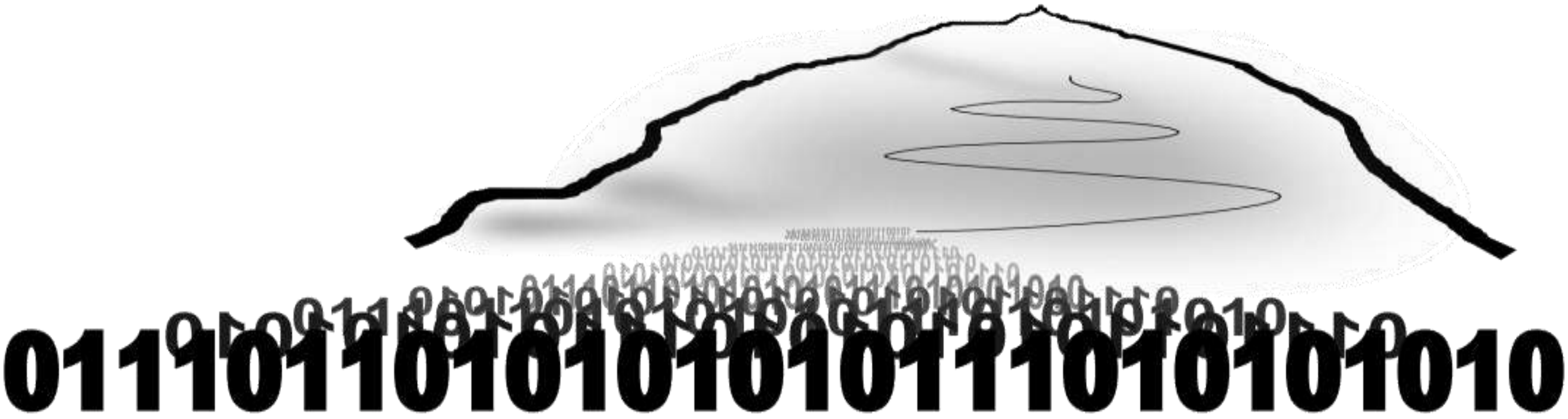
Large Scale Data Facility:
Design of meta data and community-specific services

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



Objectives of this talk

Title: Large Scale Data Facility:
Design of meta data and community-specific services

- Why is the LSDF different?
- Why is meta data important?
- Data and meta data management
- Advantages for the user:
  - Long term sustainability
  - Additional services
  - High throughput data analysis
- Examples

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

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)

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

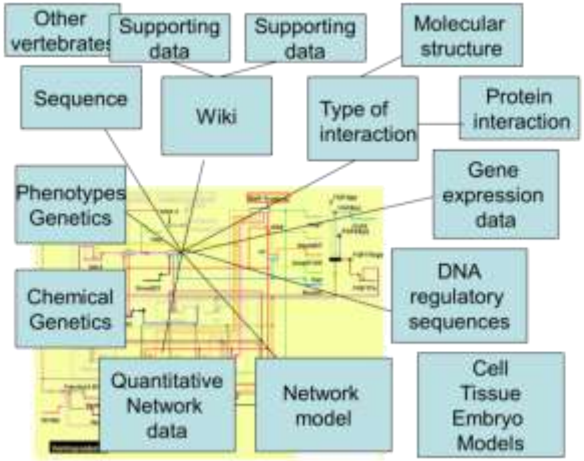
Have you ever tried to locate a file or info-somewhere-in-a-file-system

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

PANIC ?

Applications requiring meta data

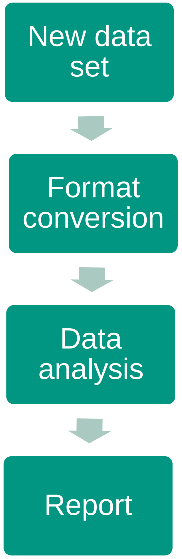
- Data archiving and retrieval (libraries)
- Fusion of complex data from various sources (data integration)



Source: Uwe Strähle

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.


