Report on Data Management and Online Data Analysis Session

John Maclean

Advanced Photon Source

February 2nd 2012

Speakers

Talk	Speaker
Data management: PaN-Data (and CRISP) initiative	Rudolf Dimper ESRF
New data-intensive experiments and scientific opportunities for X-ray micro-tomography	Francesco De Carlo APS
High data rate initiative in the Helmholtz association (HDRI)	Rainer Gehrke Petra-III
Next generation data exploration: Intelligence in data analysis, visualization and mining	Stefan Vogt APS
Data analysis workbench (DAWB)	Olof Svensson ESRF
Experiment workflow pipelines at APS: message queuing and HDF5	Claude Saunders APS
icat metadata catalogue (from CRISP project)	D. Porte and A. Goetz ESRF

I have included slides from all these speakers in this presentation – Thank you

Common Themes

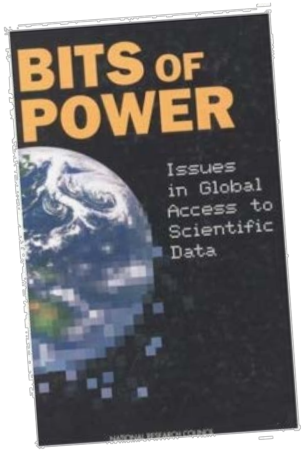
- Data volume:
  - Prepare for the Deluge, Tsunami, Avalanche, Flood and synonyms thereof
- Data Management and curation
- Meta data and data provenance
- HDF5 emerging as a common data container
- The need to provide data analysis infrastructure for users
- The need for Integration of data from multiple techniques, instruments

Data Volumes

- ESRF: *New detectors → high data rates (GB/s), high frame rates (<1ms/frame)*
- APS: Tomography - ~13GB per sample, HEDM ~1TB/day
- DESY: CFEL/PETRA III+/FLASH → 1.6 PB/year
- Cern: ATLAS 100 MB/s → 3 PB/year

- Computing infrastructure is under pressure

1. Scientific data is often considered private property


BITS OF POWER
Issues in Global Access to Scientific Data
NATIONAL RESEARCH COUNCIL

US National Research Council, Study: “Bits of Power, Issues in Global Access to Scientific Data”, 1997

“The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for … data derived from publicly funded research”

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007):

“Sharing and open access to publicly funded research data not only helps to maximise the research potential but provides greater returns from the public investment in research”

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES

OCDE

1. Scientific data is often considered private property

ESFRI Position Paper on Digital Repositories:

“Research Infrastructures should guarantee that raw research data are made available through portals and databases.”

06/09/2007 – e-IRG ESFRI

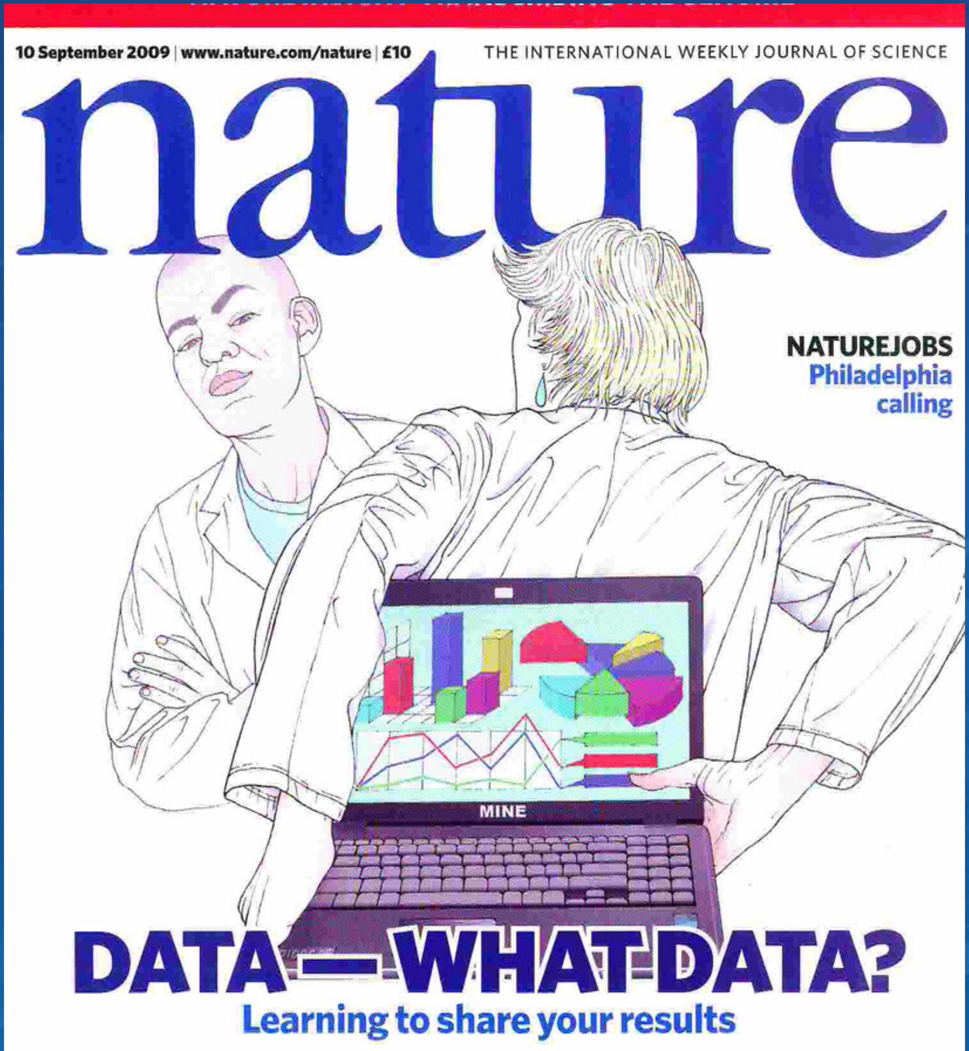
European Strategy Forum on Research Infrastructures
ESFRI
EUROPEAN ROADMAP FOR RESEARCH INFRASTRUCTURES
Roadmap 2008
UPDATE 2008

nature International weekly journal of science

Data's shameful neglect

“Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly”

Nature **461**, 145 (10 September 2009) | doi:10.1038/461145a


10 September 2009 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
NATUREJOBS
Philadelphia calling
MINE
DATA—WHAT DATA?
Learning to share your results

"Research cannot flourish if data are not preserved and made accessible.

All concerned must act accordingly."

Nature 461, 145 (10 September 2009)

We will also ask authors to provide a specific statement regarding the availability and curation of data as part of their acknowledgements ...

Science 11 February 2011

Science 2 December 2011

We must all accept that science is data and that data are science, and thus provide for, and justify the need for the support of, much-improved data curation.

2. Open access to scientific data is almost impossible today

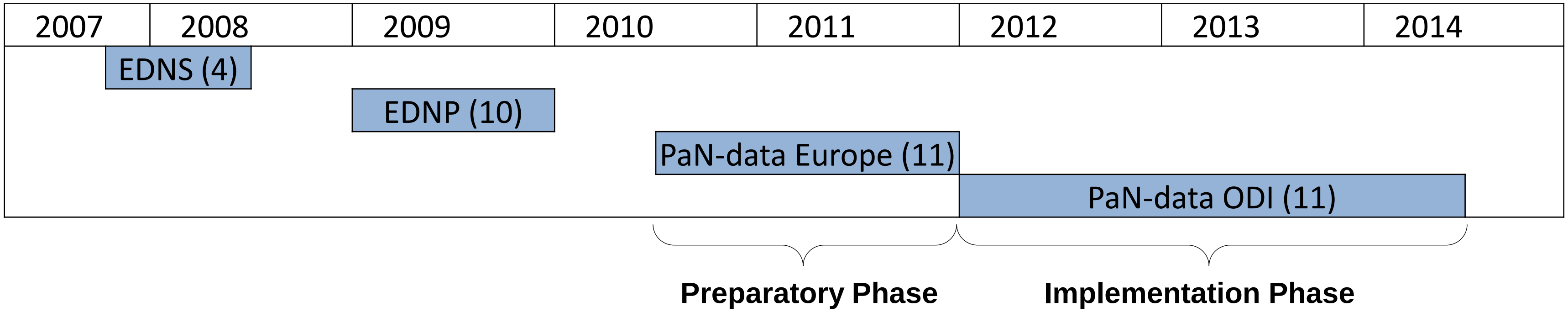
- Data is not on-line
- Data is poorly or not described
- No search tools
- No persistent identifiers
- Authentication/authorisation for scientists is difficult
- Open access data is not (yet) well accepted
- Institutions lack infrastructure

pandata

Established in 2007 with 4 facilities

Expanded since to 11 facilities

Goal: *“...to construct and operate a shared data infrastructure for Neutron and Photon laboratories...”*



