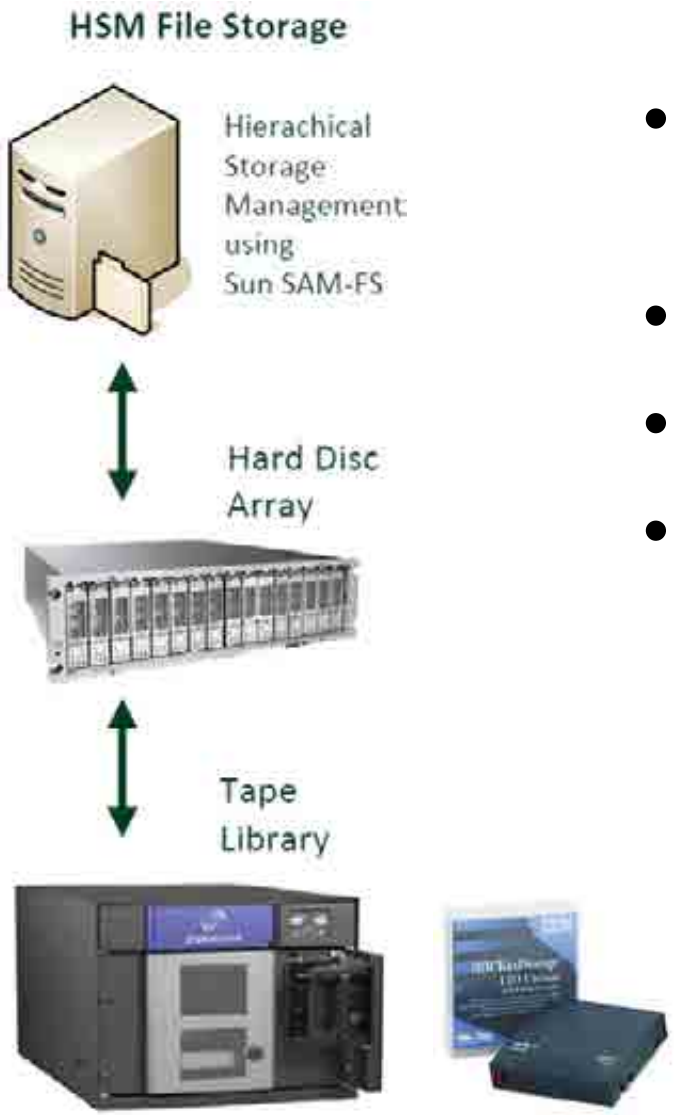


- **Data Management**
  - management of experiment and test results
  - results may be discrete values, numbers, yes/no values, text or result files; furthermore basic charts and analysis are supported

Storage Systems



Storage: Hierarchical Storage Management



- Server with special operating system Sun SAM-FS
- Arrays with hard disks
- Tape library
- Properties:
  - Fast for “new” data
  - Slow for “old” data
  - Online on demand
  - Safe (e.g. different copies à Backup)
  - (Extensible)
  - Easy to manage
  - Support different OS and protocols