New data set

Format conversion

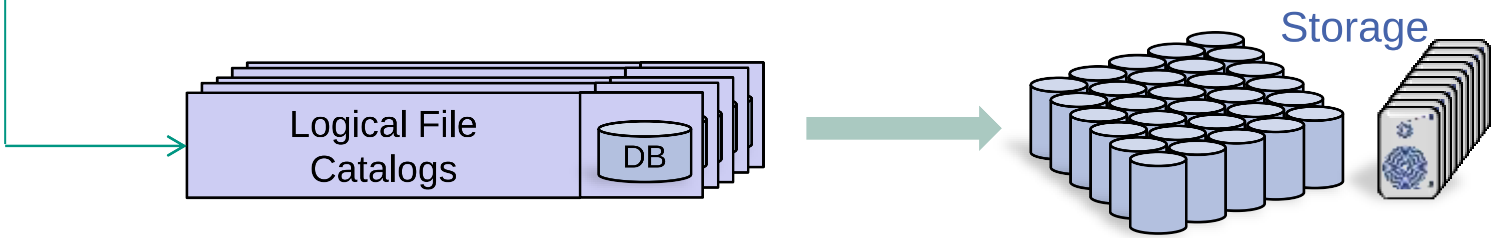
Data analysis

Report

Model of the LSDF meta data management

Idea:
Clear separation between

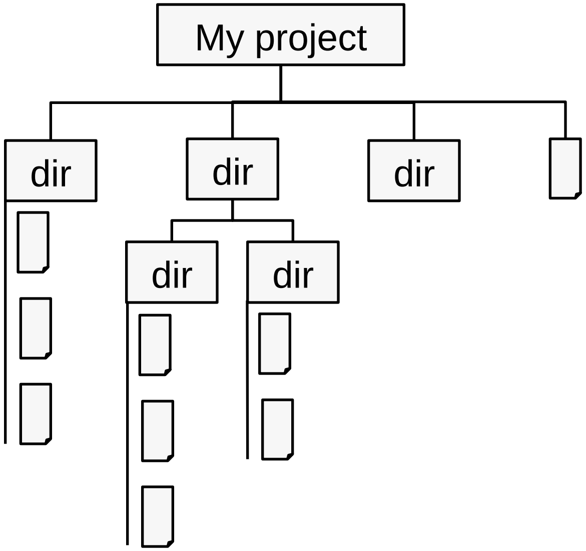
- Data (files),

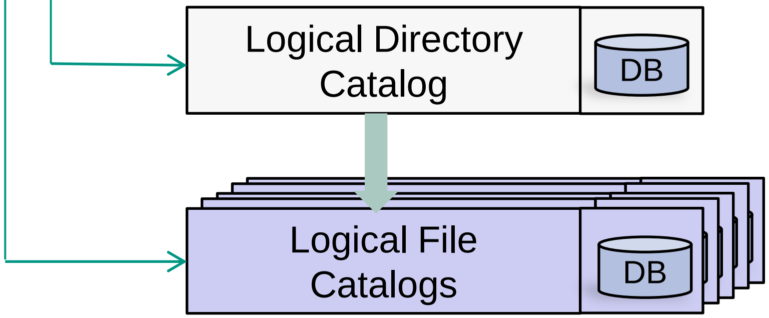


Model of the LSDF meta data management

Idea:
Clear separation between

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

