Karlsruhe Institute of Technology

HELMHOLTZ | ASSOCIATION

Fostering Data-Intensive Science in the German Helmholtz Association

Achim Streit
Director of Steinbuch Centre for Computing (SCC) and Professor in Computer Science, KIT

Steinbuch Centre for Computing (SCC)

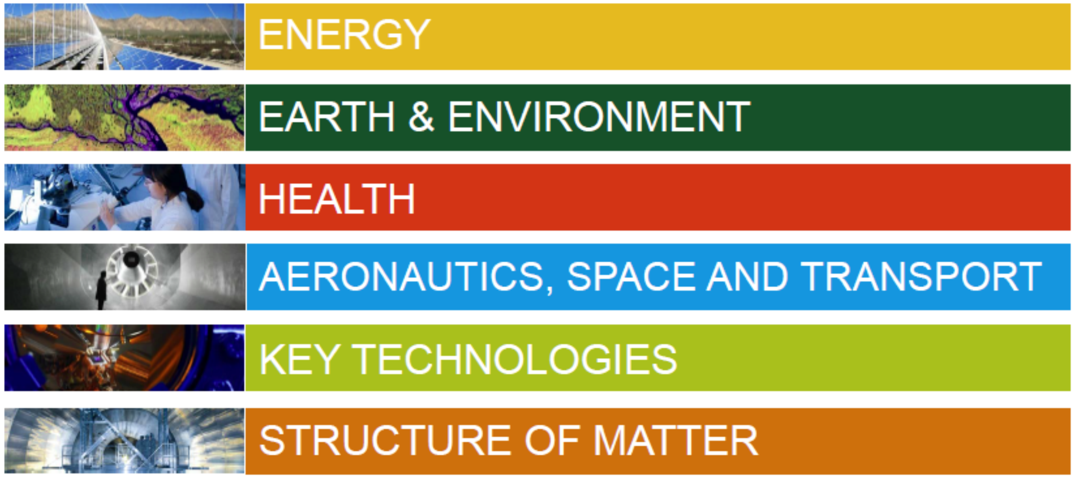
Helmholtz Mission

- **Strategic research for grand challenges** with cutting-edge research

- **Think big, act big:** Developing and operating **complex infrastructure** and **large-scale facilities** for the national and international scientific community

- **Creating wealth** for society and industry through transfer of knowledge and technology

Hermann von Helmholtz
(1821 – 1894)



ENERGY

EARTH & ENVIRONMENT

HEALTH

AERONAUTICS, SPACE AND TRANSPORT

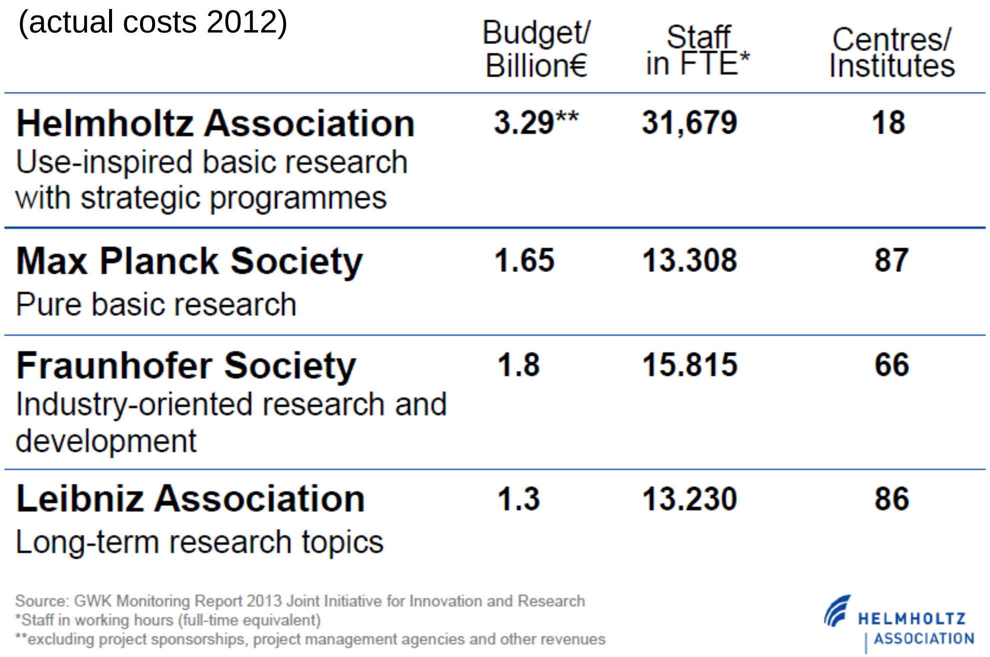
KEY TECHNOLOGIES

STRUCTURE OF MATTER

Helmholtz – Fact & Figures

- 35,672 Staff (employees 2012)
  - 12,709 scientists & engineers
  - 6,635 PhD students
  - 1,652 trainees