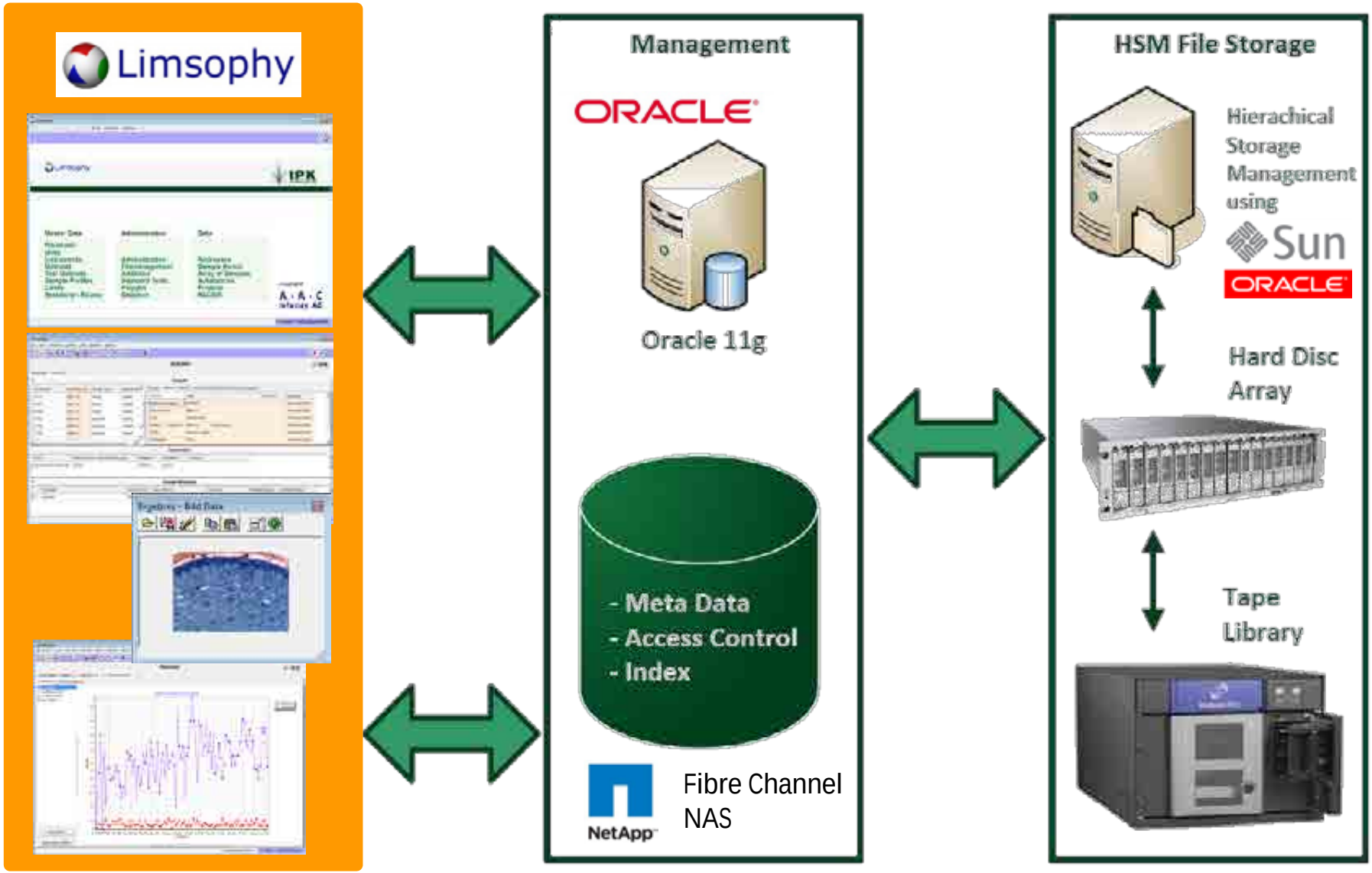
The IPK HSM - System

Some current numbers:

- Overall capacity: **263 TeraByte**:
  - **10 TeraBytes** on hard disks
  - **~253 TeraBytes** on tapes
    (allways 2 copies: à backup)

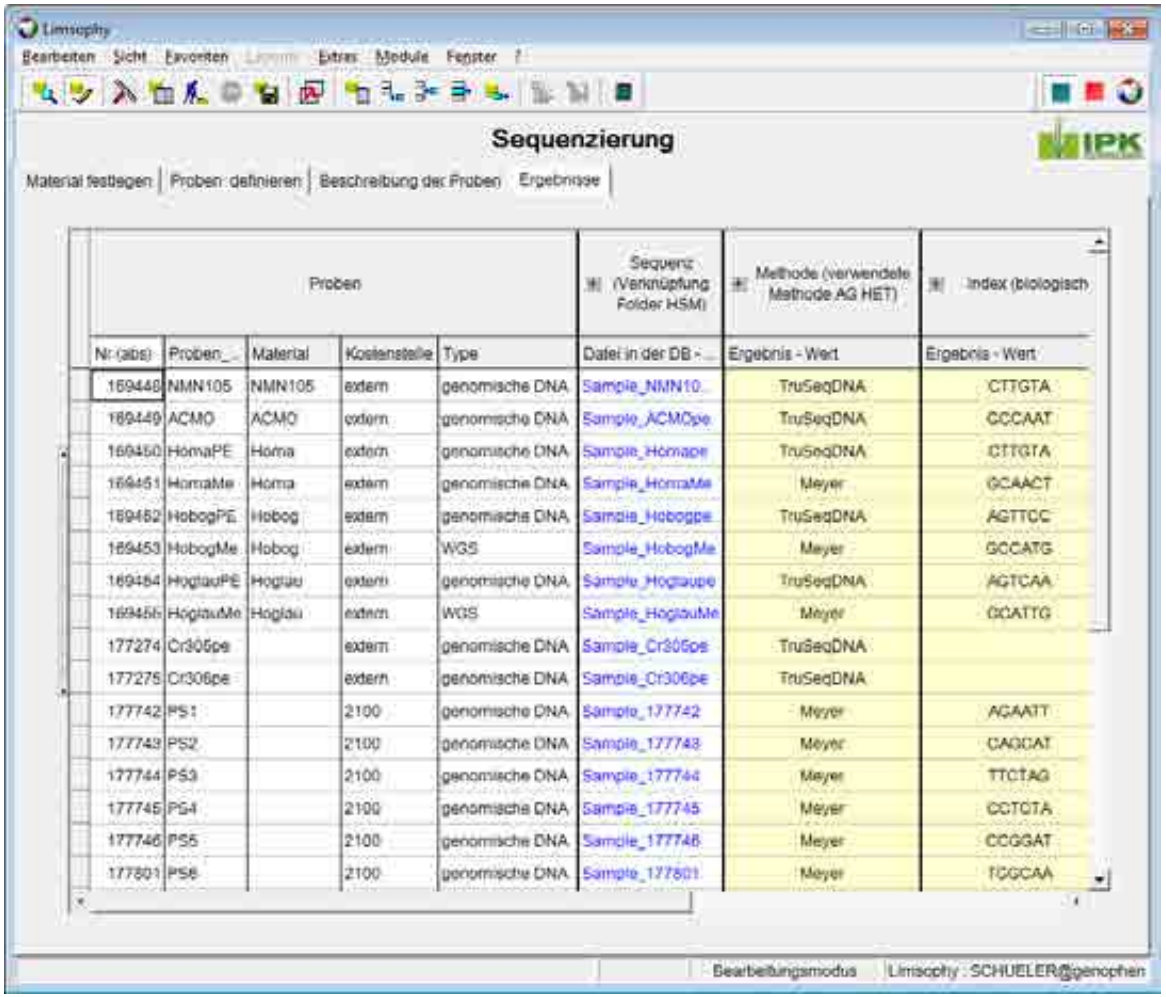
- In use:
  - Hard disks: **~7.6 TeraBytes**
  - Tapes: **approx. 110 TeraBytes**
  - In total **48%** of overall capacity