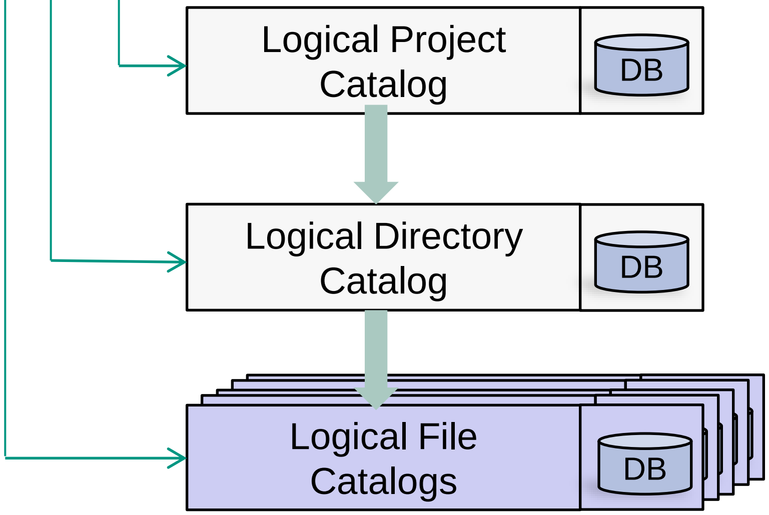




Model of the LSDF meta data management

Idea:
Clear separation between

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

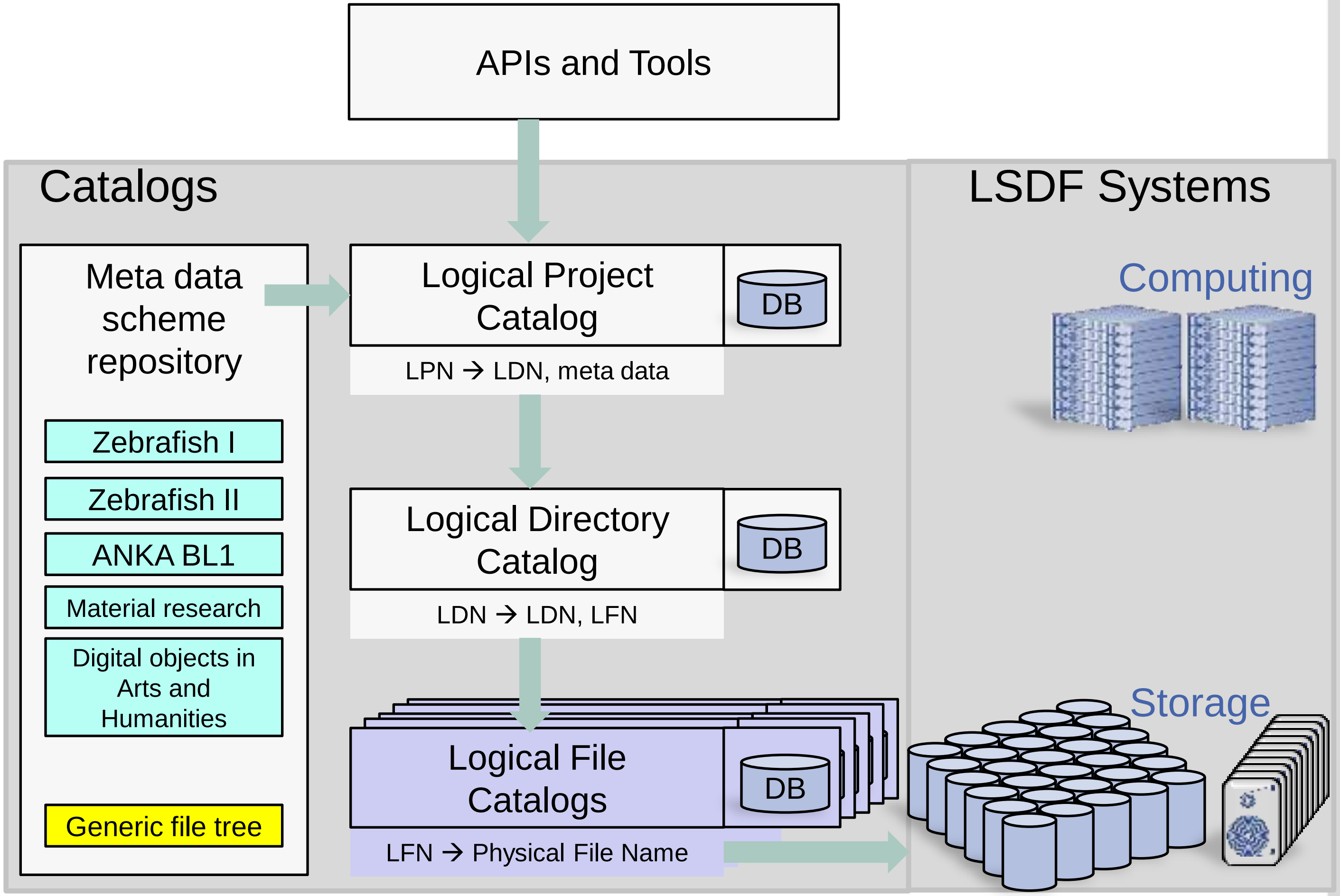


- name
- owners
- access rights
- date
    - community
    - (sub)subcommunity
        - measurement type
        - device, instrument
        - …

