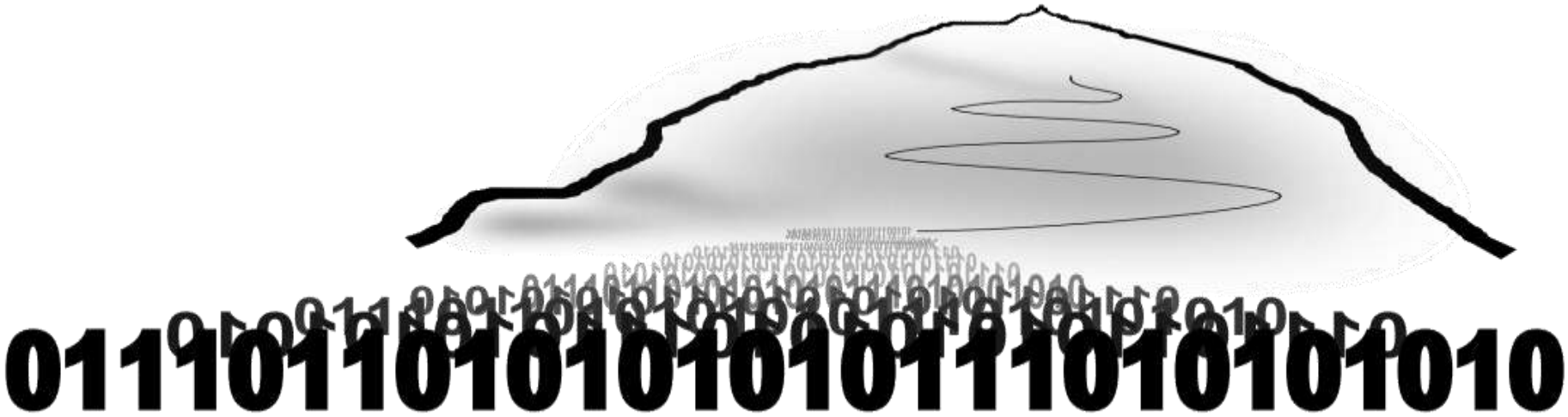
Karlsruhe Institute of Technology

Die Large Scale Data Facility des KIT

Rainer Stotzka, Jos van Wezel

Institute for Data Processing and Electronics

In close collaboration with:
Steinbuch Centre for Computing
Institute of Toxicology and Genetics
Institute for Applied Computer Science



Wissenschaftliche Experimente erzeugen Daten

USCT
zur Brustkrebsdiagnose

0,3 **Peta**Byte
pro Jahr

ANKA
Synchrotron-
Strahlenquelle

1 **Peta**Byte
pro Jahr

Robotische **Mikroskopie**
in der Biologie

1 **Peta**Byte
pro Jahr

… und viele andere mehr …
(FAIR, X-FEL, ITER, KATRIN, …)