LIMS @ IPK = Limsophy



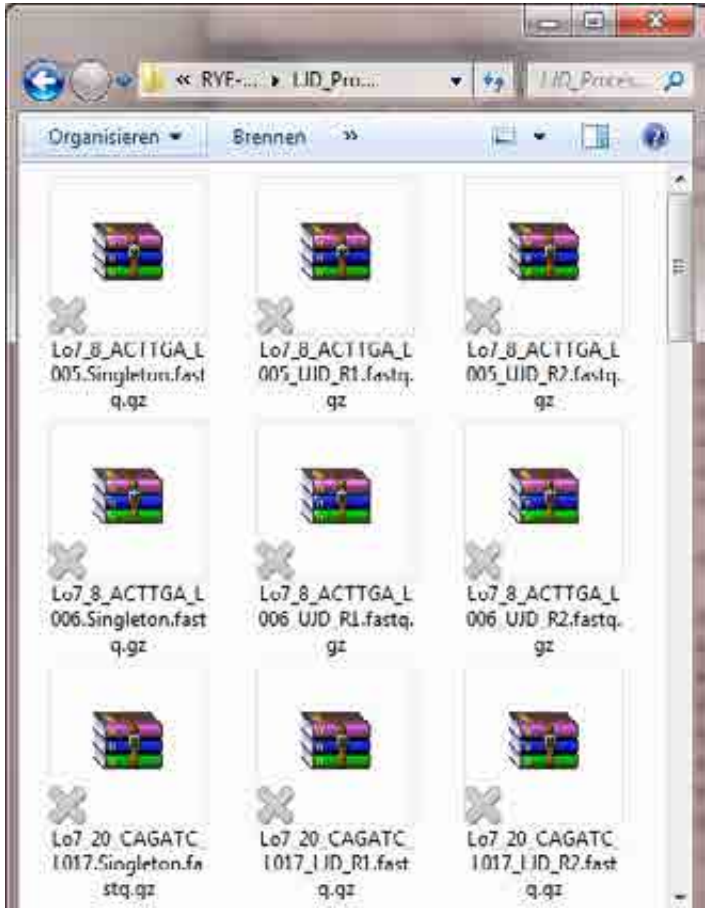
Illumina Sequencing Example

Overview about the results in LIMSOPHY:

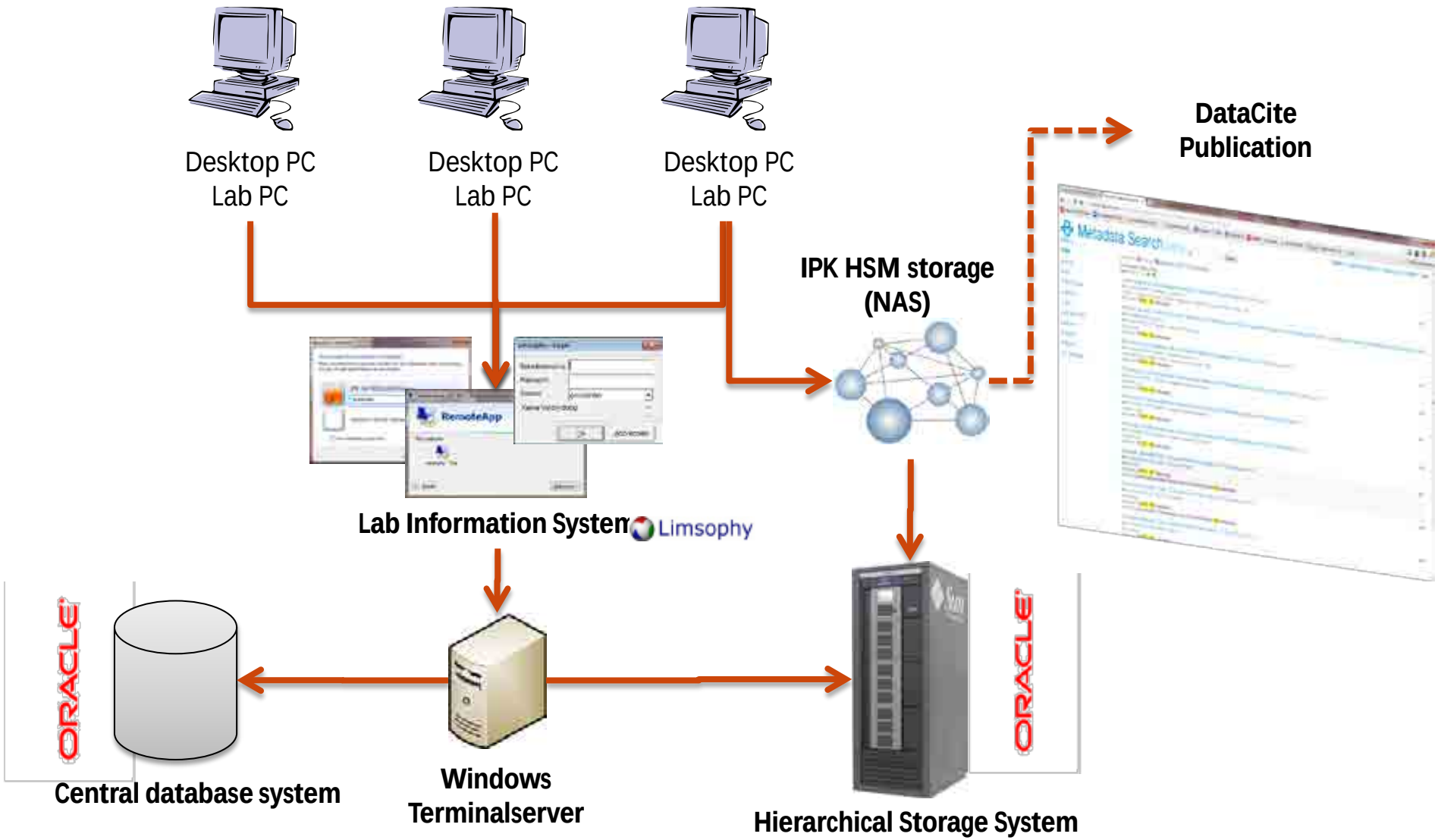


Proben					Sequenz (Verknüpfung Folder HSM)	Methode (verwendete Methode AG HET)	Index (biologisch)
Nr (abs)	Proben_...	Material	Kostenstelle	Type	Datei in der DB - ...	Ergebnis - Wert	Ergebnis - Wert
169448	NMN105	NMN105	extern	genomische DNA	Sample_NMN10...	TruSeqDNA	CTTGTA
169449	ACMO	ACMO	extern	genomische DNA	Sample_ACMOpe	TruSeqDNA	GCCAAT
169450	HomaPE	Homa	extern	genomische DNA	Sample_Homape	TruSeqDNA	CTTGTA
169451	HomaMe	Homa	extern	genomische DNA	Sample_HomaMe	Meyer	GCAACT
169452	HobogPE	Hobog	extern	genomische DNA	Sample_Hobogpe	TruSeqDNA	AGTTCC
169453	HobogMe	Hobog	extern	WGS	Sample_HobogMe	Meyer	GCCATG
169454	HogtauPE	Hogtau	extern	genomische DNA	Sample_Hogtaupe	TruSeqDNA	AGTCAA
169455	HogtauMe	Hogtau	extern	WGS	Sample_HogtauMe	Meyer	GCATTG
177274	Cr305pe		extern	genomische DNA	Sample_Cr305pe	TruSeqDNA	
177275	Cr306pe		extern	genomische DNA	Sample_Cr306pe	TruSeqDNA	
177742	PS1		2100	genomische DNA	Sample_177742	Meyer	AGAATT
177743	PS2		2100	genomische DNA	Sample_177743	Meyer	CAGCAT
177744	PS3		2100	genomische DNA	Sample_177744	Meyer	TTCTAG
177745	PS4		2100	genomische DNA	Sample_177745	Meyer	CCTCTA
177746	PS5		2100	genomische DNA	Sample_177746	Meyer	CCGGAT
177801	PS6		2100	genomische DNA	Sample_177801	Meyer	TCGCAA

File Explorer:



Data Publication (in Progress)

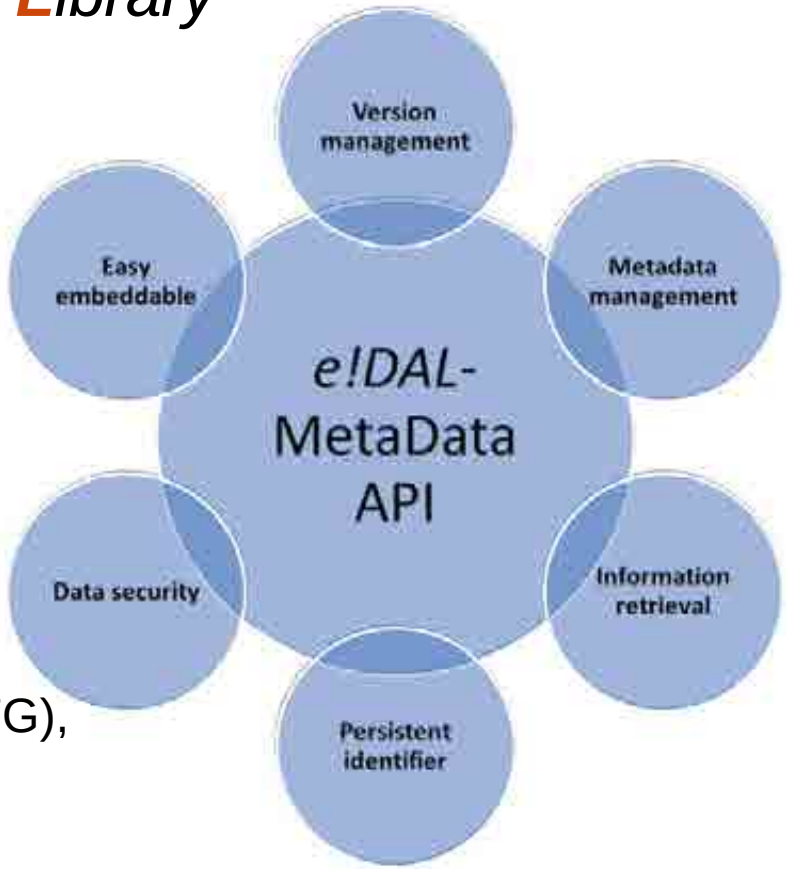


Data publication using *e!DAL*

e!DAL à ***e**lectronical **D**ata **A**rchive **L**ibrary*

Features:

- Enhanced file system like storage
- System for any kind of data
- supports for long term preservation
- Remote/local API
- Based on recommendation of DataCite, German Research Council (DFG), …