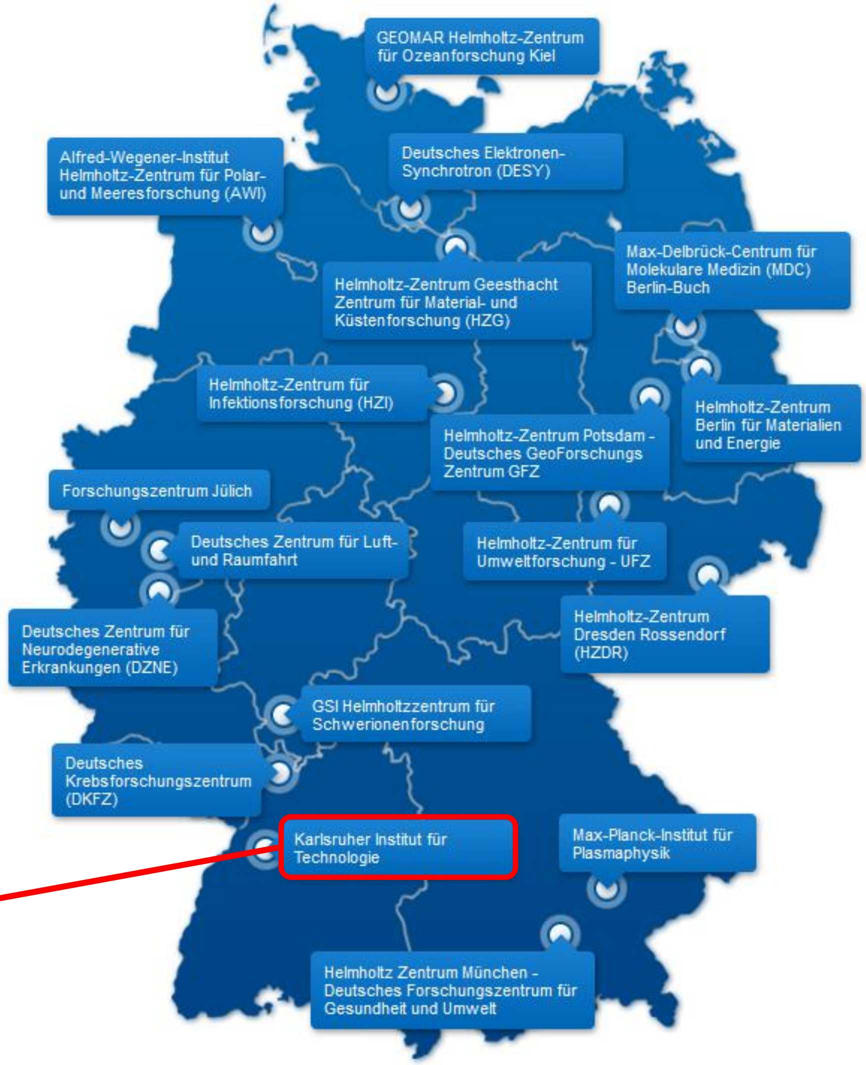
- Budget 2013: €3.6 billion
  - €2.4 bn (budget approach) Institutional funding (90% federal, 10% local)
  - €1.0 bn: Third-party funding (basis actual costs 2012)
  - €0.2 bn: Special Financing



(actual costs 2012)	Budget/ Billion€	Staff in FTE*	Centres/ Institutes
Helmholtz Association Use-inspired basic research with strategic programmes	3.29**	31,679	18
Max Planck Society Pure basic research	1.65	13.308	87
Fraunhofer Society Industry-oriented research and development	1.8	15.815	66
Leibniz Association Long-term research topics	1.3	13.230	86

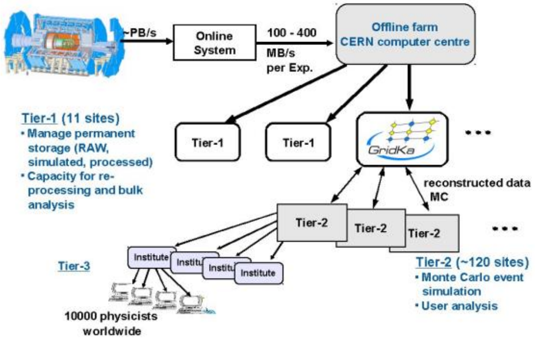
Source: GWK Monitoring Report 2013 Joint Initiative for Innovation and Research
*Staff in working hours (full-time equivalent)
**excluding project sponsorships, project management agencies and other revenues