? **Peta**Byte
pro Jahr

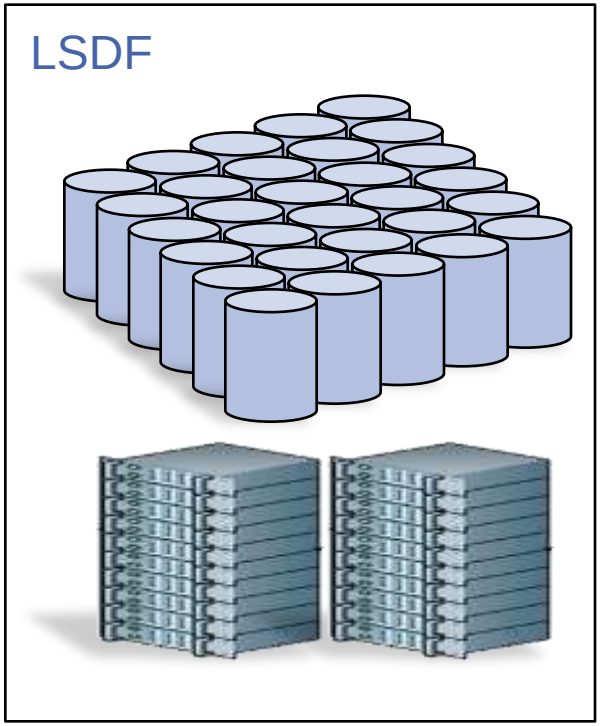
**WOHIN,
WIE
???**

Programm Supercomputing

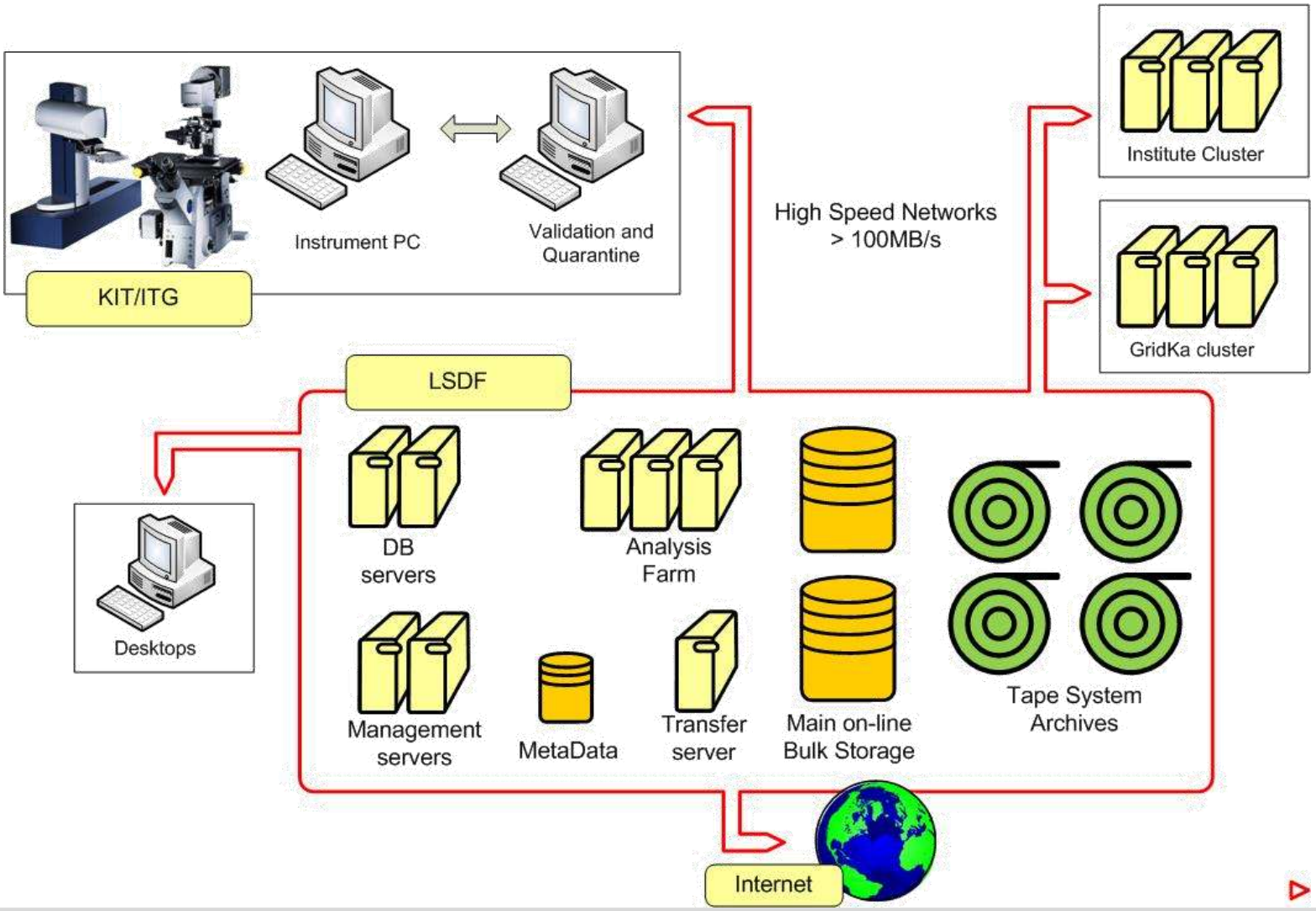
Die **Large Scale Data Facility** ist eine Einrichtung des KIT in der HGF mit dem Ziel, das bedeutendste Zentrum zur Speicherung und Analyse von Experimentdaten in Europa zu werden.

- 6 PetaByte
- 58 Knoten mit je
  - 8 Prozessor-Kernen
  - 36 GigaByte Speicher
  - 1 TeraByte Festplatte
- 10 GigaBit/s Netzwerke

(Stand 2010)



Internal stucture



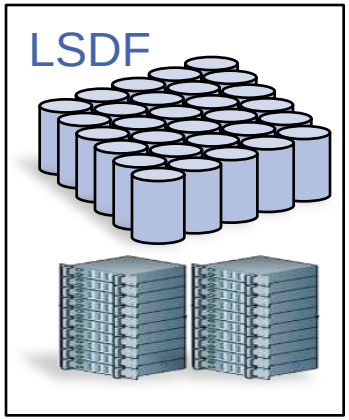
LSDF objectives (from the user‘s point of view)

- Objectives:
  - Storage
    - Dedicated for science data
    - ExaByte scale data
    - To archive data, long term sustainability (10 yrs. – ?)
  - Interactivity
    - To enable scientists to gain better scientific results by providing
      - Data intensive analysis
      - Added value services for data intensive processing
    - To provide high performance access, high throughput
    - “Barrier free” access (easy-to-use)

- Conflicting objectives:
  - Archive – high throughput
  - Enhanced functionality – simplicity

The LSDF strikes a new path

- Scientists dream: as simple as a USB hard disk



LSDF is more complex:

- ExaByte scale
- Distributed composite of various technologies:
  hard disks, IBM-systems, networks, Hadoop, …
  → internal dynamics caused by replacing or adding new components: disks, archives, techs, etc.
  → sea of data with internal flow
- Added value services,
  e.g. automatic processing of new data
  → additional information
- Security: worldwide access

Why is meta data necessary?