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

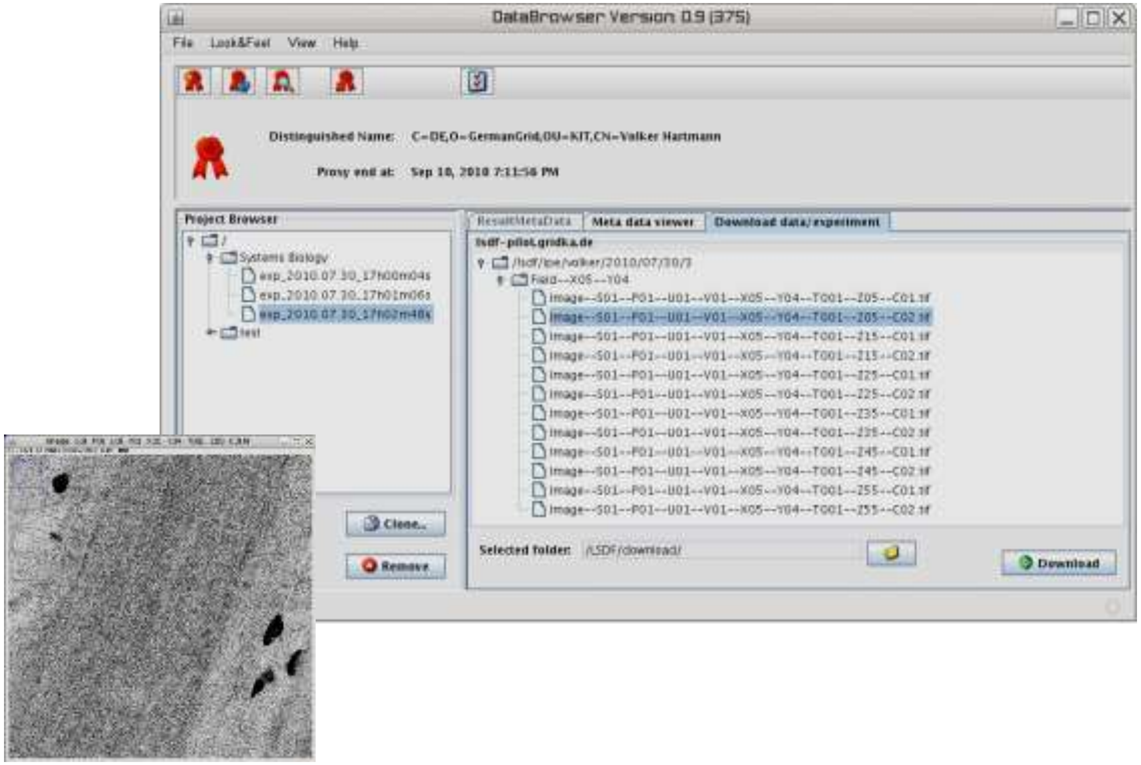
Hierarchical Catalog System

- 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
    - → **LSDF DataBrowser** is a File-, Data- and Project-Explorer



Distinguished Name: C=DE,O=GermanGrid,OU=KIT,CN=Volker Hartmann

Proxy end at: Sep 10, 2010 7:11:56 PM

```
/
  Systems Biology
    exp_2010.07.30_17h00m04s
    exp_2010.07.30_17h01m06s
    exp_2010.07.30_17h02m48s
  test
```

```
lsdf-pilot.gridka.de
/lsdf/ipe/volker/2010/07/30/3
  Field--X05--Y04
    image--S01--P01--U01--V01--X05--Y04--T001--Z05--C01.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z05--C02.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z15--C01.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z15--C02.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z25--C01.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z25--C02.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z35--C01.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z35--C02.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z45--C01.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z45--C02.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z55--C01.tif
    image--S01--P01--U01--V01--X05--Y04--T001--Z55--C02.tif
```