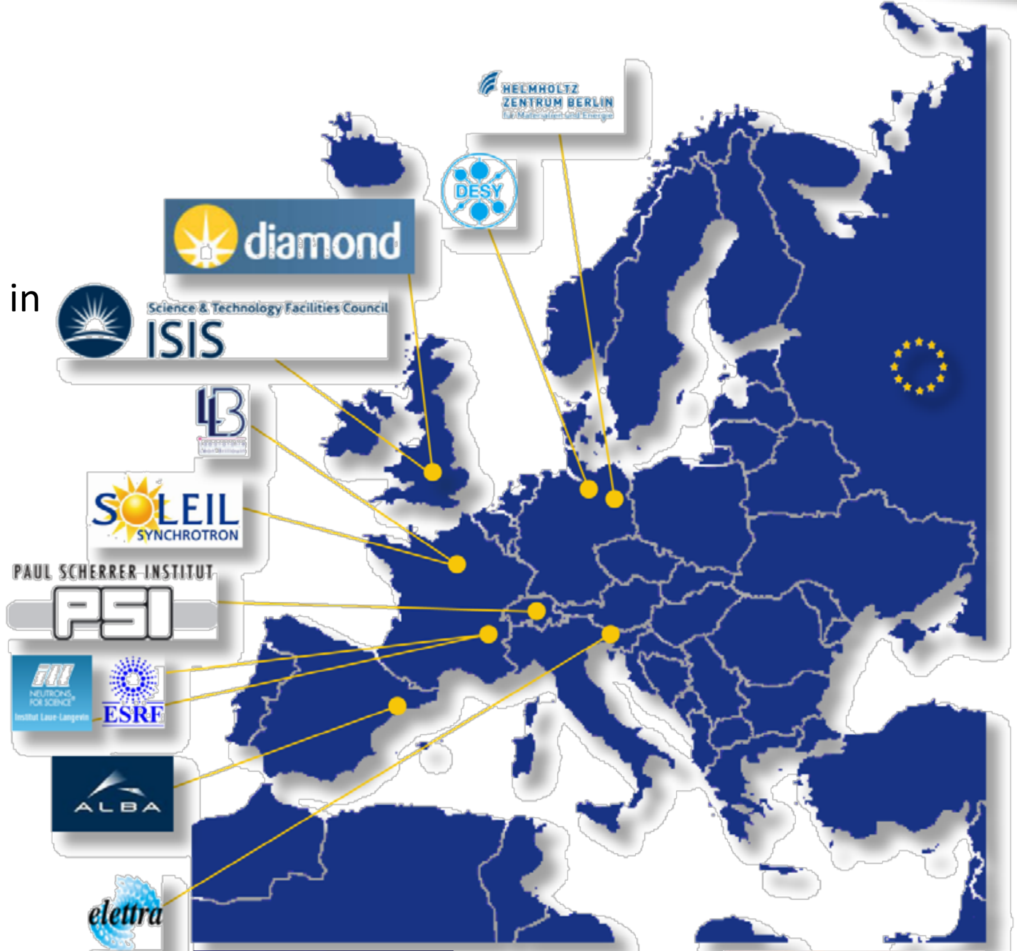
Preparatory Phase **Implementation Phase**

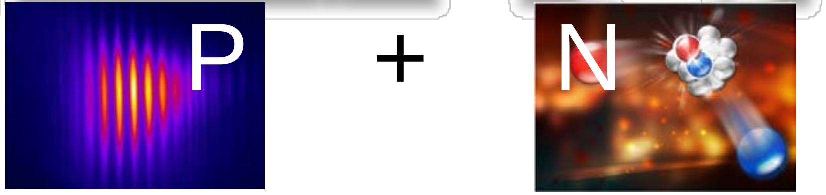
The PaN-data initiative

- Photons and Neutrons are complementary investigation tools
- Cross discipline experiments are increasing in number
- Neutron labs have built up data catalogues
- Synergy is essential for the project

Five P+N sites in Europe are in PaN-Data:

- ISIS + DIAMOND
- SINQ + SLS
- ILL + ESRF
- HMI + BESSY, now the HZB
- LLB + SOLEIL
- (+ DESY, ELETTRA, and ALBA)

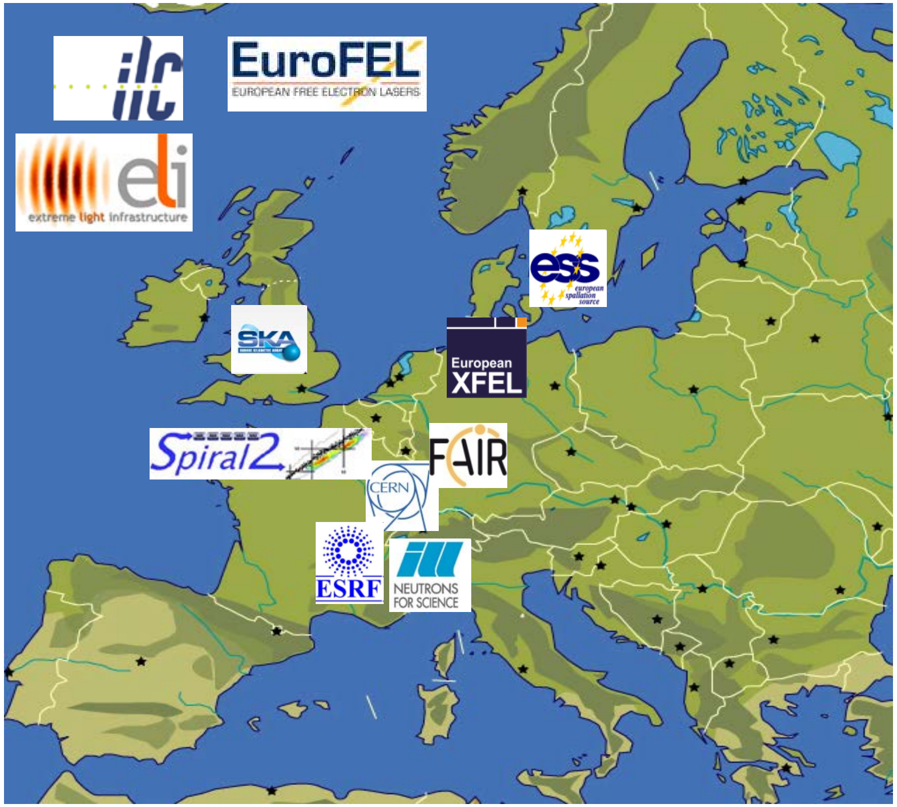


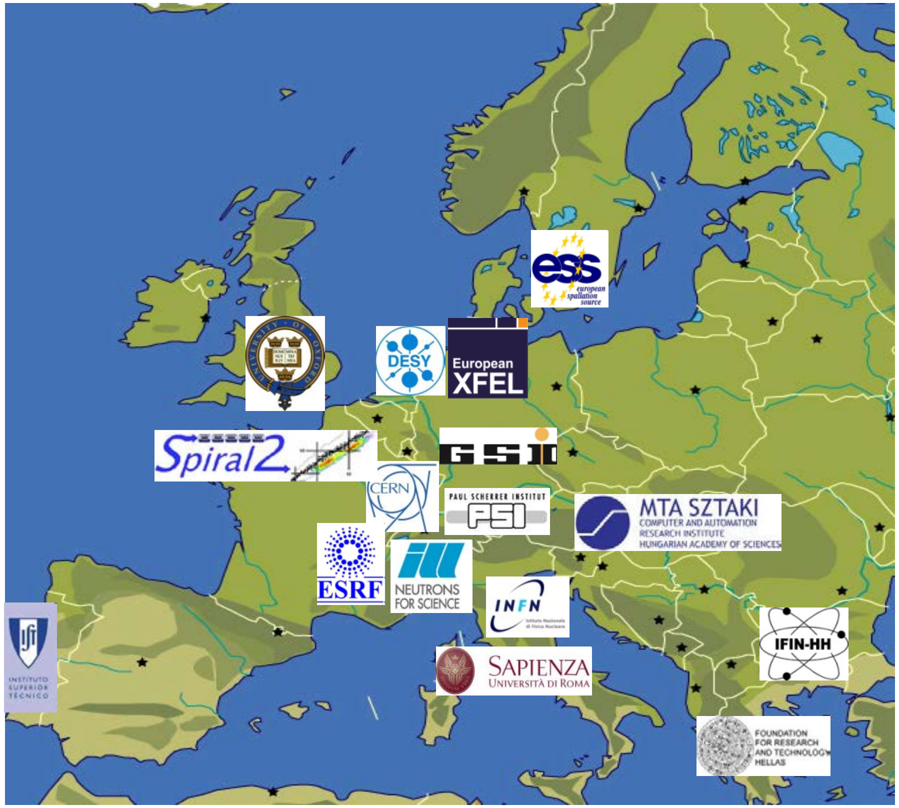


CRISP – Cluster of Research Infrastructures for Synergies in Physics

FP7 Project established in 2011 – 12M€/3 years

CRISP - Research Infrastructures and Participants


ilc
EuroFEL
EUROPEAN FREE ELECTRON LASERS
eli
extreme light infrastructure
ess
european spallation source
SKA
European XFEL
Spiral2
FAIR
CERN
ESRF
ill
NEUTRONS FOR SCIENCE