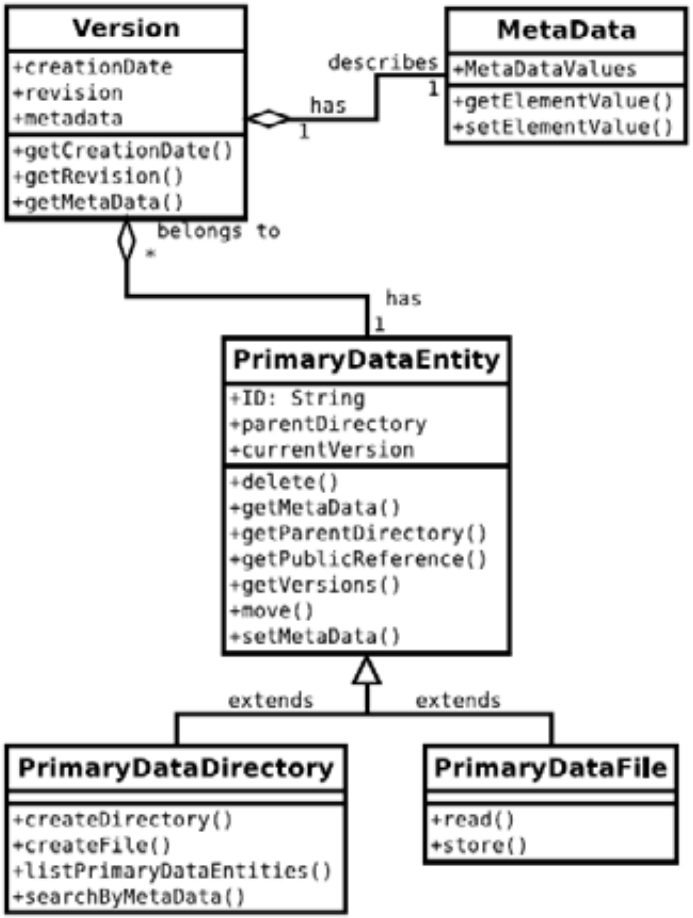


More Information:

Arend D, Lange M, Colmsee C, Flemming S, Chen J & Scholz U: **The e!DAL JAVA-API: Store, Share and Cite Primary Data in Life Sciences.** In IEEE International Conference on Bioinformatics and Biomedicine 2012; 511-515; **DOI: http://dx.doi.org/10.5447/IPK/2012/13**

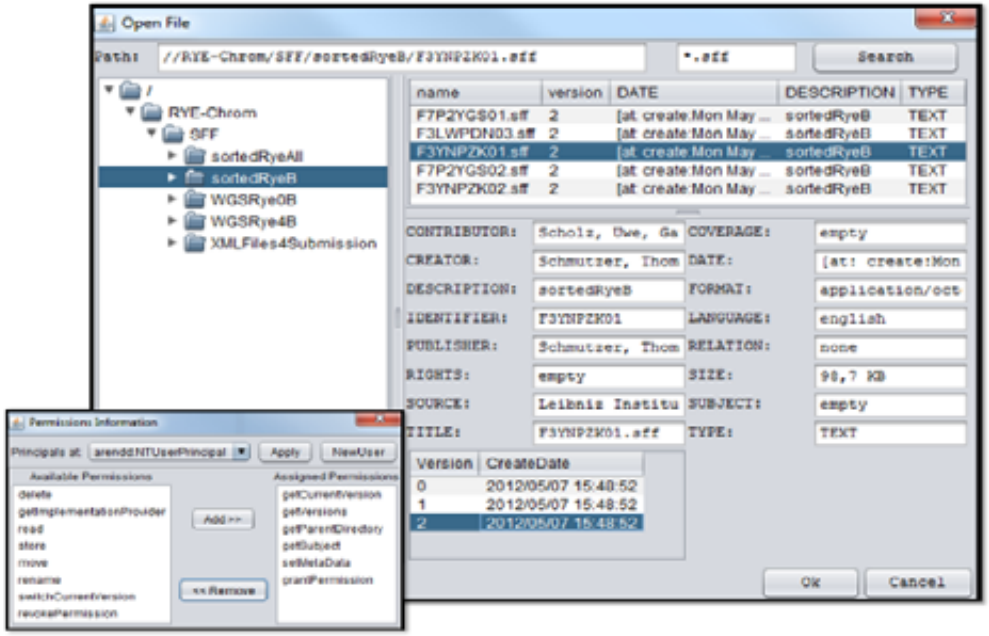
e!DAL Design

- store entities in a file system like organization
- manage different versions for every data object
- e.g. store different processing states of a dataset
- support authentication and authorization of users/groups (e.g. Windows/Unix login. . .)



