Data Management Experiences and Best Practices from the Perspective of a Plant Research Institute

Daniel Arend, Christian Colmsee, Helmut Knüpffer, Markus Oppermann, Uwe Scholz, Danuta Schüler, Stephan Weise and Matthias Lange

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben

July 17, 2013

Outline

Leibniz Institute of Plant Genetics and Crop Plant Research

IPK Gatersleben

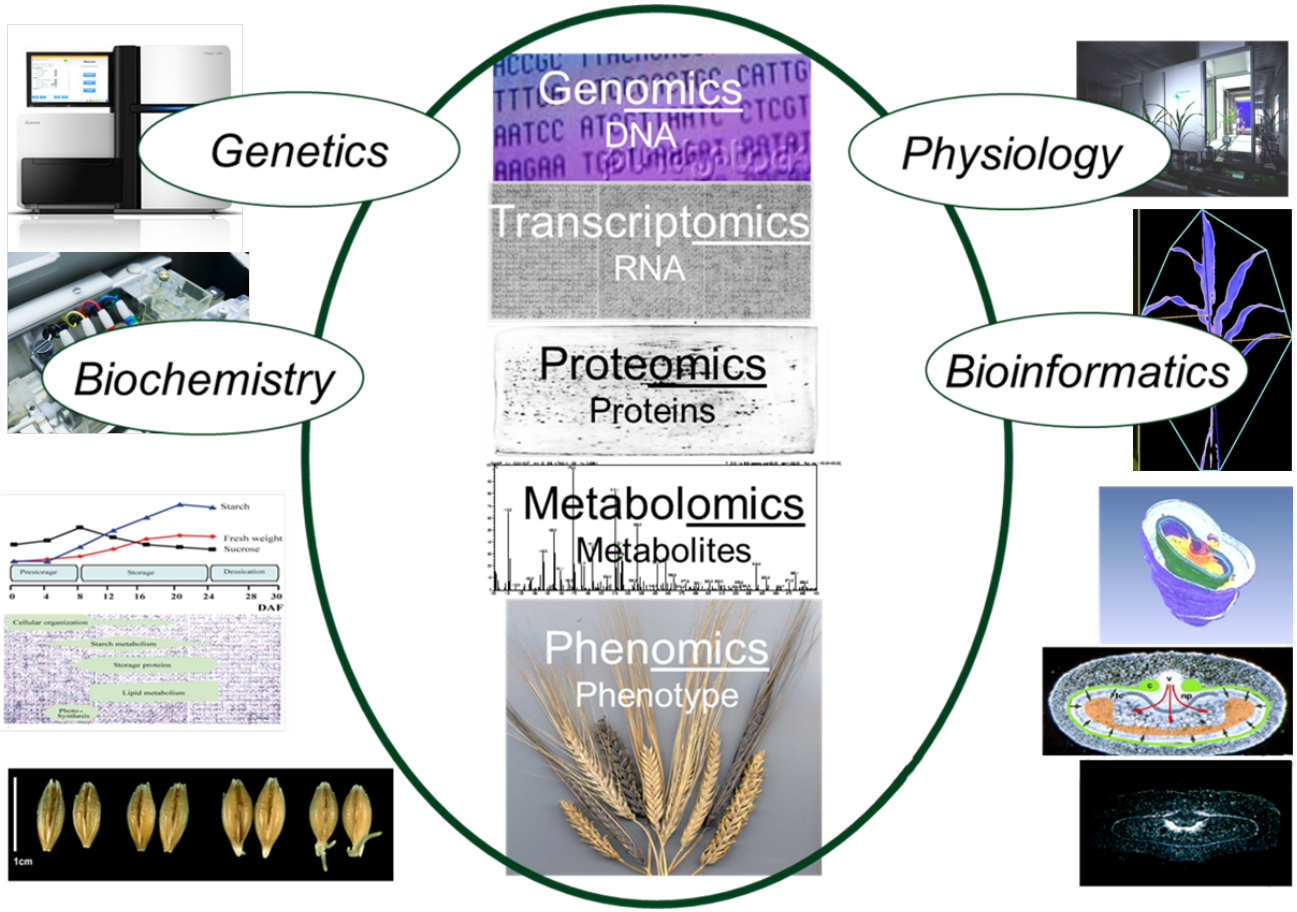
- over 70 years tradition
- Federal ex-situ Genebank of Agricultural & Horticultural Crop Species (150.000 accessions)
- source of the german breeding industry
- total staff: $\sim$ 550
- scientists: $\sim$ 200 (10 Bioinformaticians)
- about 30 research groups