ess
european spallation source
DESY
European XFEL
Spiral2
GSI
CERN
PAUL SCHERRER INSTITUT
PSI
MTA SZTAKI
COMPUTER AND AUTOMATION RESEARCH INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES
ESRF
ill
NEUTRONS FOR SCIENCE
INFN
SAPIENZA
UNIVERSITÀ DI ROMA
IFIN-HH
INSTITUTO SUPERIOR TÉCNICO
FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS

11 Research Infrastructures & 16 Participating Institutions

CRISP

The Cluster of Research Infrastructures for Synergies in Physics

AAI

Account management

Proposal management

Remote data access

Remote experiment access

Metadata management and data mining:

Enhance and deploy metadata catalogues

Implement data mining

Data continuum – traceability, DOIs

High-speed data recording

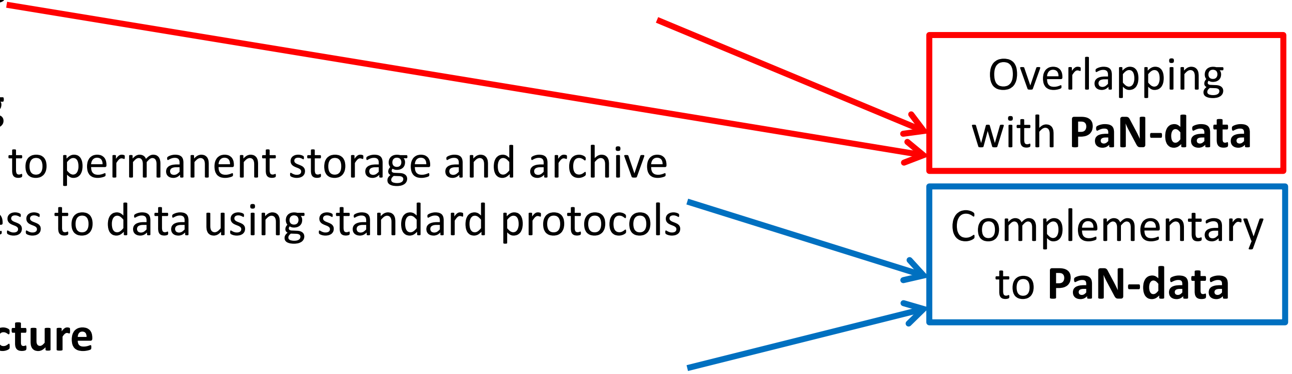
High-speed data recording to permanent storage and archive

Optimised and secure access to data using standard protocols

Distributed Data Infrastructure

Analyse existing data infrastructures from a network and technology perspective

Plan their evolution to support the expanding data management needs



Overlapping with **PaN-data**

Complementary to **PaN-data**

PaN Data and Crisp

- A common data format HDF5/Nexus
- A unique ID for scientists
  - A unique point to update user information (e.g. affiliation)
  - A unique password to access the facilities (remote data access, remote experiments)
  - A possible platform to manage proposals and facility events
  - A prototype implementation (based on Shibboleth) is operational
- ICAT (from ISIS) selected as meta data catalogue tool
  - In use at many facilities already
- Data Curation = preservation and maintenance of digital assets
  - Issues: Storage format evolution and obsolescence
    - Persistence of the digital objects and their identifiers
    - Rate of creation of new data and data sets
    - Broad access and searching flexibility
    - Obsolescence of data analysis code
- Agreement and policies to share analysis code

“Digital documents last forever -
or for five years, whichever comes first”
Jeff Rothenberg, 1997

High Data Rate Processing and Analysis Initiative (HDRI)

Helmholtz PNI Centres

DESY Hamburg	FZ Jülich
FZ Karlsruhe	HZG Geesthacht (former GKSS)
GSI Darmstadt	HZB Berlin

DESY Hamburg
HZ Geesthacht
HZ Berlin
FZ Jülich
GSI Darmstadt
FZ Karlsruhe

2010 - 2014

Work Packages

WP1: **Data Management (DESY, HZB)**

- Standardisation and Data Formats
- Data Access Strategies
- Data Lifetime Management and Archiving

WP2: **Real-time Data Processing (GSI, KIT)**

- Real-Time Data Assessment with Parallel Computing
- Analysis Methods and Applications
- Data Processing with Dedicated Hardware

WP3: **Data Analysis, Modeling, and Simulation (FZJ)**

Close co-operation with PanData

Conclusions and Outlook

WP1

- Design of standard data format has been settled
- Software development is progressing (NeXus API, Data Collector)
- Implementation at Instruments has started
- Approval of Common Data Policy in 2012
- Other issues to be solved in close co-operation with PanData (Authentification, Authorisation, Data Access Web-Portal …)
  Various Solutions for web based data access, mass storage, etc. are existing at the different centers.

WP2

- First case of GPU-processing is finished (Tomography)
- Start of second case for GPU-processing (Macromolecular Crystallography)

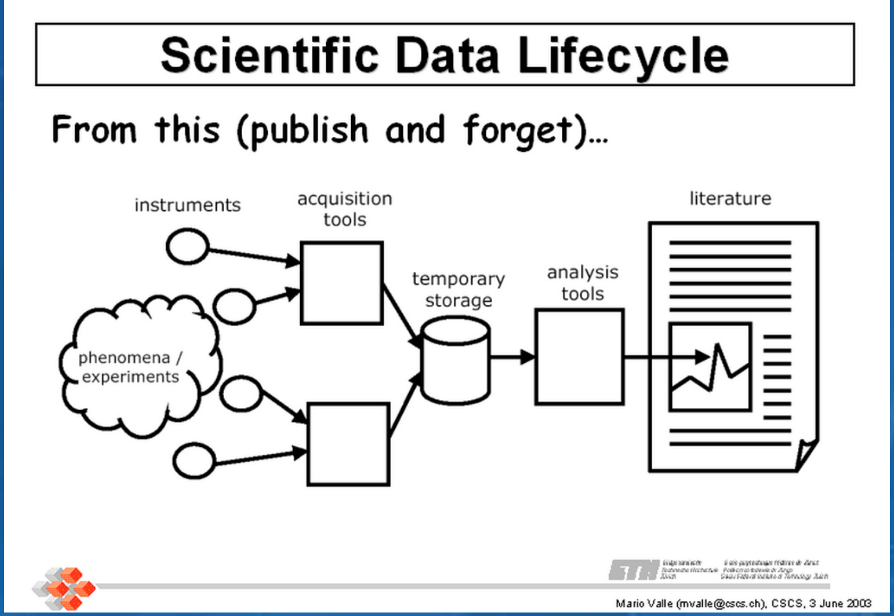
WP3

- DPDAK is operating and continuously extended (SAXS)
- Group for scientific computing at GSI has been formed

Long-term Goals: - Standard data format and fast data reduction and evaluation procedures
- Development of software for data evaluation
- Creation of a unique gateway for data access and evaluation