Helmholtz – 18 Centers


GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel
Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung (AWI)
Deutsches Elektronen-Synchrotron (DESY)
Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin-Buch
Helmholtz-Zentrum Geesthacht Zentrum für Material- und Küstenforschung (HZG)
Helmholtz-Zentrum für Infektionsforschung (HZI)
Helmholtz-Zentrum Berlin für Materialien und Energie
Helmholtz-Zentrum Potsdam - Deutsches GeoForschungs Zentrum GFZ
Forschungszentrum Jülich
Deutsches Zentrum für Luft- und Raumfahrt
Helmholtz-Zentrum für Umweltforschung - UFZ
Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE)
Helmholtz-Zentrum Dresden Rossendorf (HZDR)
GSI Helmholtzzentrum für Schwerionenforschung
Deutsches Krebsforschungszentrum (DKFZ)
Karlsruher Institut für Technologie
Max-Planck-Institut für Plasmaphysik
Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt

GridKa

- German Tier-1 Center in WLCG
  - Supports all 4 LHC experiments + Belle-II + several small {HE/AP}P communities
  - Benchmarked reliability of 99.5%
- Resources
  - >10,000 cores, utilization >95%
  - Disk space (net.): 12 PB, tape space: 18 PB
  - 6x10 Gbit/s network connectivity
- 14% of LHC data permanently stored at GridKa
- Serves > 20 T2 centers in 6 countries
- Services
  - File transfer
  - Regional workload management, file catalog
- Annual international GridKa School
- Global Grid User Support (GGUS) for WLCG



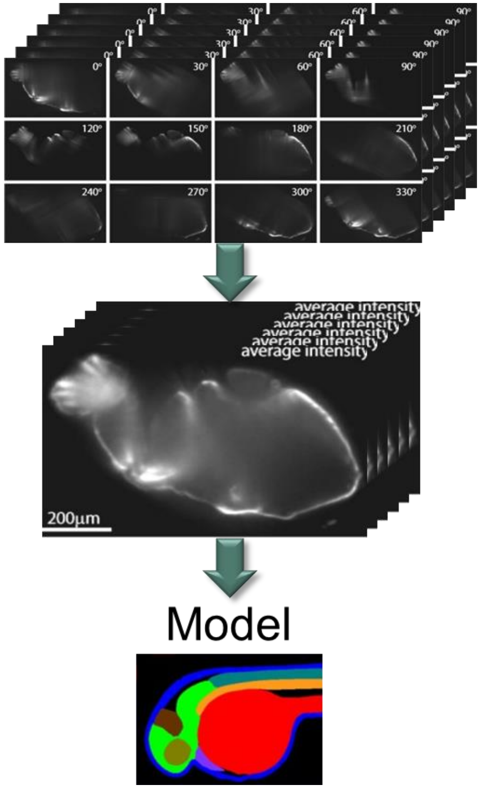
Tier-1 (11 sites)
- Manage permanent storage (RAW, simulated, processed)
- Capacity for re-processing and bulk analysis

Tier-2 (~120 sites)
- Monte Carlo event simulation
- User analysis

GridKa School

Large Scale Data Facility

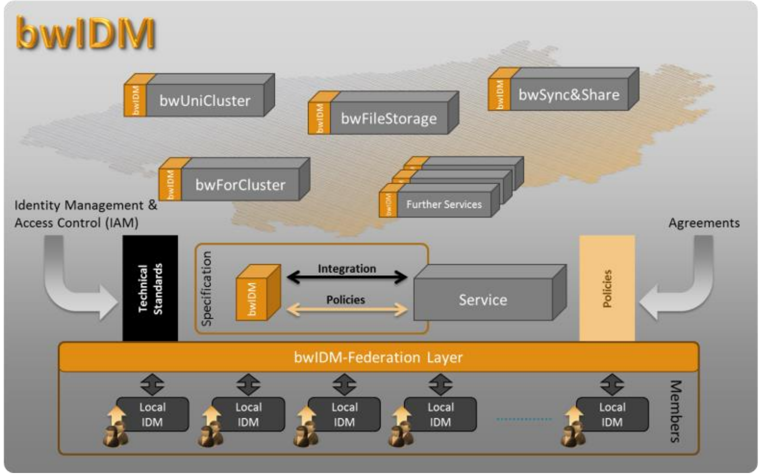
- Main goals
  - Provision of storage and archival resources for multiple research disciplines
    - Systems biology, climatology, synchrotron research/materials science, humanities, geo/earth science, …
  - Basis for BaWü-wide data services
- Resources and access
  - 6 PB of on-line storage (ext. to > 10 PB)
  - 6 PB of archival storage
  - 100 GbE connection between LSDF@KIT and U-Heidelberg
  - Hadoop analysis cluster
  - General-purpose and specialized software
  - Connection to HPC clusters in BaWü
  - Jointly funded by Helmholtz and state