Bioinformatic Research Topics

- databases & information retrieval
- sequence-, network- & image-analysis
- big data management

$\rightarrow$ **crop plants have multiple purposes: food, feed or bioenergy source**

Data Domains at the IPK Gatersleben

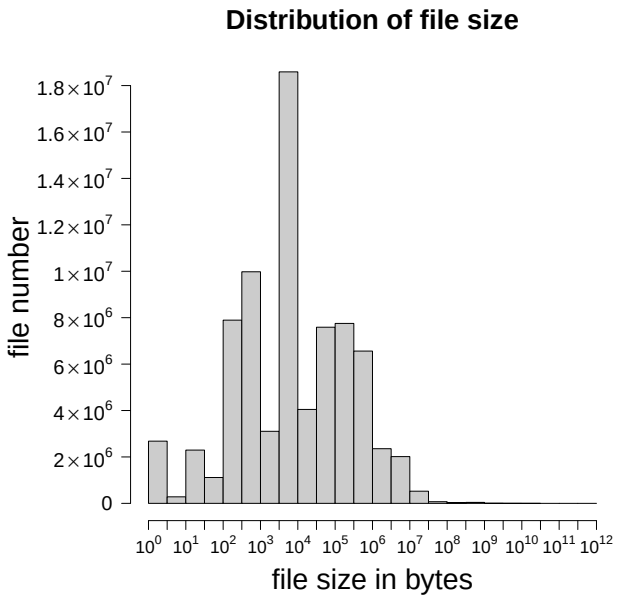


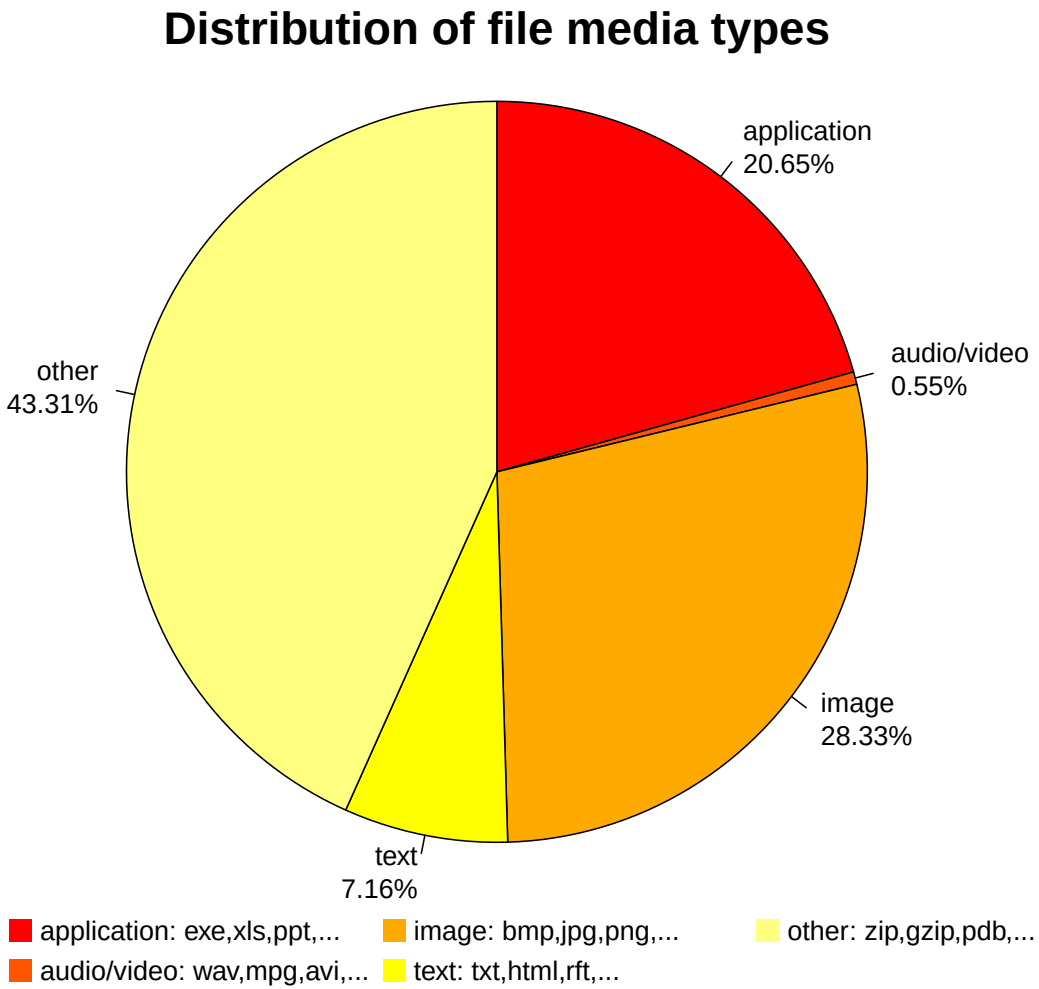
- life sciences produce a huge data volume (primary data)
- e.g. 'Next-Generation Sequencing', 'Plant Phenotyping', 'System Biology'...
- basis for modern research & publication process[1]