software design: object oriented data structure

The e!DAL API

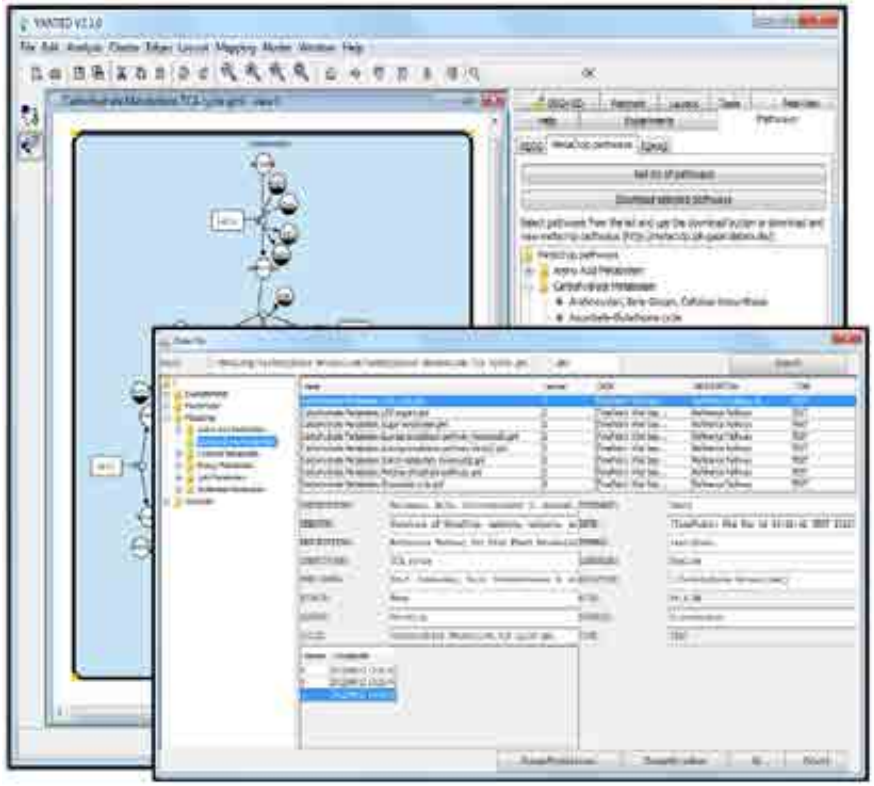


e!DAL GUI: enhanced FileChooser dialog

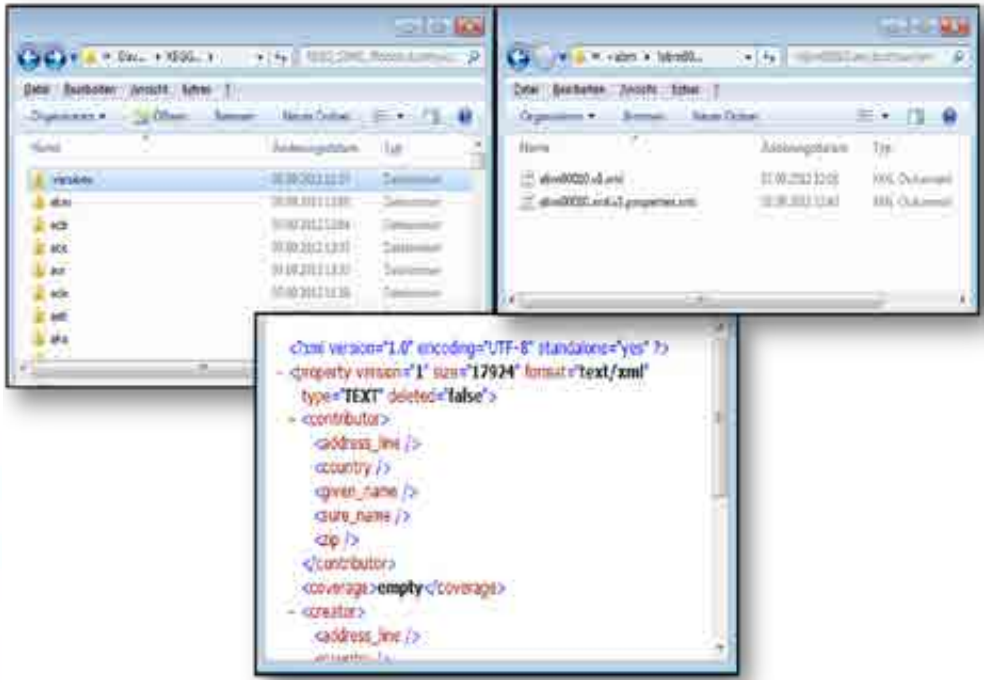
- increase data reusability and retrievability
- enrich meta data to annotate every digital object
- support *Dublin Core* standard [1] with 15 elements (e.g. CREATOR, FORMAT...)
- different data types to specify elements (e.g. Person, DataFormat...)
- support index-based file search across meta data
- fast indexing and searching (e.g. search for object name, creator...)
- additional search options (e.g. fuzzy search, wildcards...)

Embed into Application Infrastructure

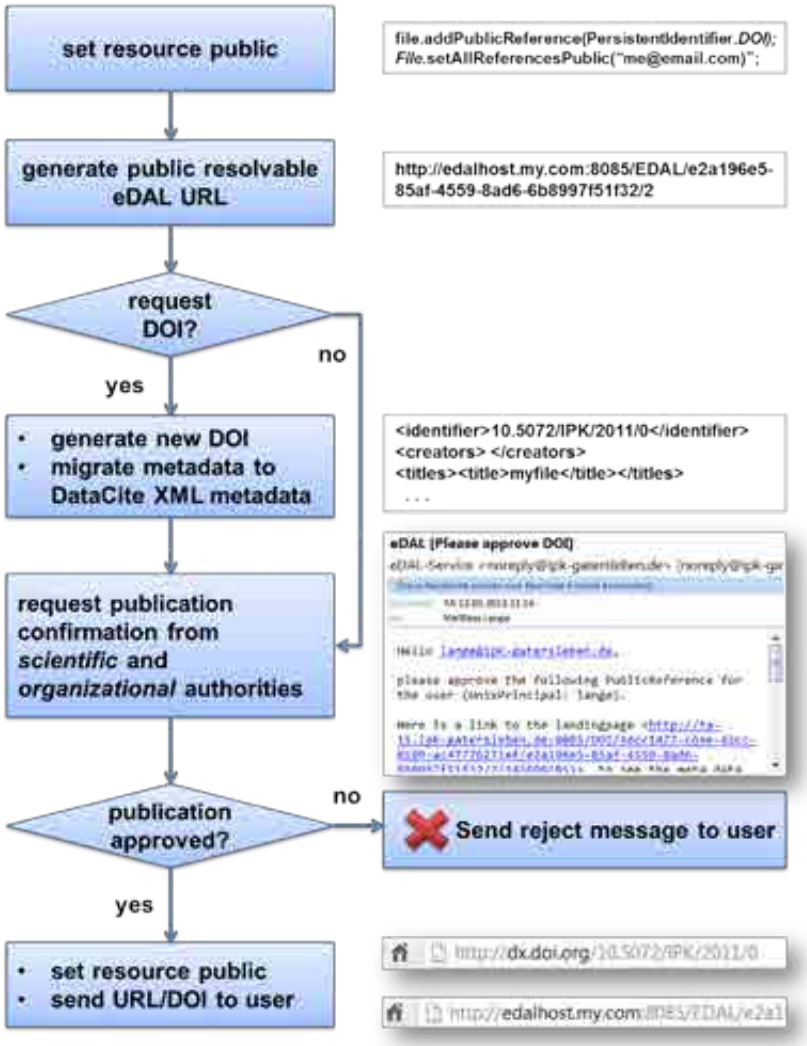
- as API into software (JAVA)