Meta data describe the contents of data

- Everybody uses meta data:
  - File name and extension
    (e.g. `rainer.jpg, budget.xls, Readme.doc`)
  - Location
    (e.g. `/…/EU-projects/2010/Fishy/budget.xls`)
  - Personal know-how

→ Sufficient for small file systems

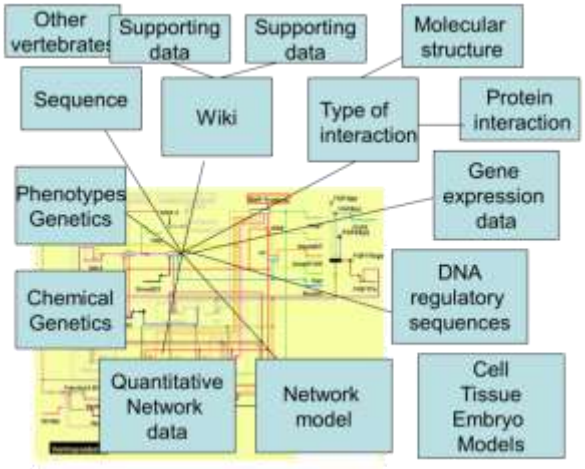
Have you ever tried to locate a file or info-somewhere-in-a-file-system

- 15 years old ?
- in the file system of a colleague ?
- in a 100 PetaByte file system ?

PANIC ?

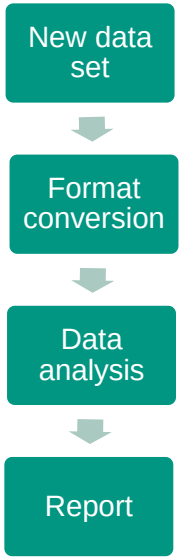
Applications requiring meta data

- Data archiving and retrieval (libraries)
- Fusion of complex data from various sources (data integration)



Community-specific services:

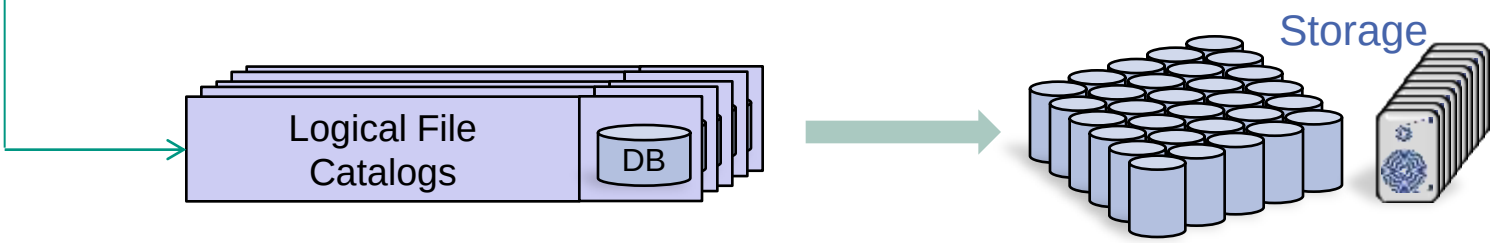
- Automatic processing
  (e.g. automatic analysis starts when data appears)
- Analysis chains
  (reporting analysis workflow, results and errors)
- Google and Yacy
- Etc.