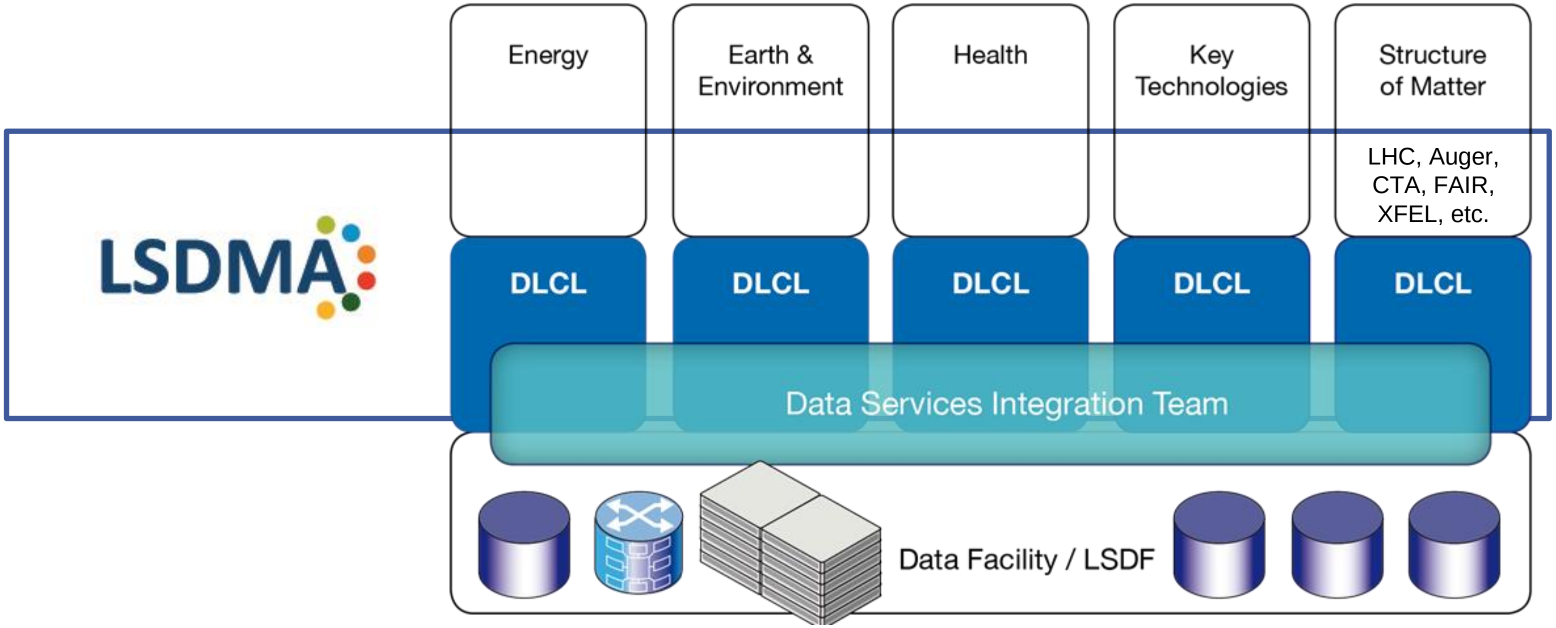
bwSync&Share on-top of LSDF

- bwSync&Share
  - Dropbox-like data storage for employees and students at Universities in Baden-Württemberg (500k people)
  - On-premise solution (data is stored at KIT)
  - Data can be shared with users not registered for the service
  - Access via webbrowser, clients for windows, linux and mobile phones
  - Based on PowerFoulder (German startup company)

- Based on state-wide identity federation
  - Building on Shibboleth
  - Uplink to DFN-AAI
  - LDAP Façade allows command line based clients to use web authentication



Large-Scale Data Management and Analysis – Dual Approach



Data Life Cycle Labs

Joint R&D with scientific user communities

- Optimization of the data life cycle
- Community-specific data analysis tools and services

Data Services Integration Team

Generic methods R&D

- Data analysis tools and services common to several DLCLs
- Interface between federated data infrastructures and DLCLs/communities

Facts and Figures

LSDMA

KIT Karlsruhe Institute of Technology

HELMHOLTZ ASSOCIATION

- Helmholtz portfolio extension
- Initial project duration: 2012-2016
- Partners:

KIT Karlsruhe Institute of Technology · JÜLICH FORSCHUNGSZENTRUM · DESY · GSI

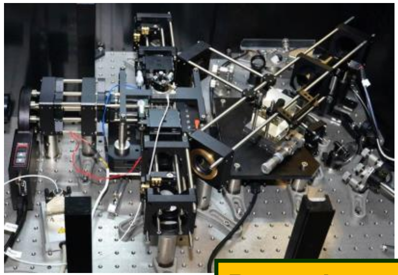
TECHNISCHE UNIVERSITÄT DRESDEN · DKRZ DEUTSCHES KLIMARECHENZENTRUM · RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG · ulm university universität uulm · UHH Universität Hamburg · htw. Hochschule für Technik und Wirtschaft Berlin University of Applied Sciences · GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Sustainability
  - Inclusion of activities into Helmholtz programme-oriented funding (PoF) “Supercomputing & Big Data” from 2015 onwards
  - Cross-programme iniative with other Helmholtz research fields
- Annual international symposium

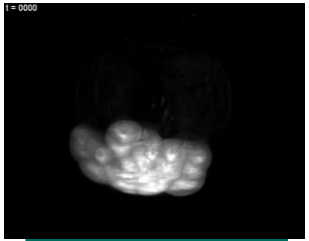
Selected Scientific Highlights

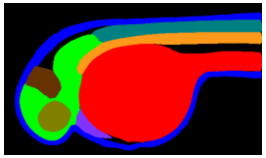
- **DLCL Key Technologies** (with U-Heidelberg, U-Dresden)
  - Light Sheet Microscopy

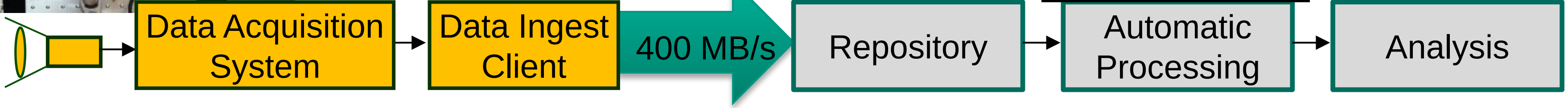
Mikut et al. (2013), Automated processing of zebrafish imaging data - a survey, **Zebrafish 10(3)**, DOI:10.1089/zeb.2013.0886



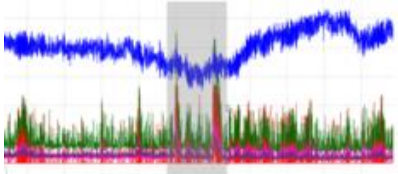
Experimental Microscope 16 TB/d

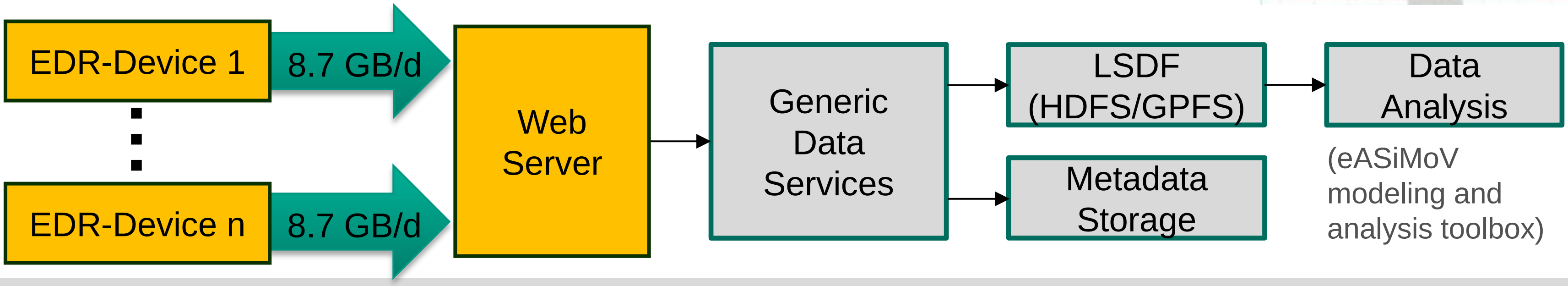






- **DLCL Energy** (with U-Ulm)
  - Secure data management for tech. & eco. data analysis
  - Electrical Data Recorder: 3-phase voltage measurement