[1] Craddock et al. *Nature Review Microbiology*, 2008

Data Storage at the IPK Gatersleben

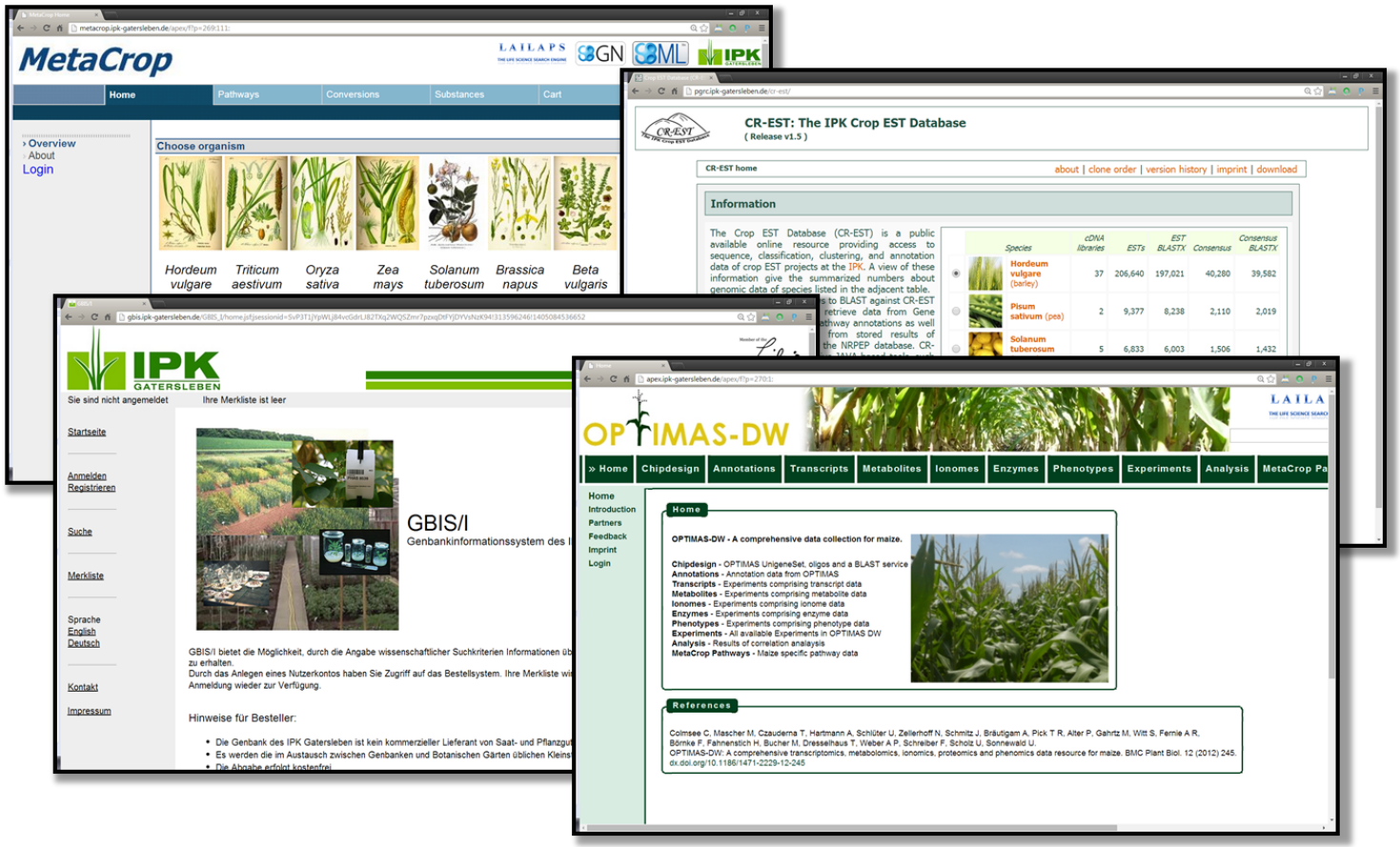
- huge amount of heterogeneous data
- $\sim$ 80 million data files (230 TB) stored on a HSM
- over 600 different file types
- most of the files $\leq$ 1 MB
- file size up to 600 GB