Model of the LSDF meta data management

Idea:
Clear separation between

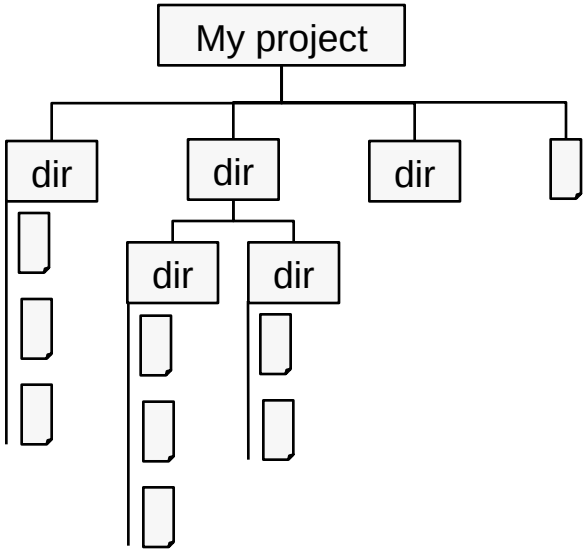
- Data (files),

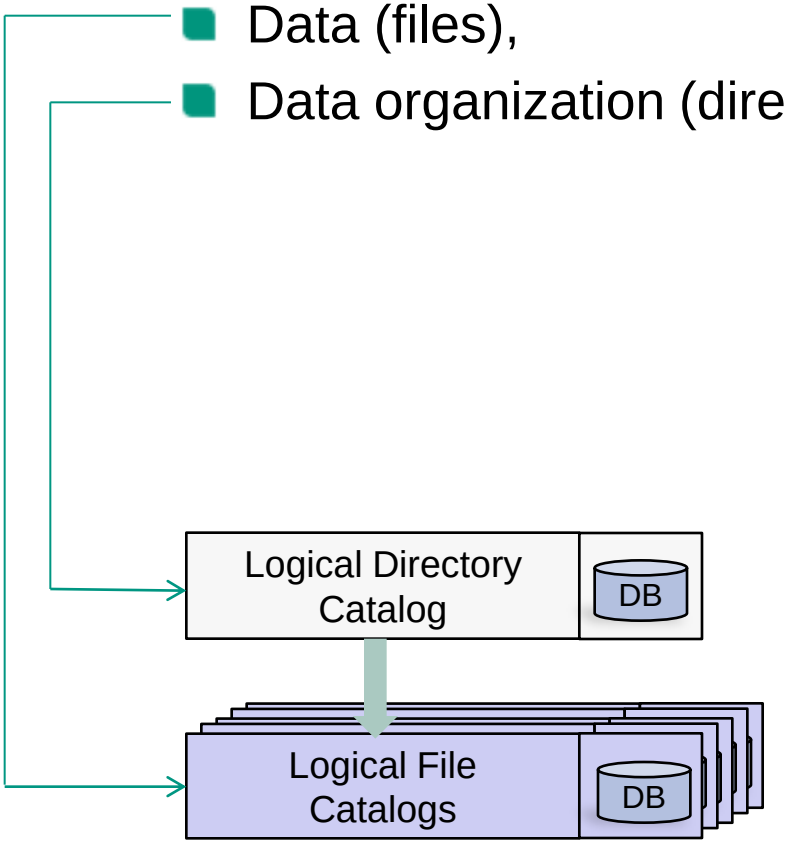


Model of the LSDF meta data management

Idea:
Clear separation between

- Data (files),
- Data organization (directory structure)

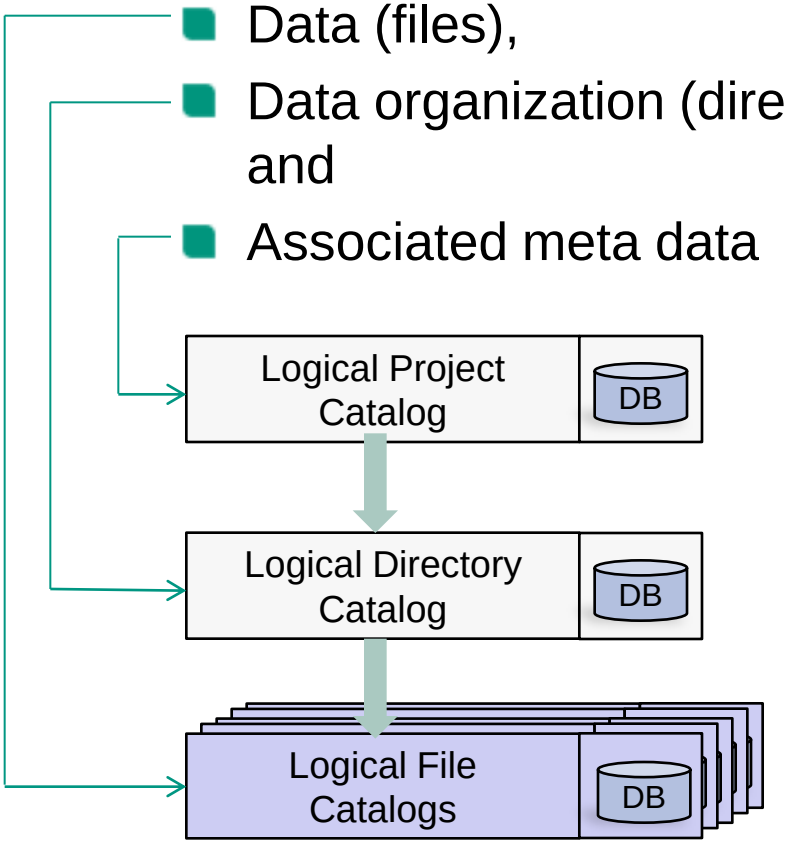




Model of the LSDF meta data management

Idea:
Clear separation between

- Data (files),
- Data organization (directory structure) and
- Associated meta data