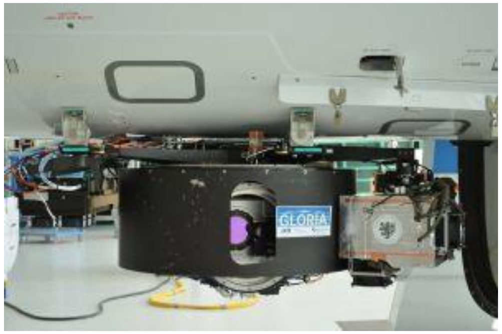


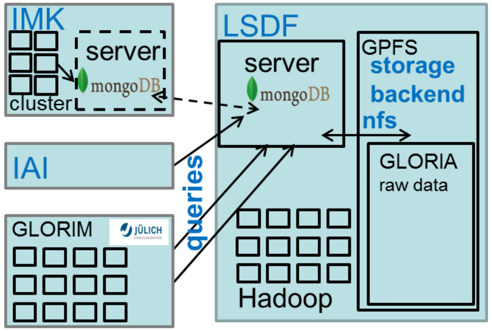


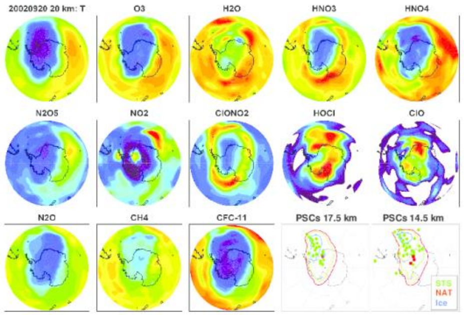
(eASiMoV modeling and analysis toolbox)

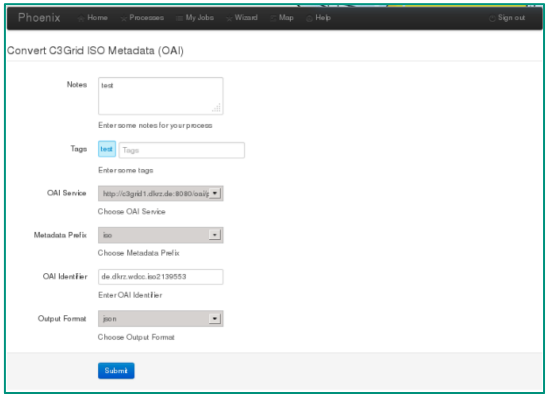
Selected Scientific Highlights (II)

- **DLCL Climatology**
  - GLORIA project
    - Spectrometer on aircraft (DLR)
    - NoSQL-DB
    - Measurements campaign in 2015
  - SAT project
    - Data life cycle management for satellite data
    - MIPAS/Envisat, MLS/Aura
    - Real-time response analysis framework
  - Geospatial data life cycle framework
    - Joint project with DKRZ
    - Portal for data transport, quality control, PIDs, publication of results and data









Convert C3Grid ISO Metadata (OAI)

Notes: test
Enter some notes for your process

Tags: test Tags
Enter some tags

OAI Service: http://c3grid1.dkrz.de:8080/oai/p
Choose OAI Service

Metadata Prefix: iso
Choose Metadata Prefix

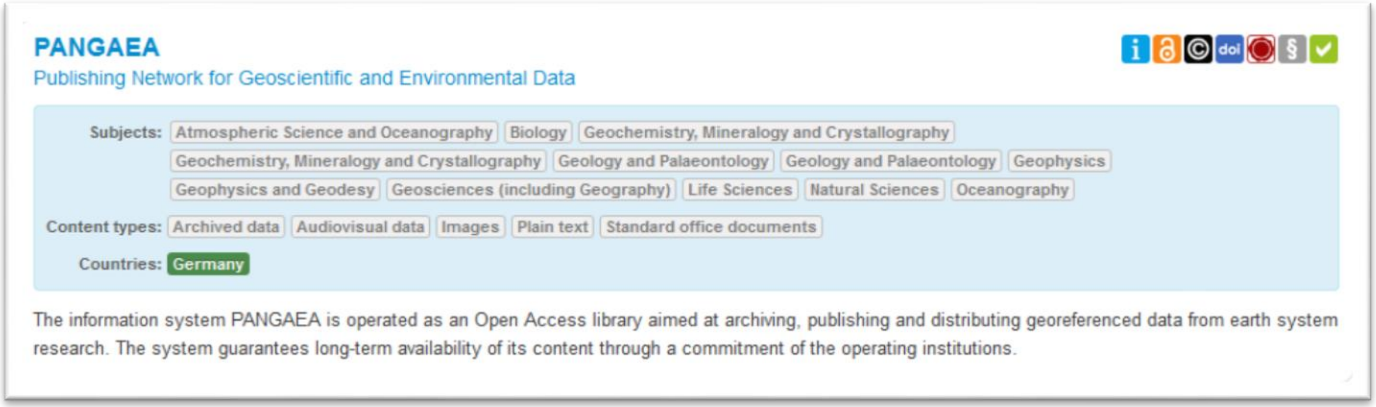
OAI Identifier: de.dkrz.wdcc.iso2139553
Enter OAI Identifier

Output Format: json
Choose Output Format

Submit

Re3data.org (KIT-Library, Frank Scholze)

- Goals
  - Linking existing research data repositories
  - In-depth analysis of quality requirements for research data repositories
  - Define a draft set of criteria for their quality assurance
- Currently 634 research data repositories from around the world covering all academic disciplines are listed
  - 586 of these are described using the re3data.org schema, http://doi.org/10.2312/re3.005



PANGAEA

Publishing Network for Geoscientific and Environmental Data

Subjects: Atmospheric Science and Oceanography | Biology | Geochemistry, Mineralogy and Crystallography | Geochemistry, Mineralogy and Crystallography | Geology and Palaeontology | Geology and Palaeontology | Geophysics | Geophysics and Geodesy | Geosciences (including Geography) | Life Sciences | Natural Sciences | Oceanography

Content types: Archived data | Audiovisual data | Images | Plain text | Standard office documents

Countries: Germany

The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. The system guarantees long-term availability of its content through a commitment of the operating institutions.