Selected folder: /LSDF/download/

DataBrowser allows:

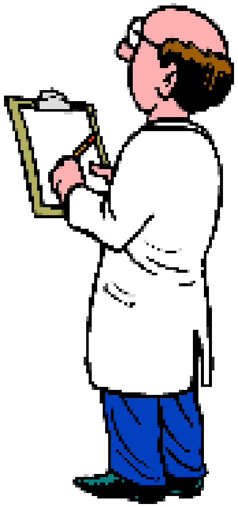
- Authentication
- Project and file browsing
- Upload
- Download
- Edit meta data
- Data visualization
- Control data analysis

Features:

- Extensible
- Huge variety of communication protocols
- Open source

How to handle the complexity?

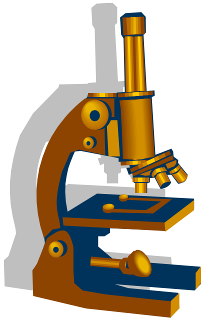
- *How do I insert a new scientific project ?*
  - → Data and meta data organization experts for projects with specific needs
  - → Generic meta data format for simple file trees



- *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.)

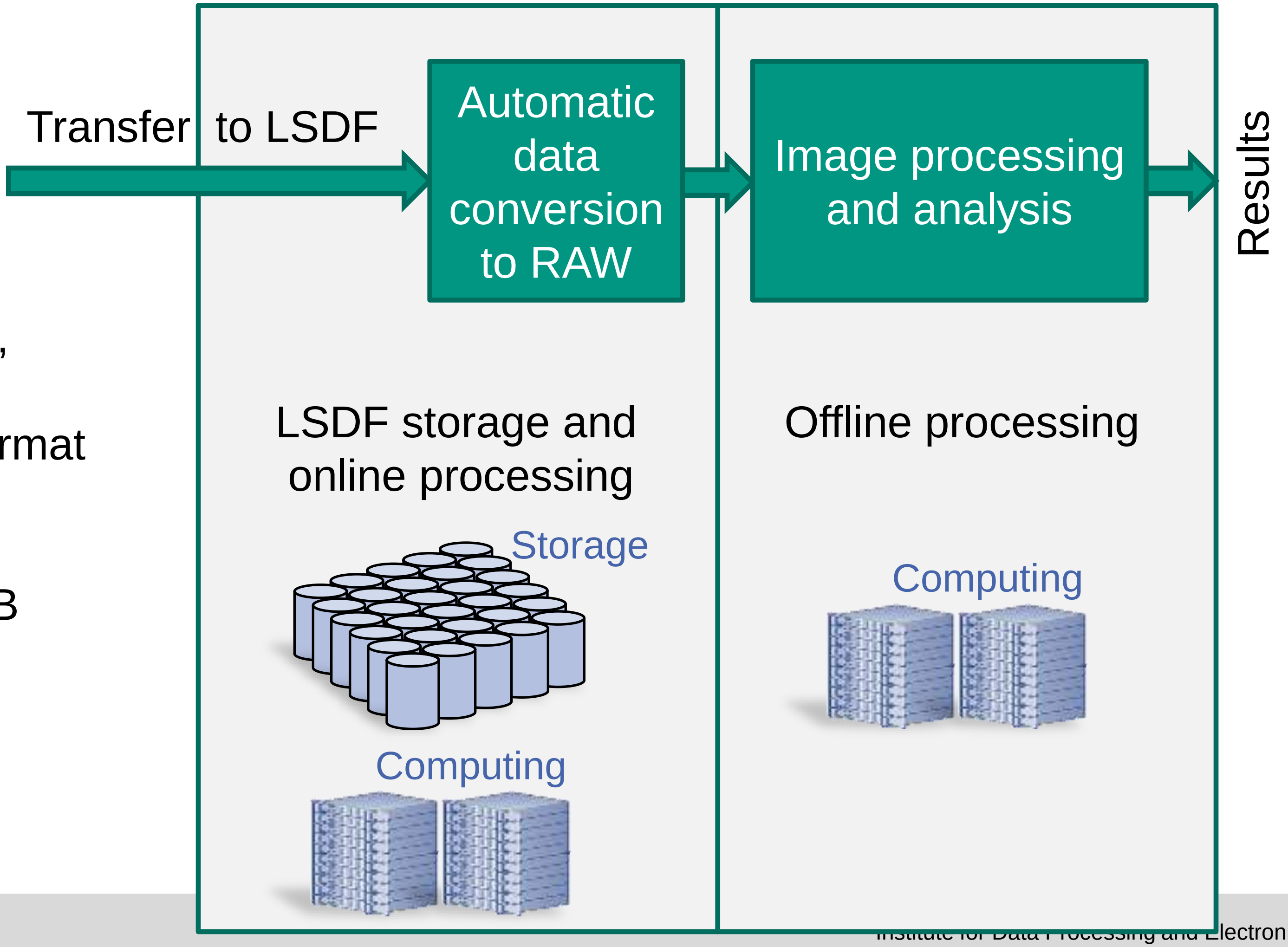
Example: Toxicology in early life stages *in vivo*