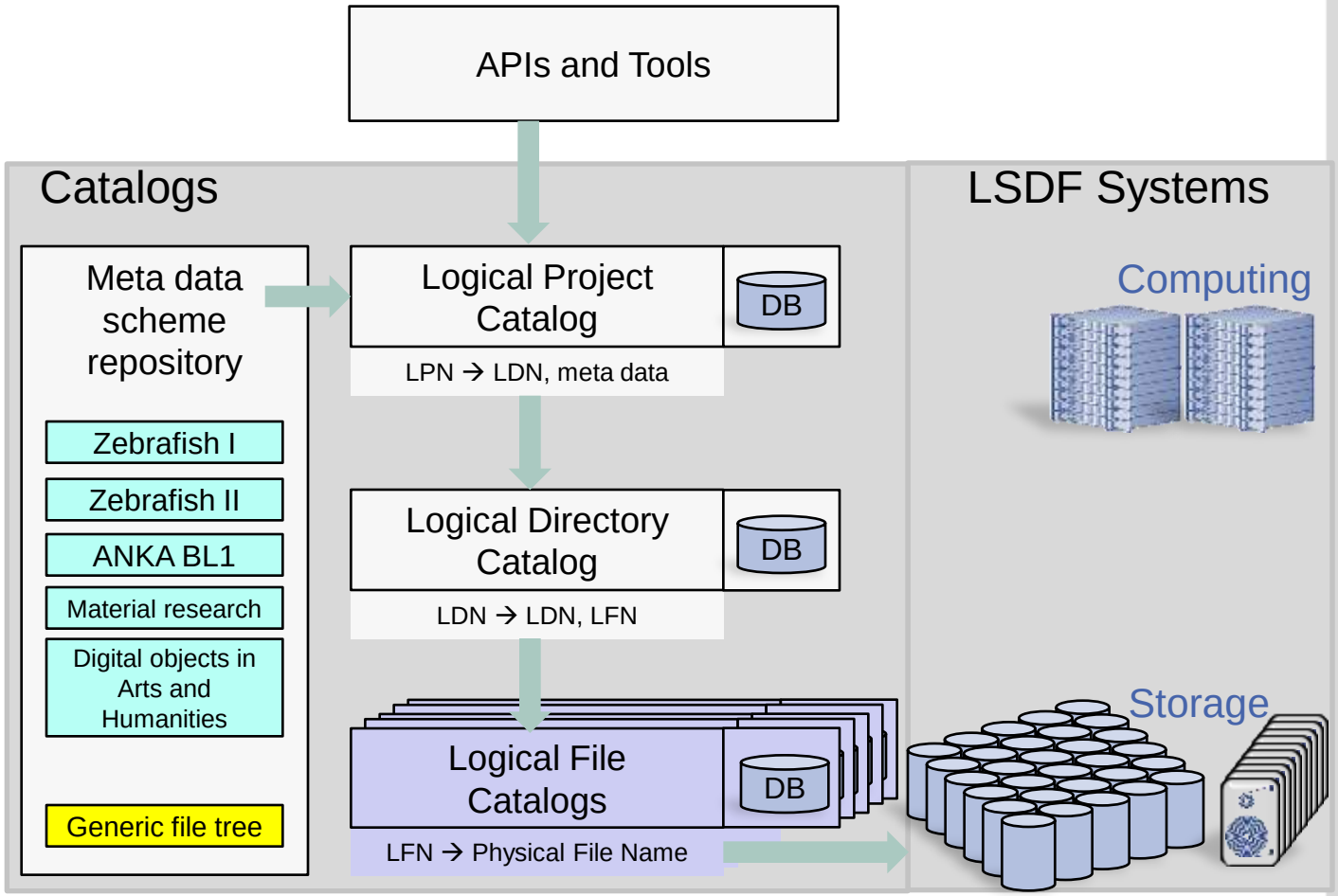
- name
- owners
- access rights
- date
  - community
  - (sub)subcommunity
    - measurement type
    - device, instrument
    - …