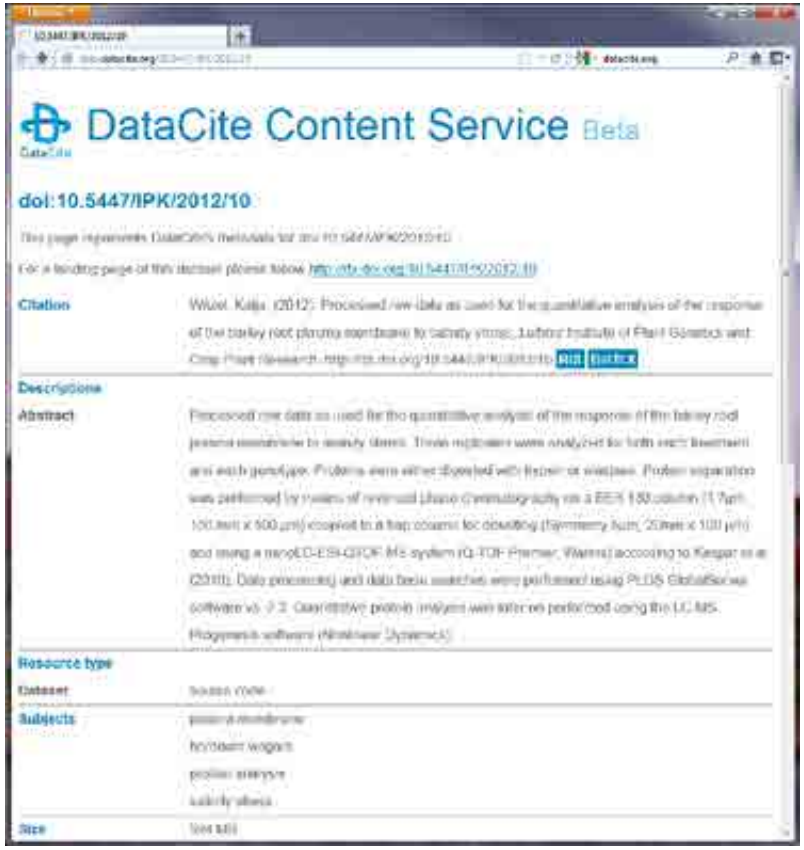


- as file system (using webDAV)



Data Publication Workflow

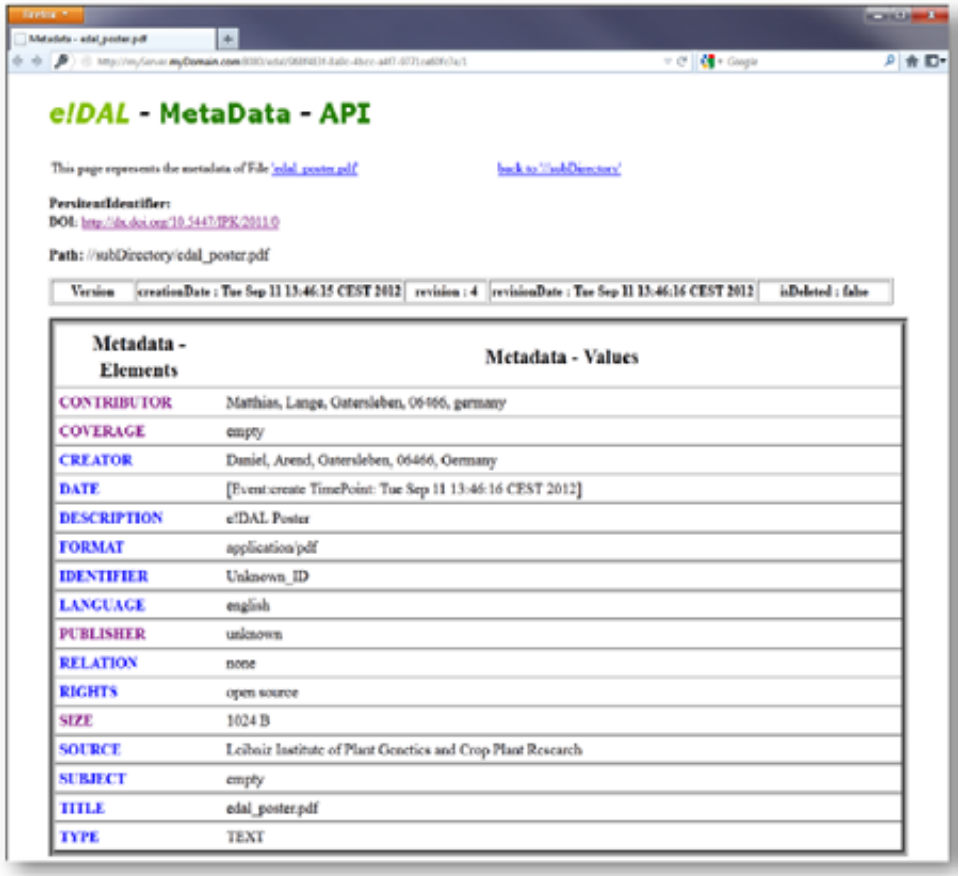




→ guarantee long-term stable
DOI: 10.5447/IPK/2012/10

Published e!DAL Data

- Interface: *PublicReference*
- registration of global persistent IDs, e.g. DOI, URN
- beta stadium
- next release provide possibility to request DOIs



e!DAL - MetaData - API

This page represents the metadata of File 'edal_poster.pdf' back to '//subDirectory'