- Complex image analysis chain:



3D image stack,
time series,
Leica Image Format

data set size:
100 GB
# data sets:
> 100



Transfer to LSDF

Automatic data conversion to RAW

Image processing and analysis

Results

LSDF storage and online processing

Storage

Computing

Offline processing

Computing

Example: Toxicology in early life stages *in vivo*

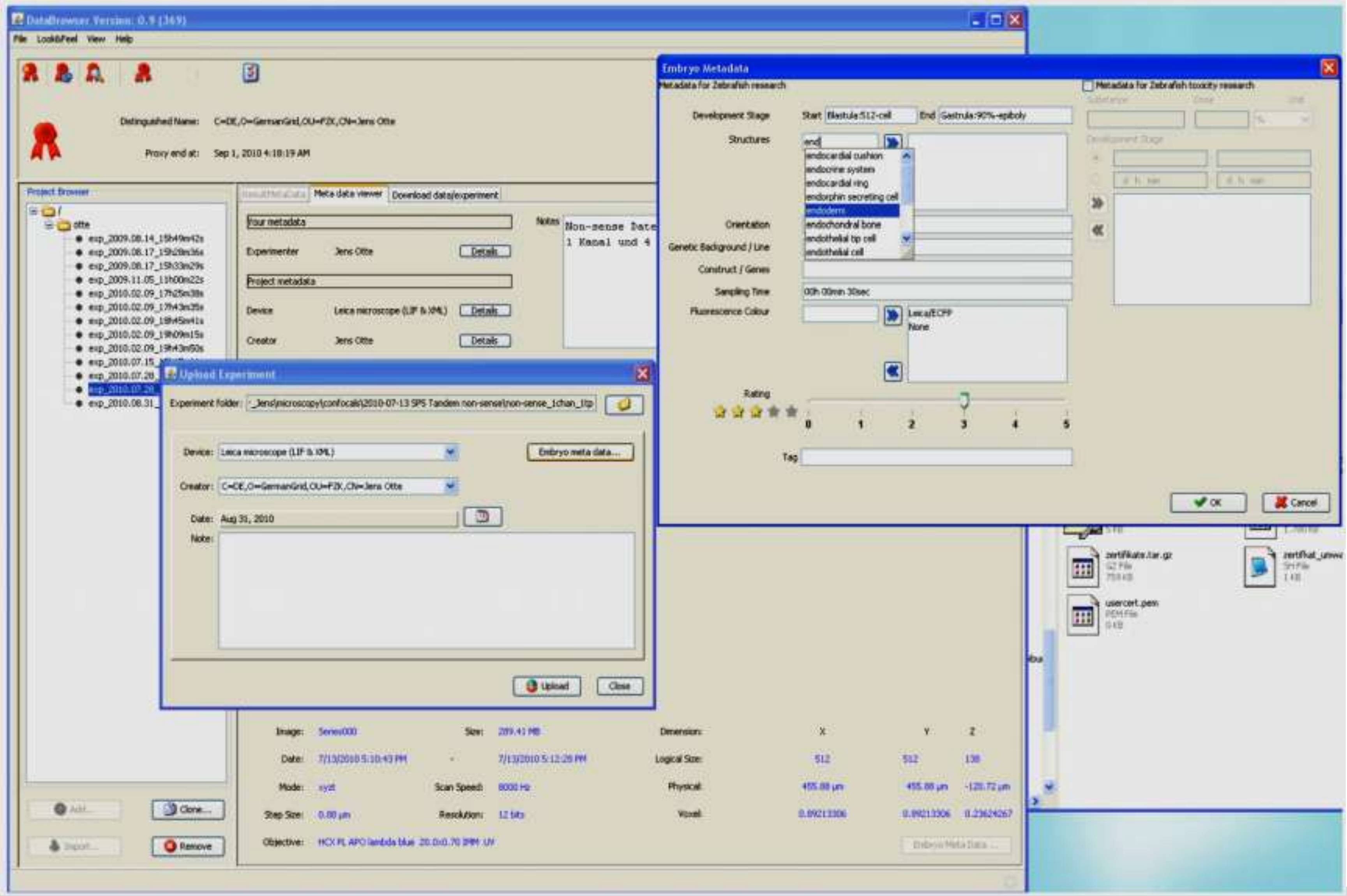
- Close cooperation ITG, IAI, SCC and IPE
  (Thanks to Jens C. Otte for the images)