Meta data structure depends on project, instruments, **time**, …

Hierarchical Catalog System

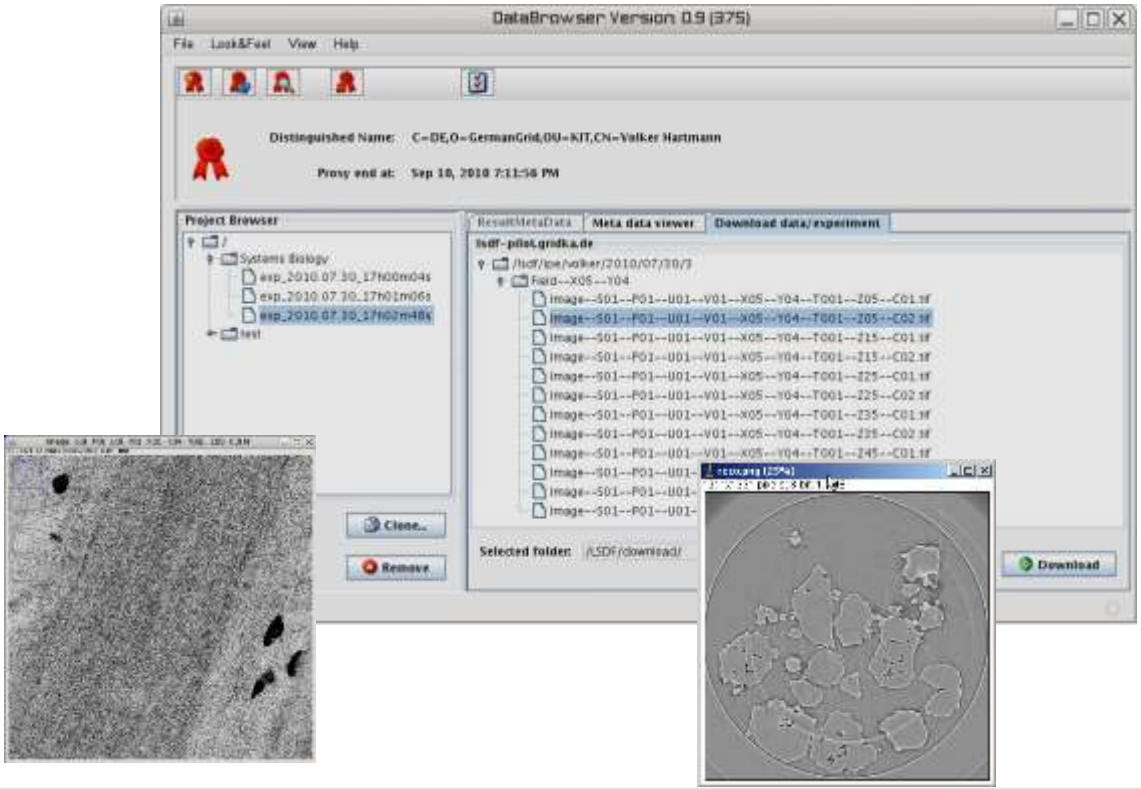
- Sustainable
- Easily extensible
- Independent of data formats
- Enhanced performance: distribution of access
- Safety by redundancy

- Easy-to-use?



How to handle the complexity?

- *Apparently more complex: how do I use it?*
  - → Simple access tools, which can be easily adapted to your specific needs
    - → **Data Browser** is a File-, Data- and Project-Explorer



Data Browser allows:

- Authentication
- Project and file browsing
- Upload
- Download
- Edit meta data
- Data visualization
- Control data analysis

Features:

- Extensible
- Huge variety of communication protocols
- Open source

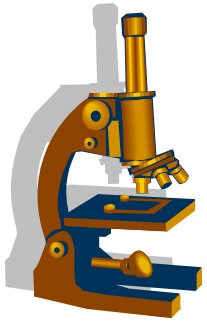
How to handle the complexity?

- *How do I insert a new scientific project ?*
  - → Data and meta data organization experts for projects with specific needs
  - → Generic meta data format for simple file trees

- *How do I transfer my data to a different location? Do I loose my meta data?*
  - → Import-export to standard data and meta data formats
  - → Archive-in-a-box (Web installer or DVD, zip-archive, etc.)