Scientific Data Management @ ESRF

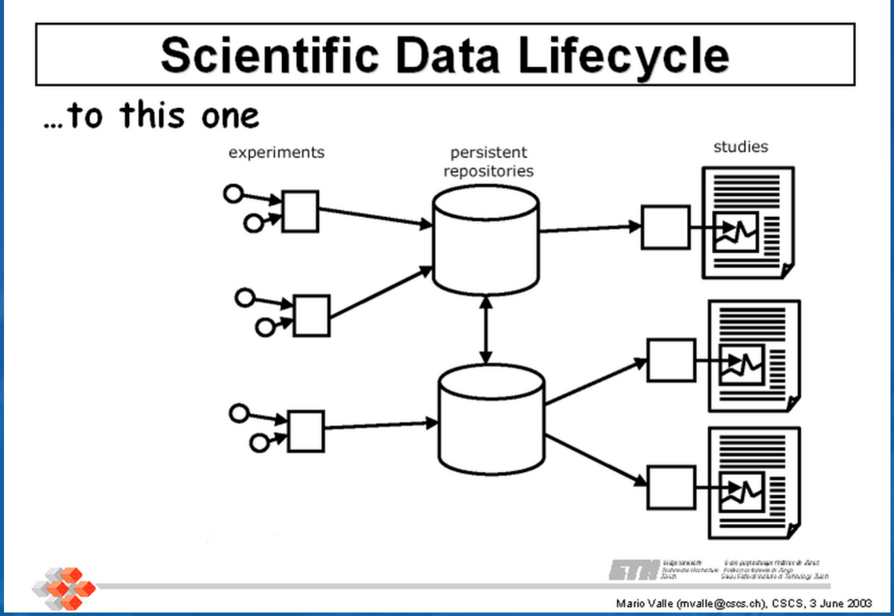
BEFORE **2012** **AFTER**



Scientific Data Lifecycle

From this (publish and forget)...

Mario Valle (mvalle@cscs.ch), CSCS, 3 June 2003



Scientific Data Lifecycle

...to this one

Mario Valle (mvalle@cscs.ch), CSCS, 3 June 2003

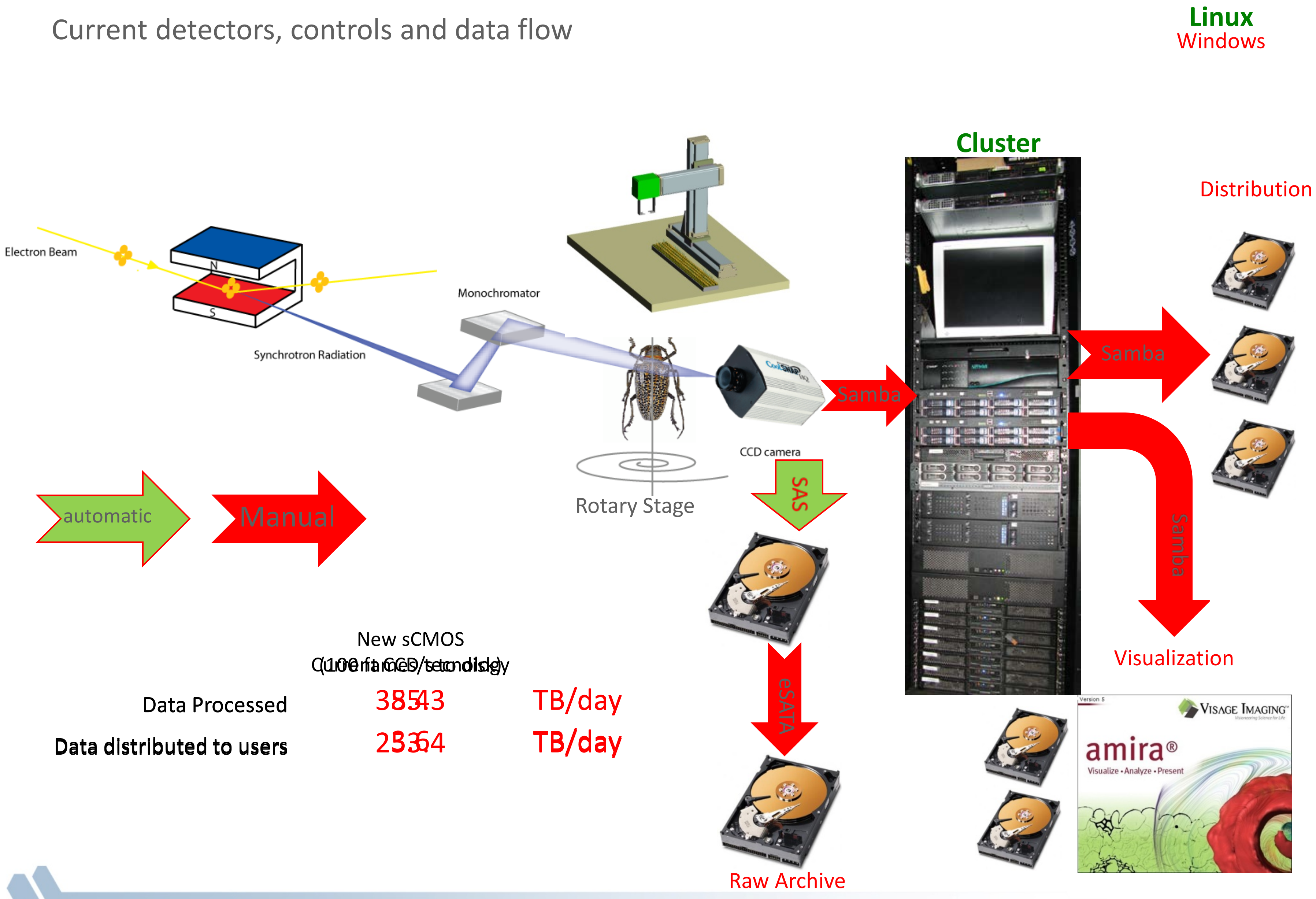
*http://personal.cscs.ch/~mvalle/sdm/scientific-data-management.html

Metadata Catalog - Added Value

- **Keep a permanent record of metadata (<< 1TB/yr) for all experiments, samples and conditions**
- **Enable automatic migration of data from online to archive storage**
- **Make public data available for download**
- **Answer questions like:**
  - What data did I take for Experiment X?
  - What experiments have been done on Sample Y?
  - What experiments have studied Sample Y at condition Z?
  - What public data are available for Sample Y at condition Z?

Micro tomography of static samples

Current detectors, controls and data flow



New sCMOS

Current CCD/technology (100 frames/second disk)

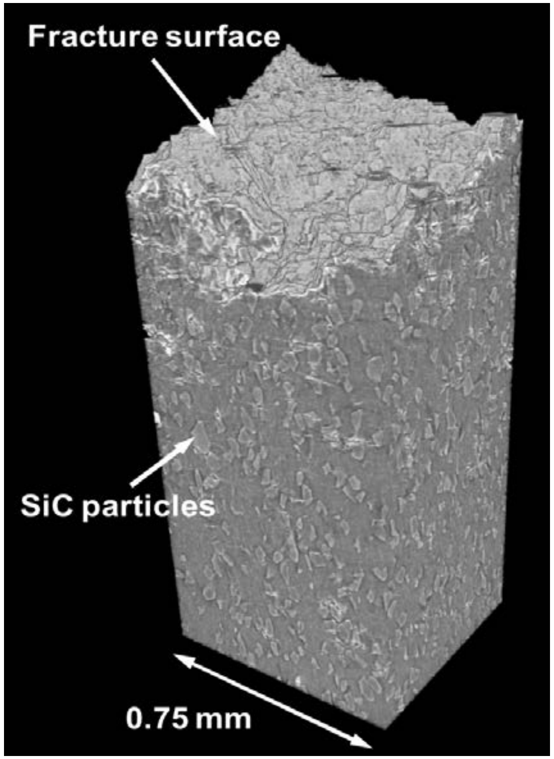
Data Processed	38543	TB/day
Data distributed to users	25364	**TB/day**

Micro Tomography Science

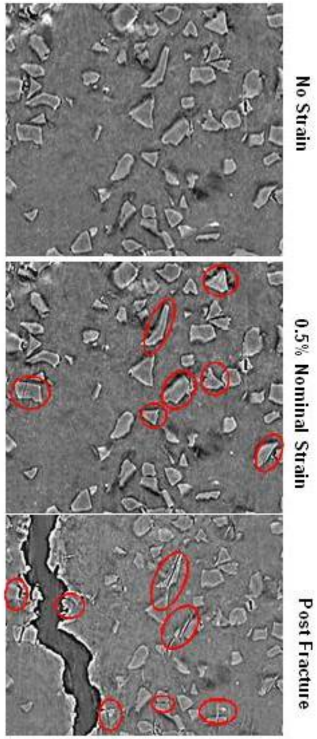
real size samples in real operational conditions