PersitentIdentifier:
DOI: http://dx.doi.org/10.5447/IPK/2011/0

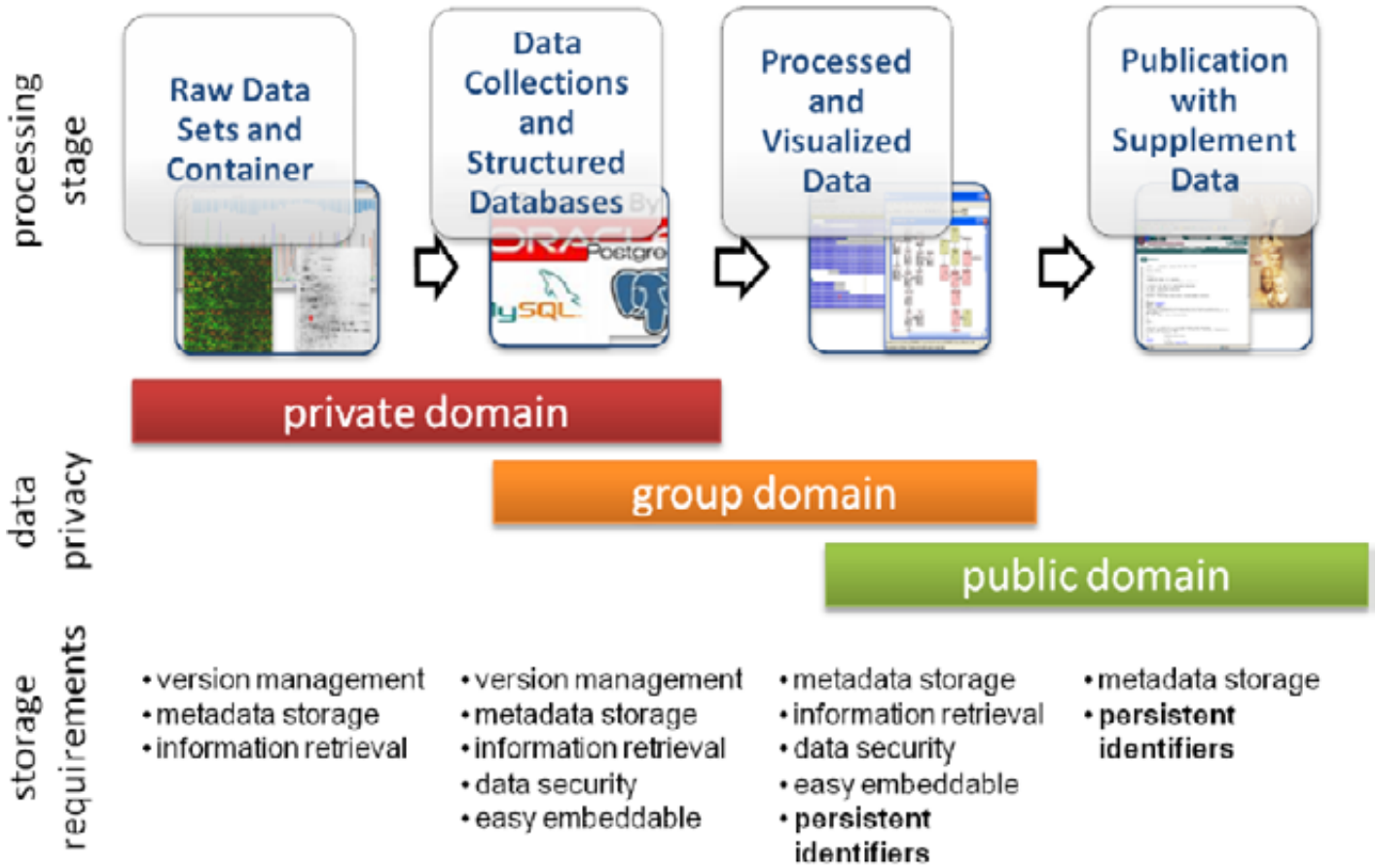
Path: //subDirectory/edal_poster.pdf

Version	creationDate : Tue Sep 11 13:46:15 CEST 2012	revision : 4	revisionDate : Tue Sep 11 13:46:16 CEST 2012	isDeleted : false

Metadata - Elements	Metadata - Values
CONTRIBUTOR	Matthias, Lange, Gatersleben, 06466, germany
COVERAGE	empty
CREATOR	Daniel, Arend, Gatersleben, 06466, Germany
DATE	[Event:create TimePoint: Tue Sep 11 13:46:16 CEST 2012]
DESCRIPTION	e!DAL Poster
FORMAT	application/pdf
IDENTIFIER	Unknown_ID
LANGUAGE	english
PUBLISHER	unknown
RELATION	none
RIGHTS	open source
SIZE	1024 B
SOURCE	Leibniz Institute of Plant Genetics and Crop Plant Research
SUBJECT	empty
TITLE	edal_poster.pdf
TYPE	TEXT

generated landing page: URL's for DOI resolver

Summary: 3-Tier Storage



Conclusions

- Use commercial software and industry standards with in-house maintenance and close support contracts!
- Minimize the in-house self developments!
- Nevertheless the implementation of an efficient data management pipeline is specific task for each institution!
- An institution policy is essential!
  - What data should stored?
  - How the primary data should stored
  - How long the data should stored?
- Use LIMS for management of meta data / project data! **IPK:** Limsophy
- Use database management systems! **IPK:** ORACLE
- Separation of meta data and primary data in data bases and storage solutions **IPK:** Sun
- Support primary data publication! **IPK:** DataCite *e!DAL*

Acknowledgements

IPK
Bioinformatics and Information Technology (BIT):

- Jinbo Chen
- Christian Colmsee
- Steffen Flemming
- Uwe Scholz
- Heiko Miehe
- Thomas Münch
- Daniel

Plant Bioinformatics (PBI):

- Anja Hartmann
- Tobias Czauderna

This work was supported by the European Commission within its 7th Framework Program, under the thematic area “Infrastructures", contract number 283496.





Thank you for your attention!