Data Browser:

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

Estimated effort: ~ 2 PM

Example: ITG adapted DataBrowser



Distinguished Name: C=DE,O=GermanGrid,OU=FZK,CN=Jens Otte

Proxy end at: Sep 1, 2010 4:18:19 AM

- /
  - otte
    - exp_2009.08.14_15h49m42s
    - exp_2009.08.17_15h28m36s
    - exp_2009.08.17_15h33m29s
    - exp_2009.11.05_11h00m22s
    - exp_2010.02.09_17h25m38s
    - exp_2010.02.09_17h43m35s
    - exp_2010.02.09_18h45m41s
    - exp_2010.02.09_19h09m15s
    - exp_2010.02.09_19h43m50s
    - exp_2010.07.15_
    - exp_2010.07.28_
    - exp_2010.07.29_
    - exp_2010.08.31_

Your metadata

Experimenter: Jens Otte

Project metadata

Device: Leica microscope (LIF & XML)

Creator: Jens Otte

Notes:

```
Non-sense Date
1 Kanal und 4
```

Upload Experiment

Experiment folder: `_Jens\microscopy\confocals\2010-07-13 SPS Tandem non-sense\non-sense_1chan_1tp`

Device: Leica microscope (LIF & XML)

Creator: C=DE,O=GermanGrid,OU=FZK,CN=Jens Otte

Date: Aug 31, 2010

Note:

Embryo Metadata

Metadata for Zebrafish research

Development Stage: Start Blastula:512-cell End Gastrula:90%-epiboly

Structures: end

- endocardial cushion
- endocrine system
- endocardial ring
- endorphin secreting cell
- endoderm
- endochondral bone
- endothelial tip cell
- endothelial cell

Orientation:

Genetic Background / Line:

Construct / Genes:

Sampling Time: 00h 00min 30sec

Fluorescence Colour: Leica/ECFP
None

Rating: ★★★☆☆ 0 1 2 3 4 5

Tag:

□ Metadata for Zebrafish toxicity research

Substance / Dose / Unit: %

Development Stage

```
Image:      Series000                     Size:         289.41 MB              Dimension:      X             Y             Z
Date:       7/13/2010 5:10:43 PM   -      7/13/2010 5:12:28 PM                 Logical Size:   512           512           138
Mode:       xyzt                          Scan Speed:   8000 Hz                Physical:       455.88 µm     455.88 µm     -120.72 µm
Step Size:  0.88 µm                       Resolution:   12 bits                Voxel:          0.89213306    0.89213306    0.23624267
Objective:  HCX PL APO lambda blue  20.0x0.70 IMM UV
```