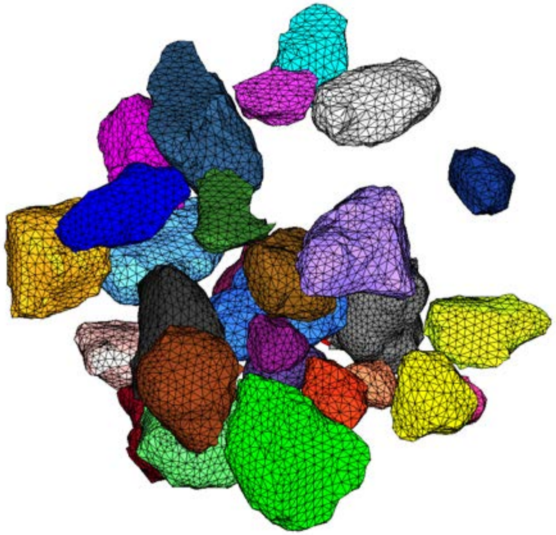
Mechanical Properties of Metal Matrix Composite Materials

transportation technology, new material, industrial applications



N. Chawla J. Williams ASU





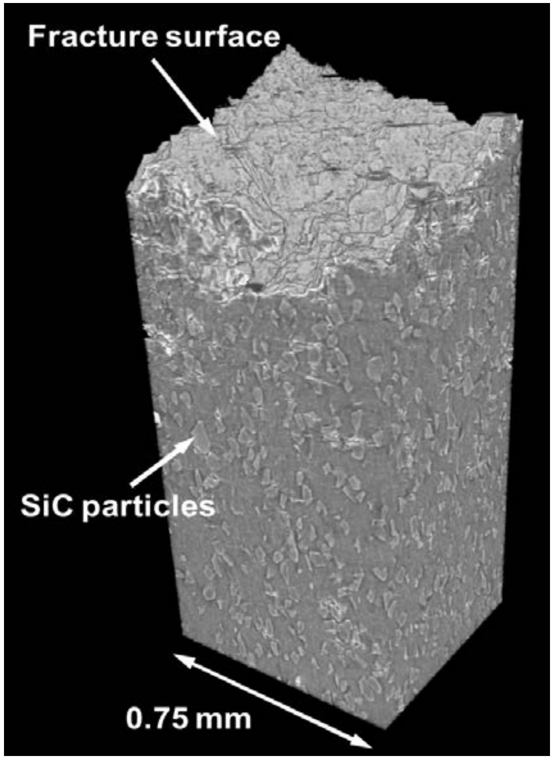
Metal Matrix Composite

Micro Tomography Science

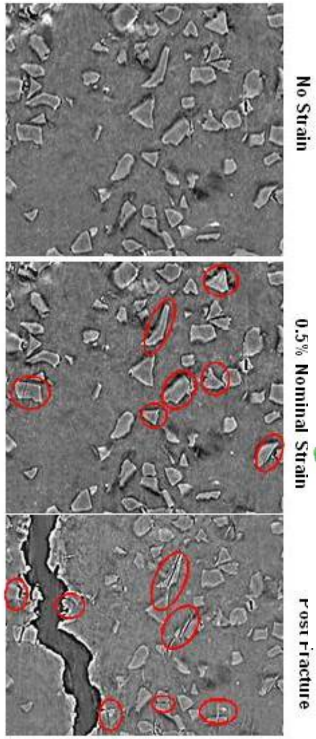
real size samples in real operational conditions

Mechanical Properties of Metal Matrix Composite Materials

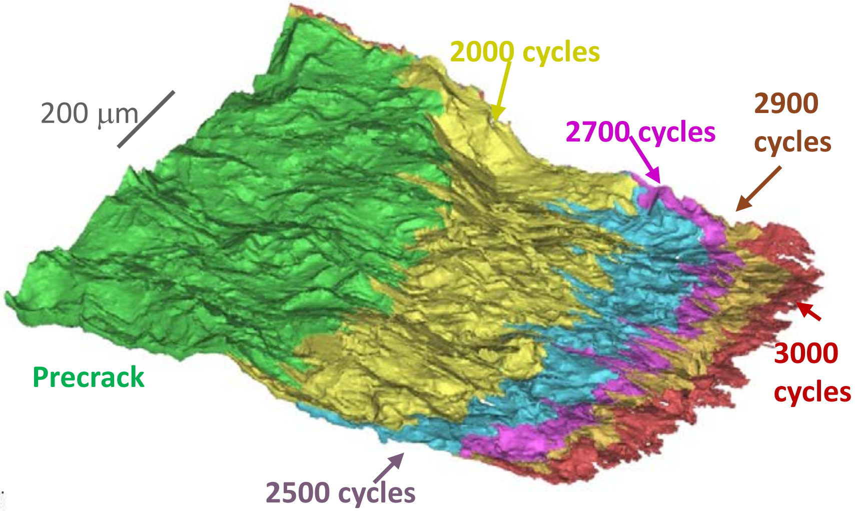
transportation technology, new material, industrial applications



N. Chawla J. Williams ASU



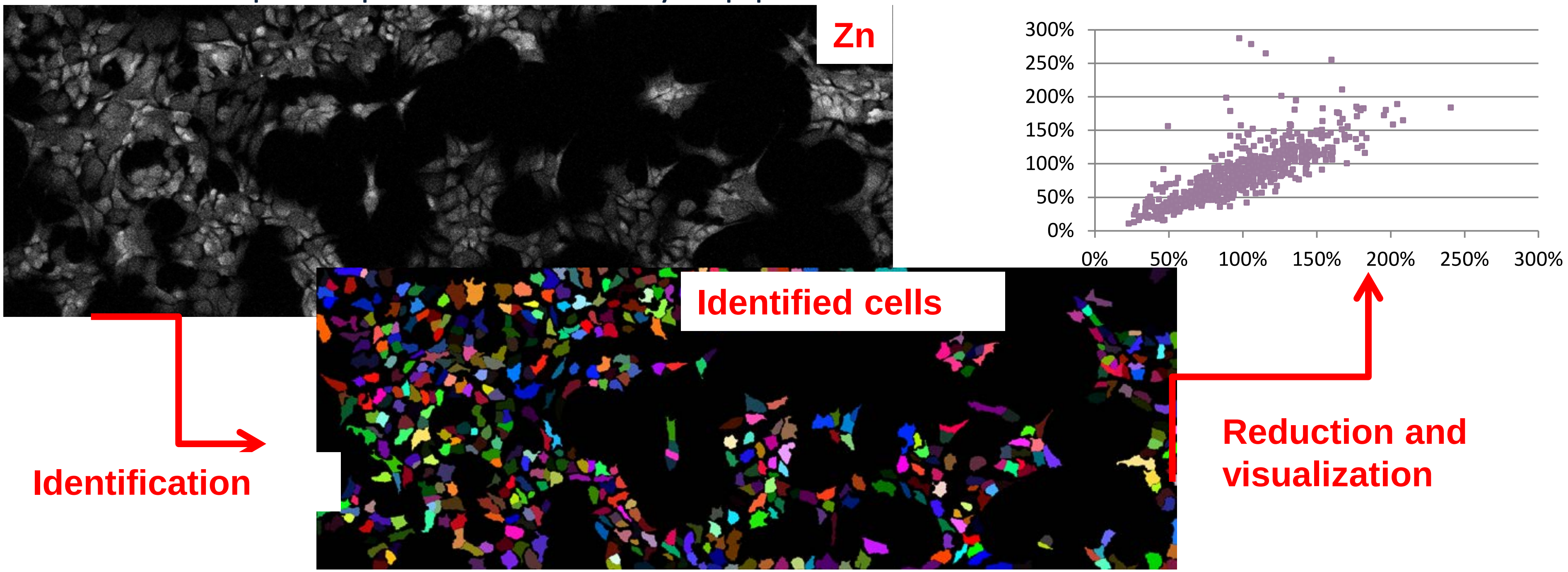
Acta Mater. 58 (18), 6194-6205 (2010)



Next Generation Data Exploration: Intelligence in Data Analysis, Visualization and Mining

- develop new generation of data analysis and visualization tools for multidimensional microscopy
  - Automatic identification and classification of objects
  - Enable correlative microscopy and analysis across a range of complementary instruments (light microscopy, electron microscopy, …)
  - Enable comprehensive datamining, with robust, rapid, and unsupervised analysis
  - Develop unsupervised data analysis pipline