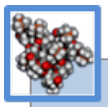
IPK Database and Information Systems

- over 20 different project-specific databases & information systems
- specialized on different research fields & data domains
- every system uses its own database schema & storage backend
- heterogeneous development & maintenance



Data Management Study

- initiated by IPK directors board in 2010
- requirement analysis for an institute-wide Laboratory Information Management System (LIMS)
- suggest general policies for sustainable data management



- **standardized vocabulary**
- **version management**
- **project planning & documentation**
- **information retrieval**
- curating process
- auditing
- data security
- plan & control operational procedures

experiments

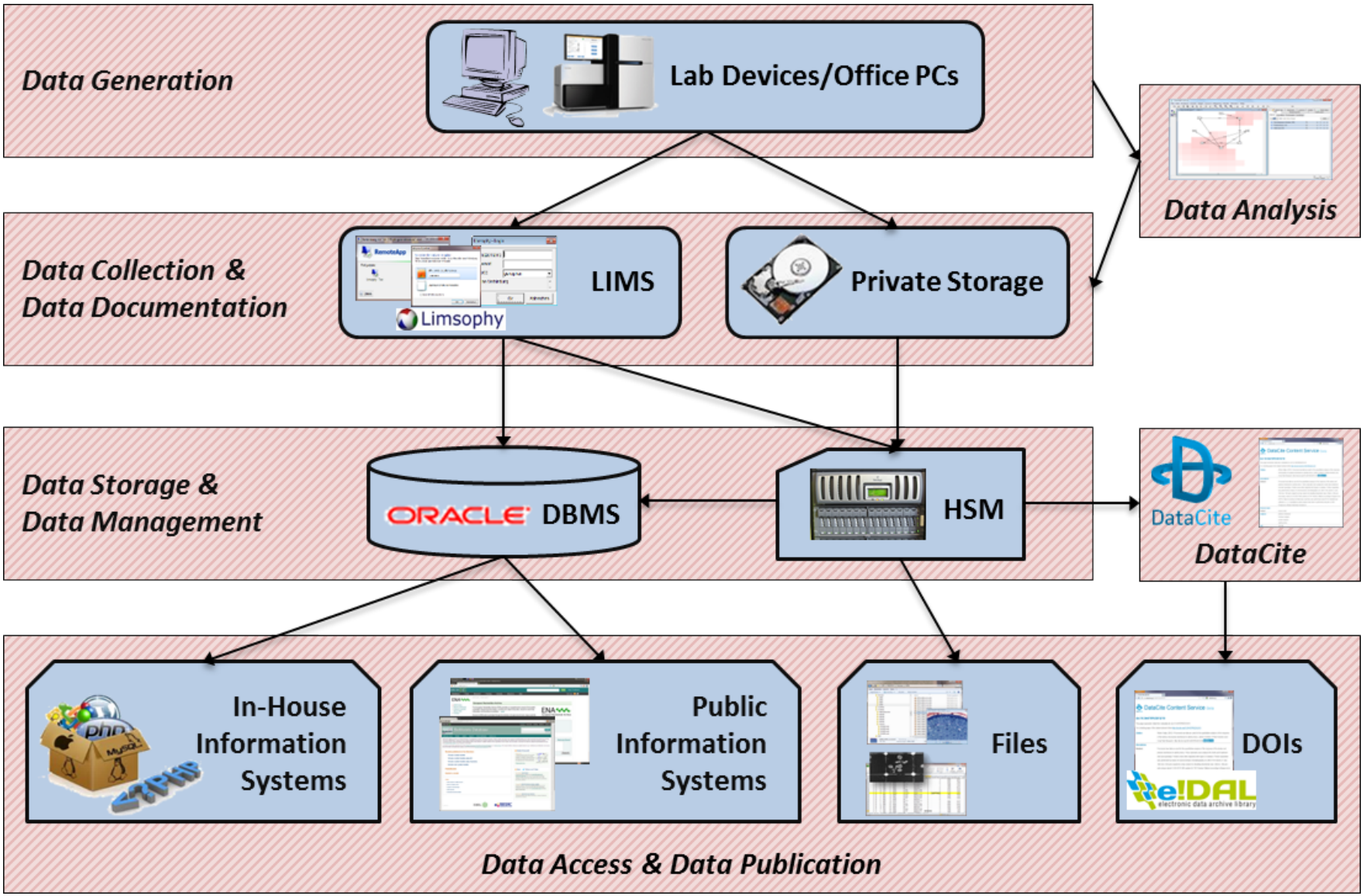
- **direct access on measurement devices**
- **central storage of primary data & central database**
- data import/export interface
- train and support by LIMS producer