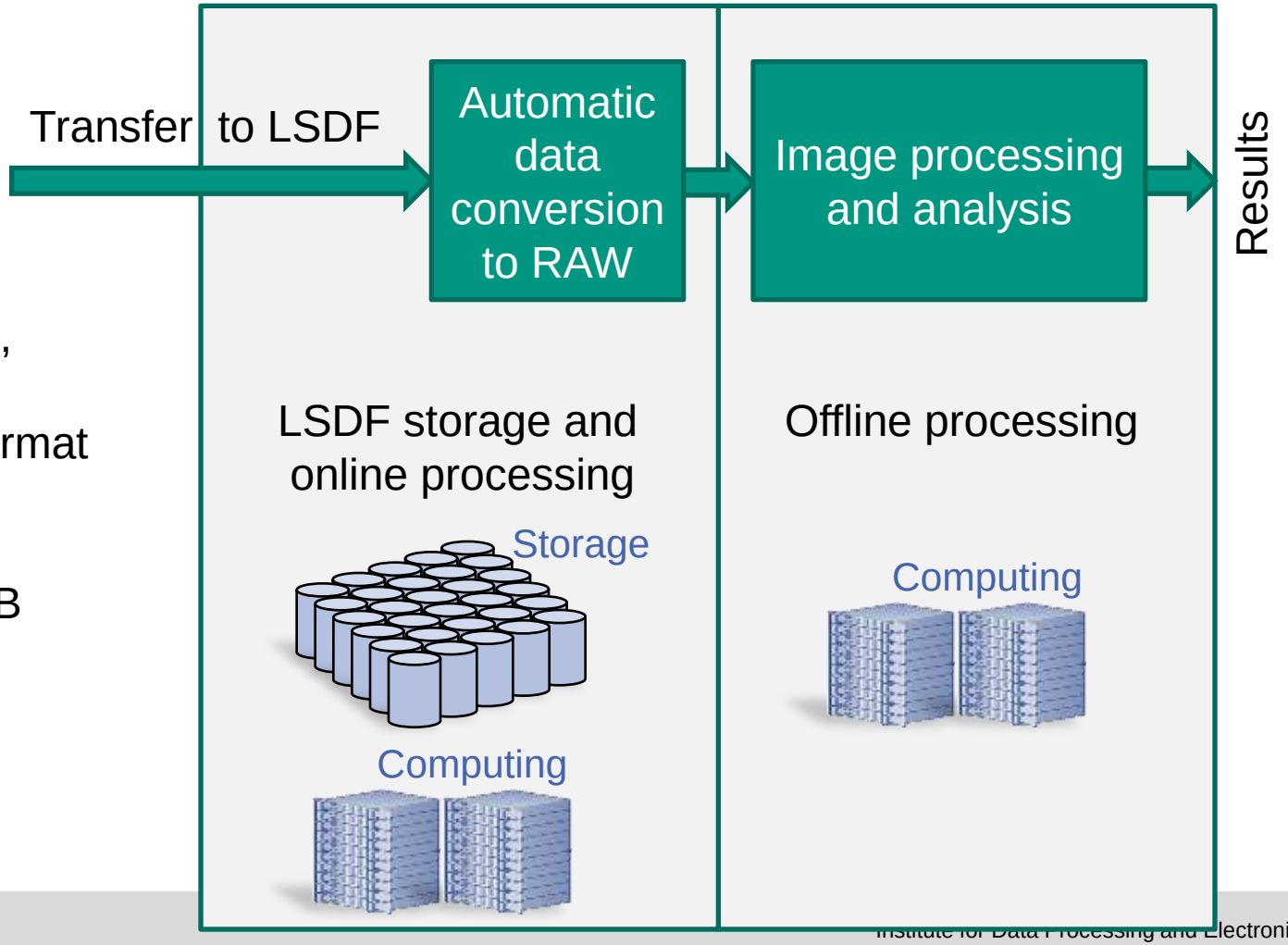
Example: ITG Vertebrate Development

- Complex image analysis chain:



3D image stack,
time series,
Leica Image Format

data set size:
100 GB



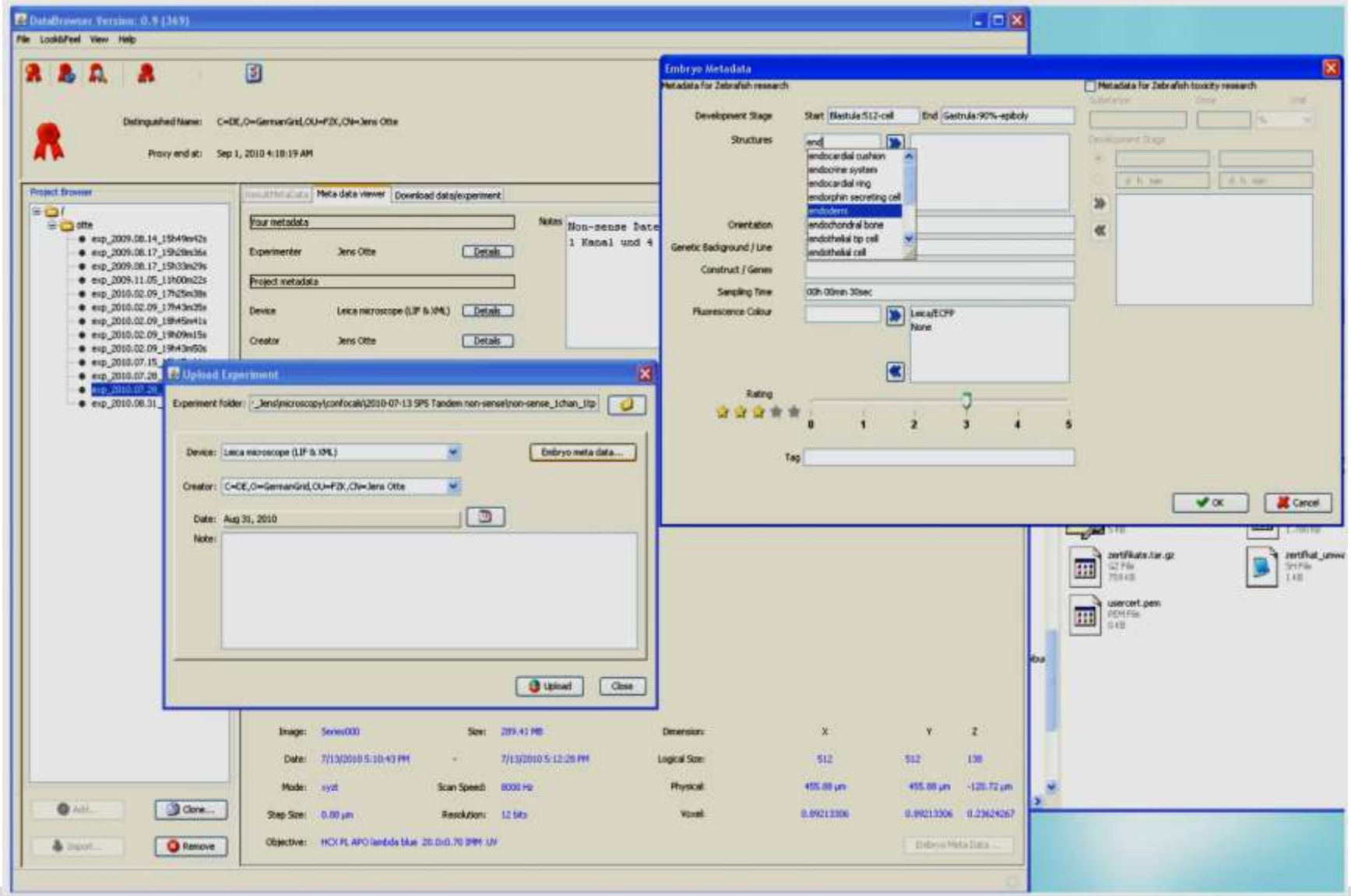
Example: ITG Vertebrate Development

- Close cooperation ITG, IAI, SCC and IPE
  (Thanks to Jens C. Otte for the images)

Data Browser:

- Meta data organization
- Adapted Data Browser implementation
- Automatic data conversion workflow at LSDF steered by meta data

Example: ITG adapted DataBrowser



Scientific communities

- Systems biology (ITG, BioQuant, Immunogenetics)
  - Vertebrate development studies and
  - Deconvolution
- Synchroton facilities and beamlines
  - ANKA data storage
  - HGF “High Data Rate Initiative”
- Climate research
- Material research
- Arts and humanities

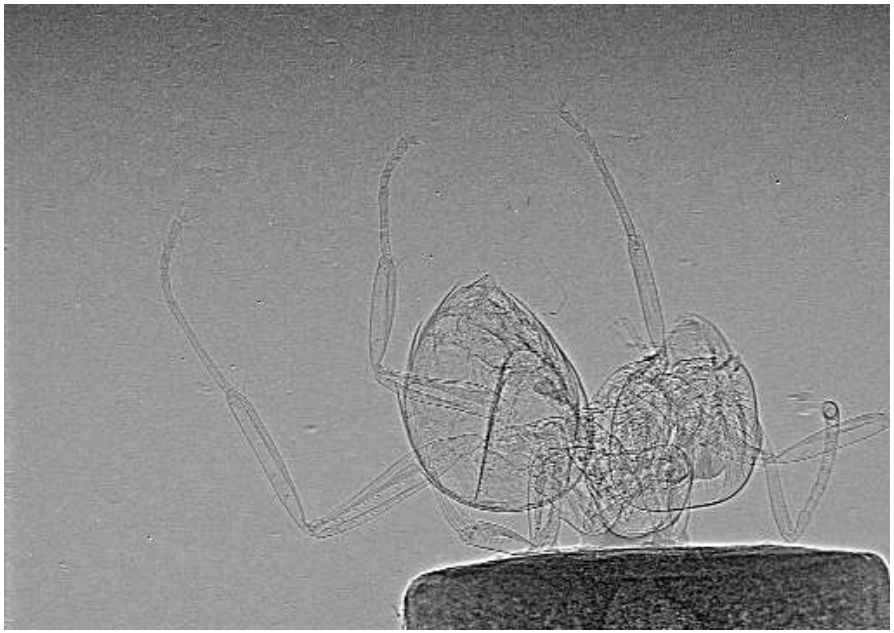
»Il Cenacolo« von Da Vinci (1494-98)

»L’ultima cena« von Julius Romanus (1754)

Data intensive science

- Remote instrumentation
  - Targeted at scientific instruments
  - Grid enabled e-infrastructure, distributed
  - Virtual control room
  - OGF RISGE-RG

- Algorithms for data analysis
- Visualization of huge 3D data sets: online visualization of 500 GB data sets



Conclusions

- LSDF is a powerful structure
- Designed for future requirements

LSDF offers

- Sustainability
- Flexibility
- Interactivity
- Community-specific services
- Support

→ To gain faster and better scientific results