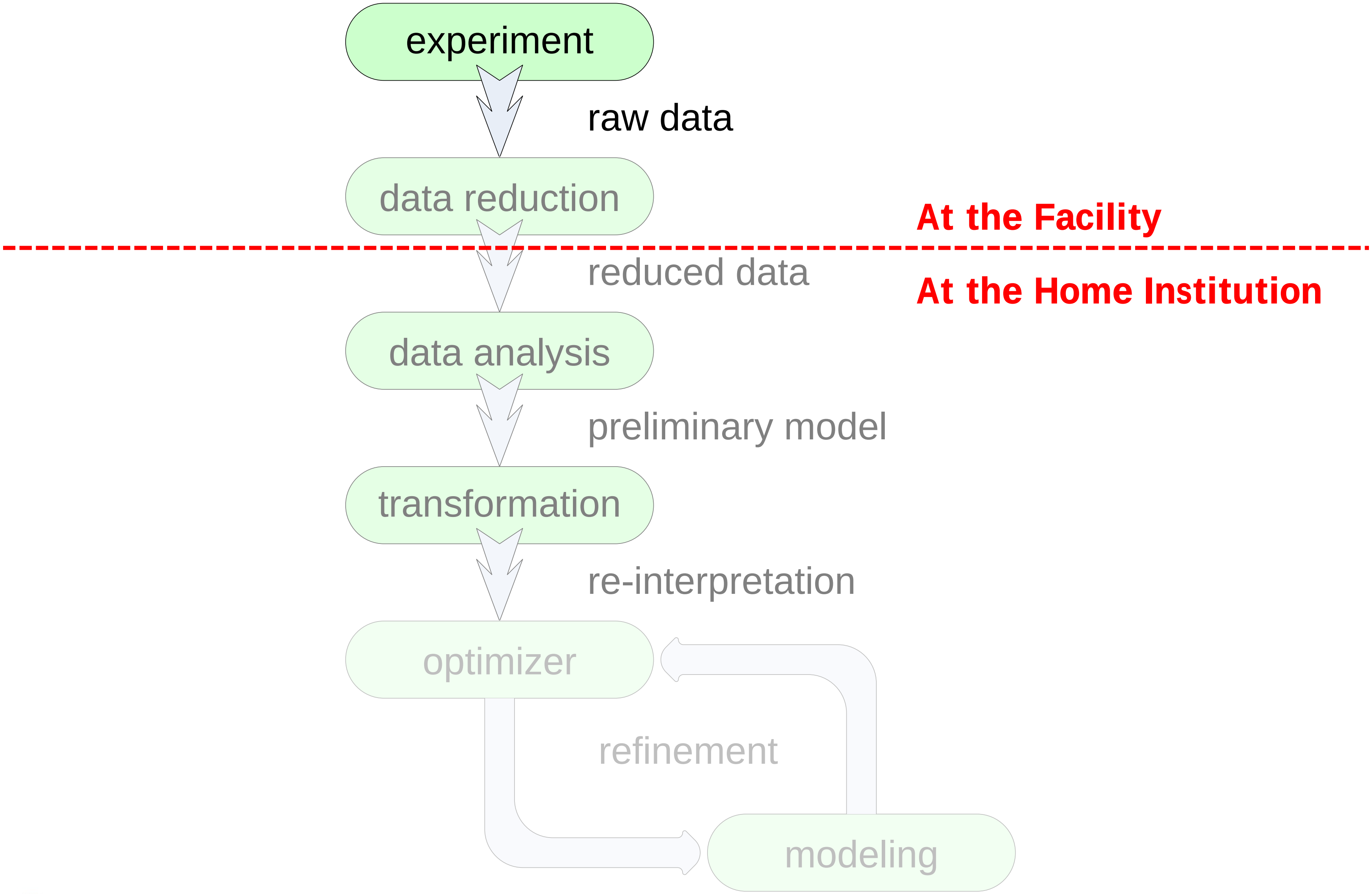
total Fe vs total Zn



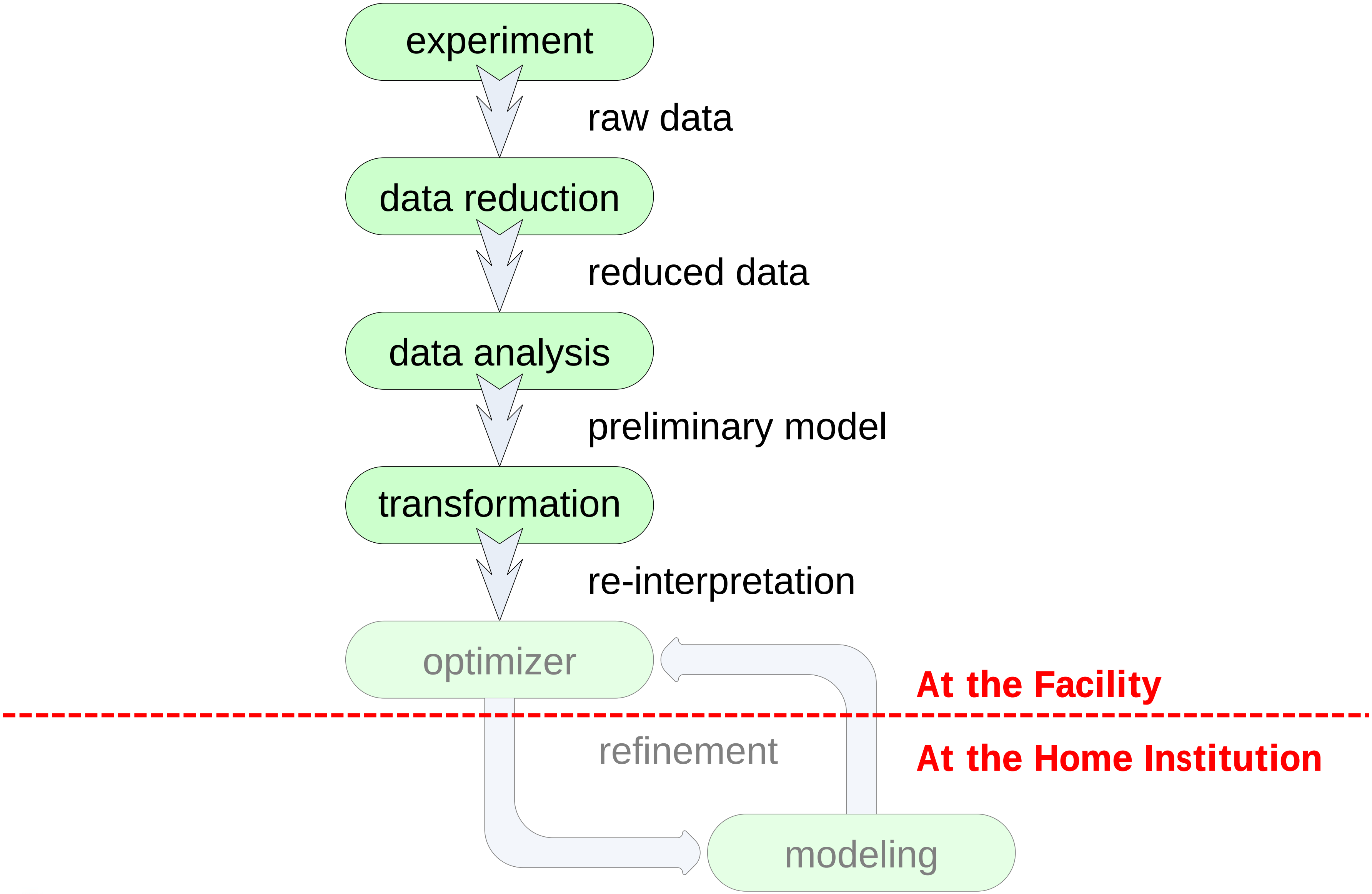
Left: Developed graph-partitioning based object identification algorithm, that was able to identify ~95% of the cells (and cell boundaries) in this complex dataset of ~ 500 HCT116 cells. Right: Zn is shown as one representative elemental map out of 10.