technologies

- **connect to central user management**
- **basic data management (devices, rooms, persons, material)**
- **user acceptance**
- adapt user interface on processes & integration in data domain
- modeling of departments
- internal & external costing department
- calculate consumption material

management

Strategic Realignment



Laboratory Information and Management System

LIMSOPHY (AAC Infotray)

- selected commercial product → no open source
- introduced at the IPK in 2011
- long term support contract

\+ flexible & expandable

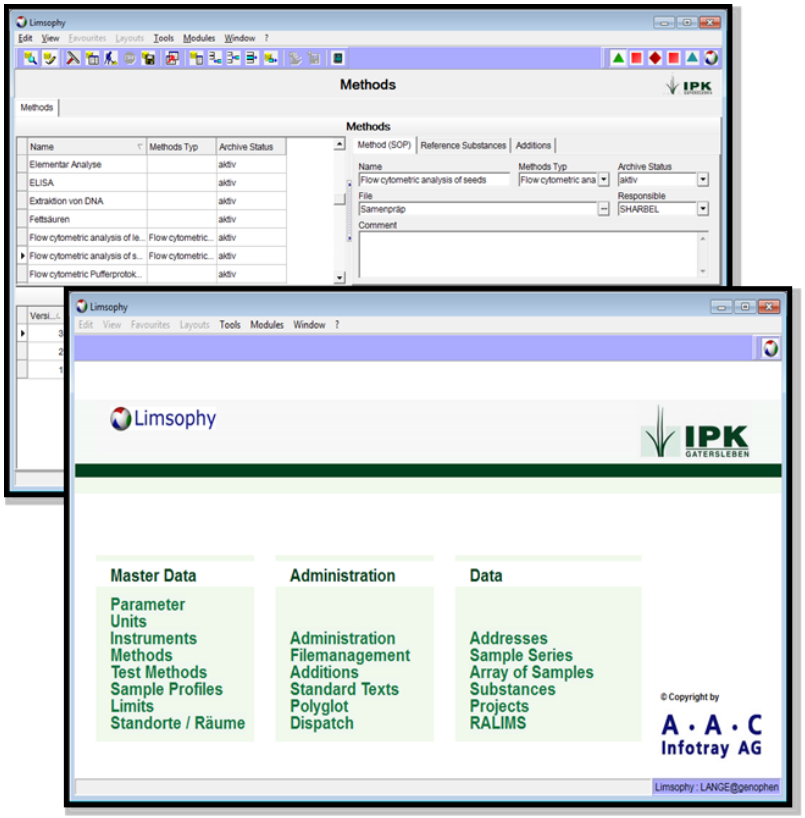
\+ additional external support

\+ less training for scientists & technicians (different computer skills)

\+ embedded into central storage backend

\- developing of new modules & user acceptance for integration of lab workflows

\- no out-of-the-box module for data publication (→ see e!DAL poster)



Summary - Strategic Realignment

Organisational Actions

- build central bioinformatics service group with a scientific administration
- financed by research funds & institutional budget
- core-financed service team for LIMS
- inter-departmental coordination of bioinformatics research

Infrastructure Actions

- central storage systems & databases
- combining in-house & public data publication systems

Summary - Some Impressions

- LIMS is used by over 40 project and 10 research groups
  - 4.100 substances & 1.400 experiments stored
  - 1.200.000 measurements & 400.000 linked files
  - goverment verified GMO managment
  - manage complete services processes, e.g. NGS sequencing
- different new web information systems
  e.g. management of chemicals, photo-archive...
  $\rightarrow$ design new system within 2 days, without experiences in programming
- already 30 manually registrated DOIs over DataCite
  $\rightarrow$ in future automatic process using e!DAL

Acknowledgment

Bioinformatics and Information Technology (BIT):

- Danuta Schüler
- Matthias Lange
- Christian Colmsee
- Uwe Scholz

Genbank Documentation: (GED):

- Stephan Weise
- Markus Oppermann
- Helmut Knüpffer

This work was performed within the German-Plant-Phenotyping Network, which is funded by the German Federal Ministry of Education and Research (project identification number: 031A053)

Thank you for your attention

http://www.ipk-gatersleben.de