Scientific communities

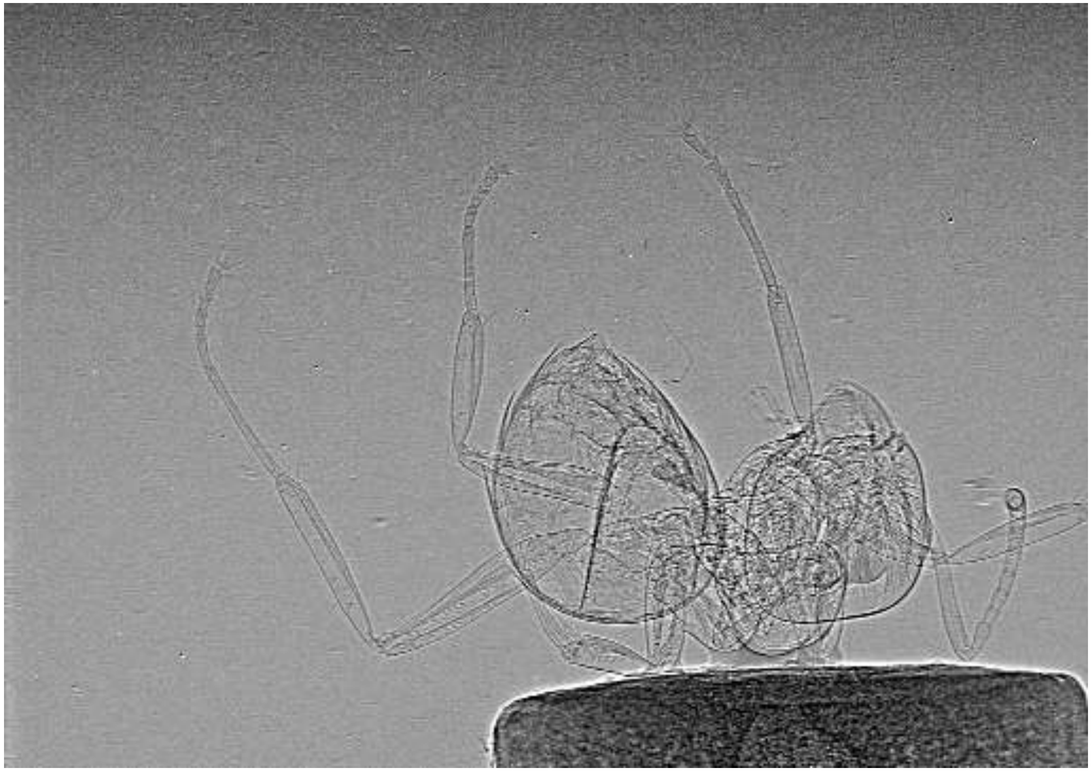
- Systems biology (ITG, BioQuant, Immunogenetics)
  - Vertebrate development studies and
  - Deconvolution (5000 data sets → <180 min.)
- Synchroton facilities and beamlines
  - ANKA data storage
  - HGF “High Data Rate Initiative”
- Climate research
- Material research
- Arts and humanities

»Il Cenacolo« von Da Vinci (1494-98)

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

Data intensive science

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



Conclusions

- LSDF is a powerful structure
  → more than data storage and cluster computing
- Design for future requirements → ExaByte storage + interactivity
  - R&D in progress

LSDF offers

- Sustainability and safety
- Flexibility for future requirements
- Interactivity
- Community-specific services
- Support

→ To gain faster and better scientific results

Thanks to

The team at IPE:
Volker Hartmann
Thomas Jejkal
Michael Sutter
Francesca Rindone
Michael Götter
Patrick Neuberger
Simon Ochsenreither

The team behind LSDF at SCC:
Serguei Bourov
Ariel Garcia
Bruno Hoeft
Rainer Kupsch
Achim Streit
Bernhard Verstege