Ultimate goal: reason with abstraction of data instead of just images

Traditional Operational Workflow



Traditional Operational Workflow



experiment

raw data

data reduction

reduced data

data analysis

preliminary model

transformation

re-interpretation

optimizer

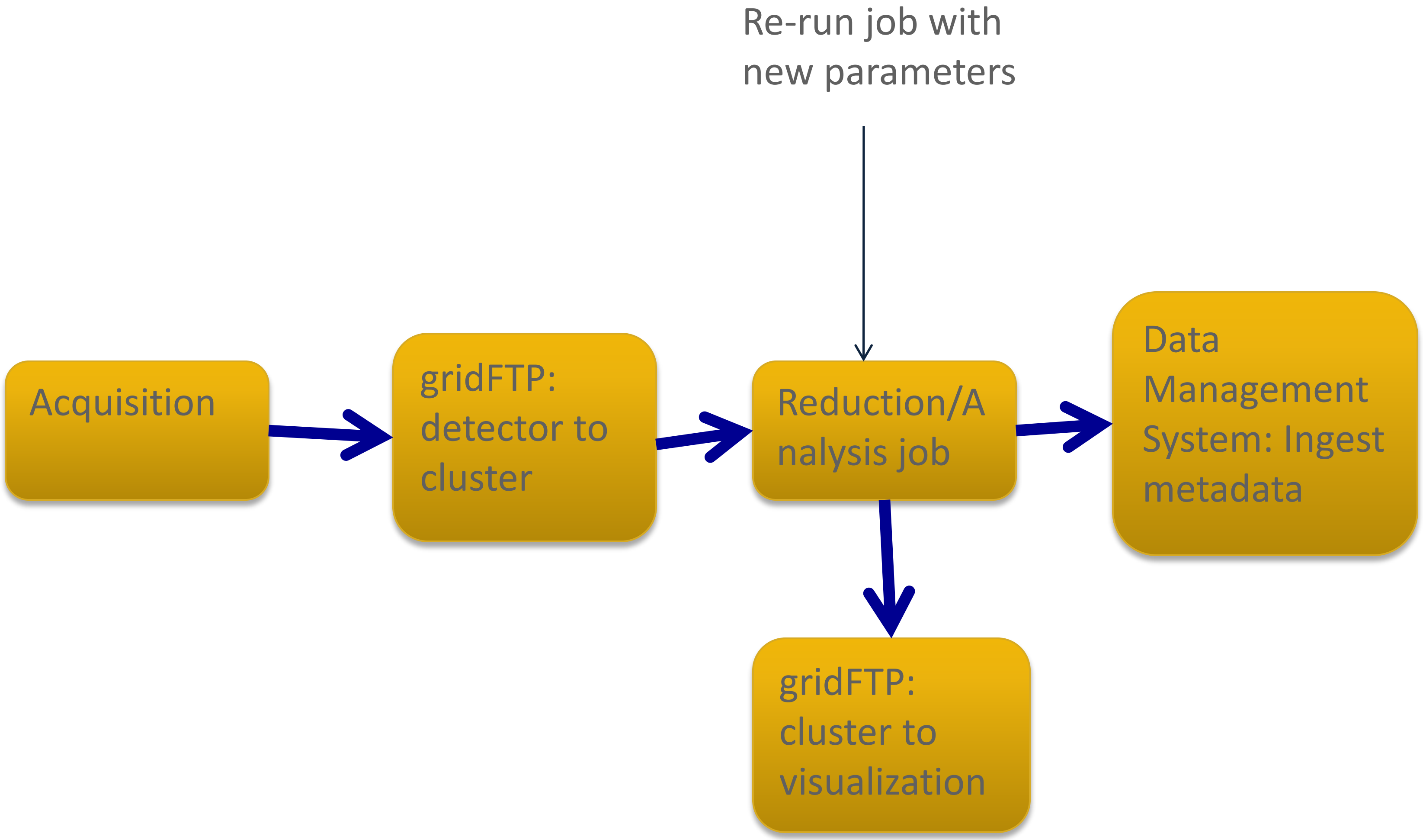
refinement

modeling

At the Facility

At the Home Institution

Abstracted Workflow



Re-run job with new parameters

Acquisition

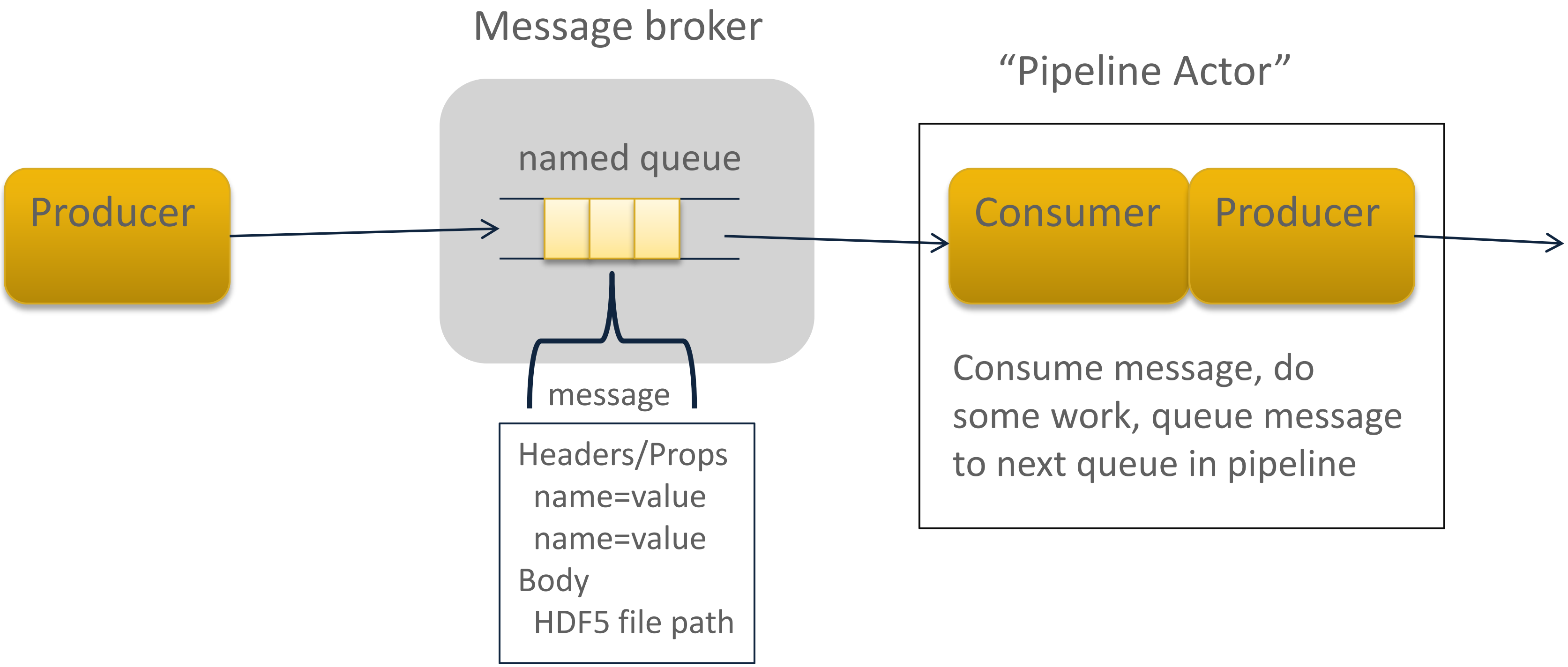
gridFTP: detector to cluster

Reduction/Analysis job

Data Management System: Ingest metadata

gridFTP: cluster to visualization

Message Queuing



- Producer and Consumer are temporally decoupled
- Message broker guarantees delivery of message
- Lots of production quality message brokers to choose from
  - We picked Apache ActiveMQ
- Can build all manner of pipelines with this

Data Exchange for Scientific Data and Metadata

Scientific Metadata

- Tomography Reconstruction
  - Iterative, analytical, interpolation type, etc.
- Instrument
  - Pixel size, orientation, etc.
- Sample
  - Temperature, pressure, etc.
- Data
  - 3D density map

All definition manual, code examples etc. in less than 20 pages !

Infrastructure Metadata

- Data transfer Status
  - End-points, progress, etc.
- Processing Status
  - Data ingestion date
- Cluster Queue status

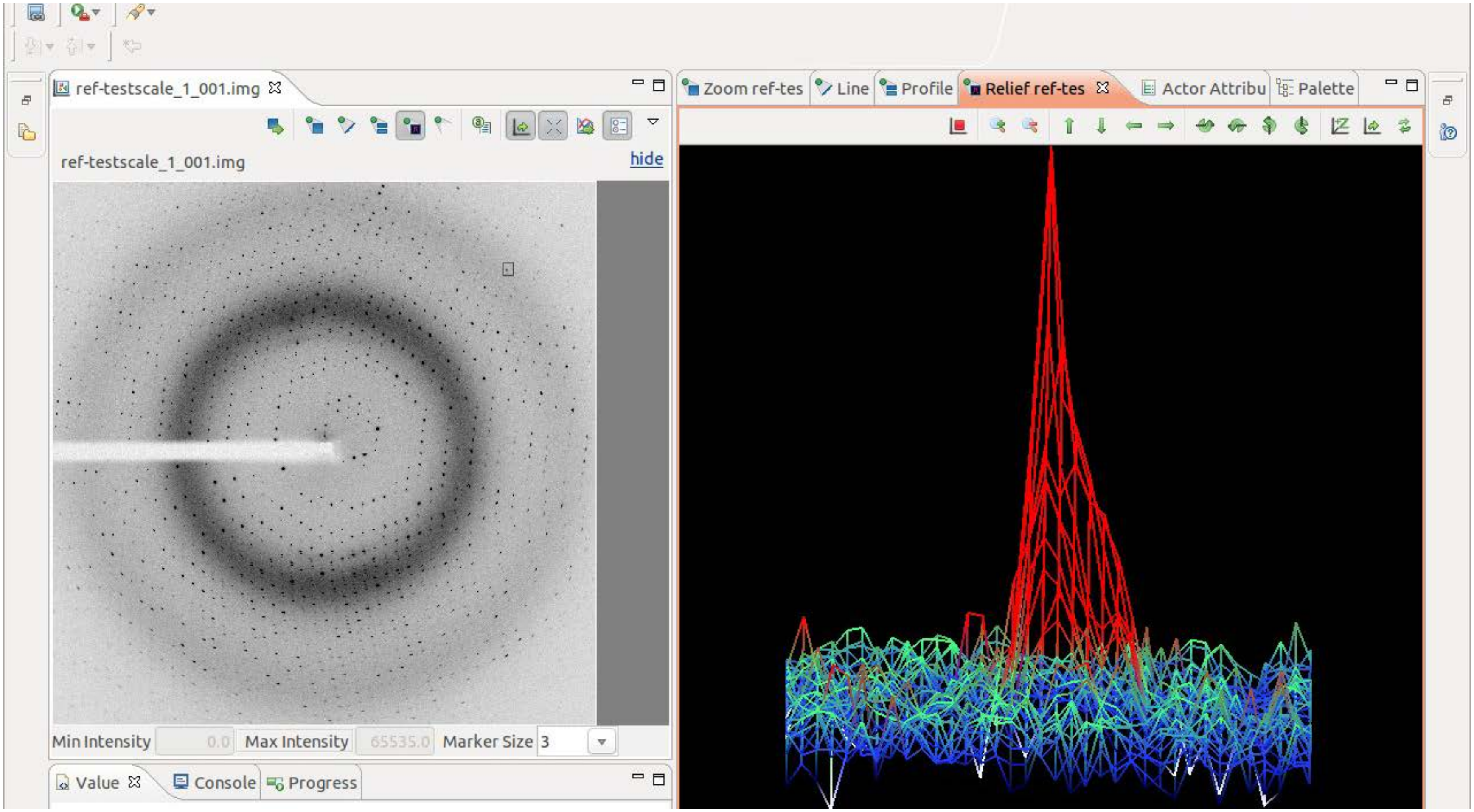
Provenance Layout

```
/provenance
          /next       "process_n"
          /process_n
                     /status
                     /ref
                     /message
          /infrastructure_n
```

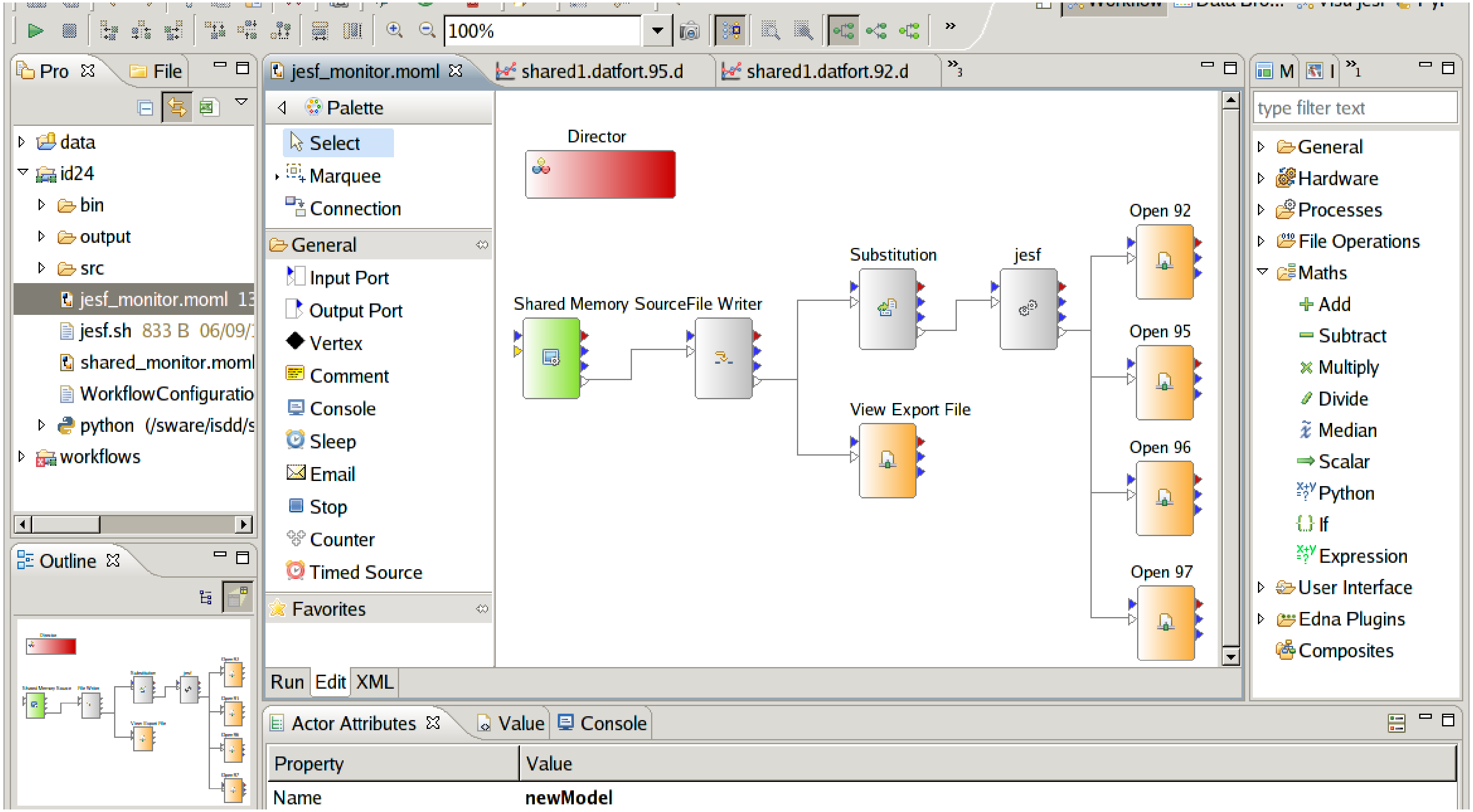
DAWB – Data Analysis Workbench

- Project goals Generic software tool for data analysis :
  - 1D, 2D and 3D visualisation
    Support for scripting languages (e.g. Python) Easy-to-use workflow tool for data reduction / analysis Off-line and on-line data analysis
- Framework for collaboration :
  - Re-use of existing components Modular project structure
- The ESRF Data Analysis Workbench (DAWB) project started in 2010
- An inter-facility collaboration around the workbench is being setup. The name of this collaboration is **DAWN.**
  - Diamond Light Source
    ESRF
    Soleil
    EMBL Grenoble
    Global Phasing ltd (Cambridge, UK) Isencia (Gent, Belgium)
- The current **DAWB** code will be migrated to **DAWN.** The code is already in the **DawnScience** Github repository:

DAWB – 3D Plotting



DAWB - GUI Workflow Design



Property	Value
Name	**newModel**

Conclusion

- Common problems
- Everyone concerned about data volume
- Talks generated much discussion
- We can have a large positive impact on science productivity
  - Improved workflow tools
  - Integrated data analysis
  - Data curation and management to become increasingly important