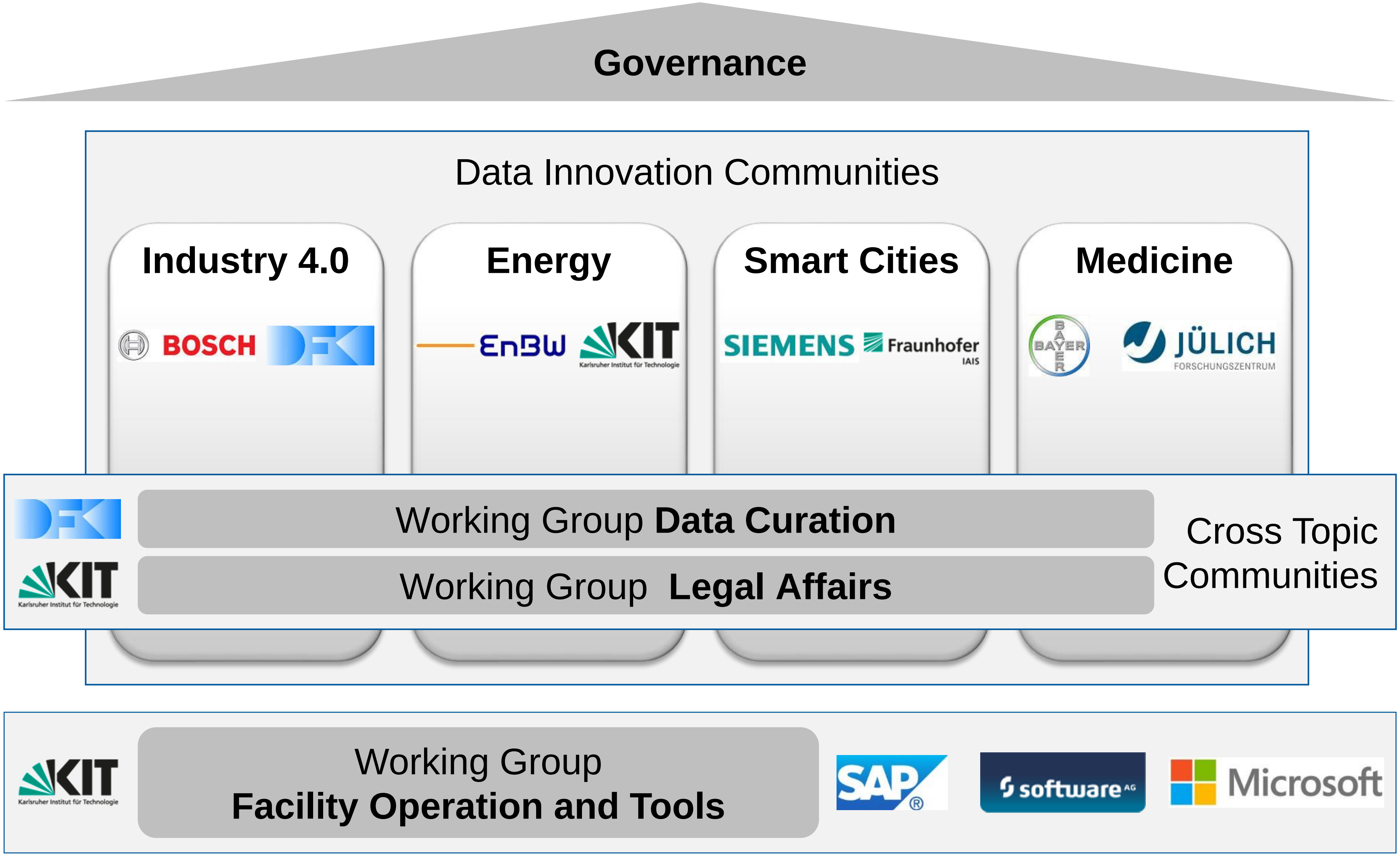
FUNDING

NETWORK

Smart Data Innovation Lab: A Data Life Cycle Hub for Industry

Smart Data Innovation Lab

KIT Karlsruhe Institute of Technology



Governance

Data Innovation Communities

Industry 4.0

BOSCH DFKI

Energy

EnBW KIT

Smart Cities

SIEMENS Fraunhofer IAIS

Medicine

BAYER JÜLICH FORSCHUNGSZENTRUM

DFKI KIT

Working Group **Data Curation**

Working Group **Legal Affairs**

Cross Topic Communities

KIT

Working Group

Facility Operation and Tools

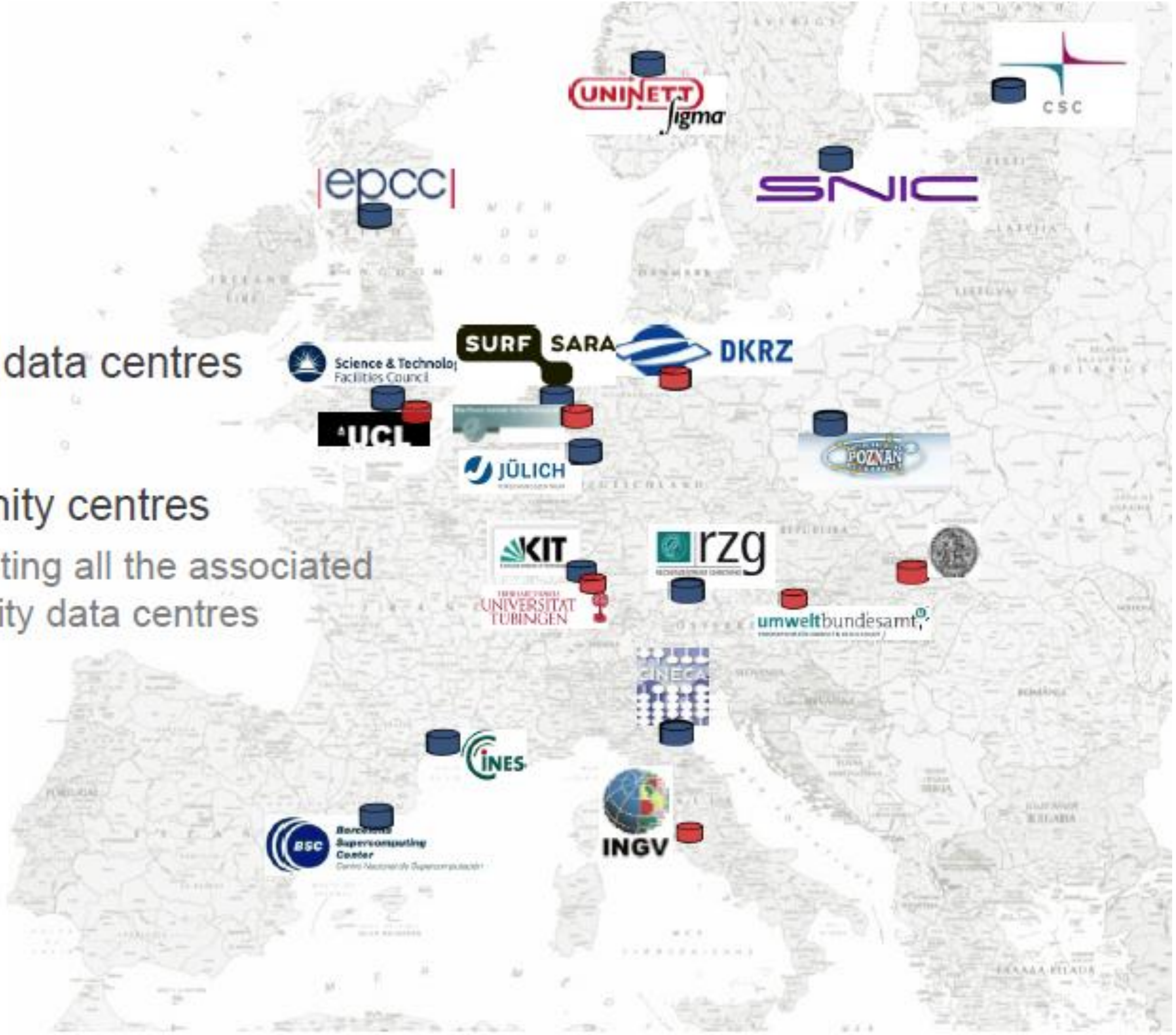
SAP software AG Microsoft

EUDAT – Building a European Collaborative Data Infrastructure

- Objectives
  - Cost-efficient and high quality Collaborative Data Infrastructure (CDI)
  - Meeting users’ needs in a flexible and sustainable way
  - Across geographical and disciplinary boundaries
- Facts & Figures
  - Started 1.10.2011
  - Duration 36+6 months
  - 16.3 M€ (9.3 M€ EC)
- Consortium
  - National data centres
  - Technology providers
  - Research communities
- www.eudat.eu

25 European partners

A pan-European Infrastructure

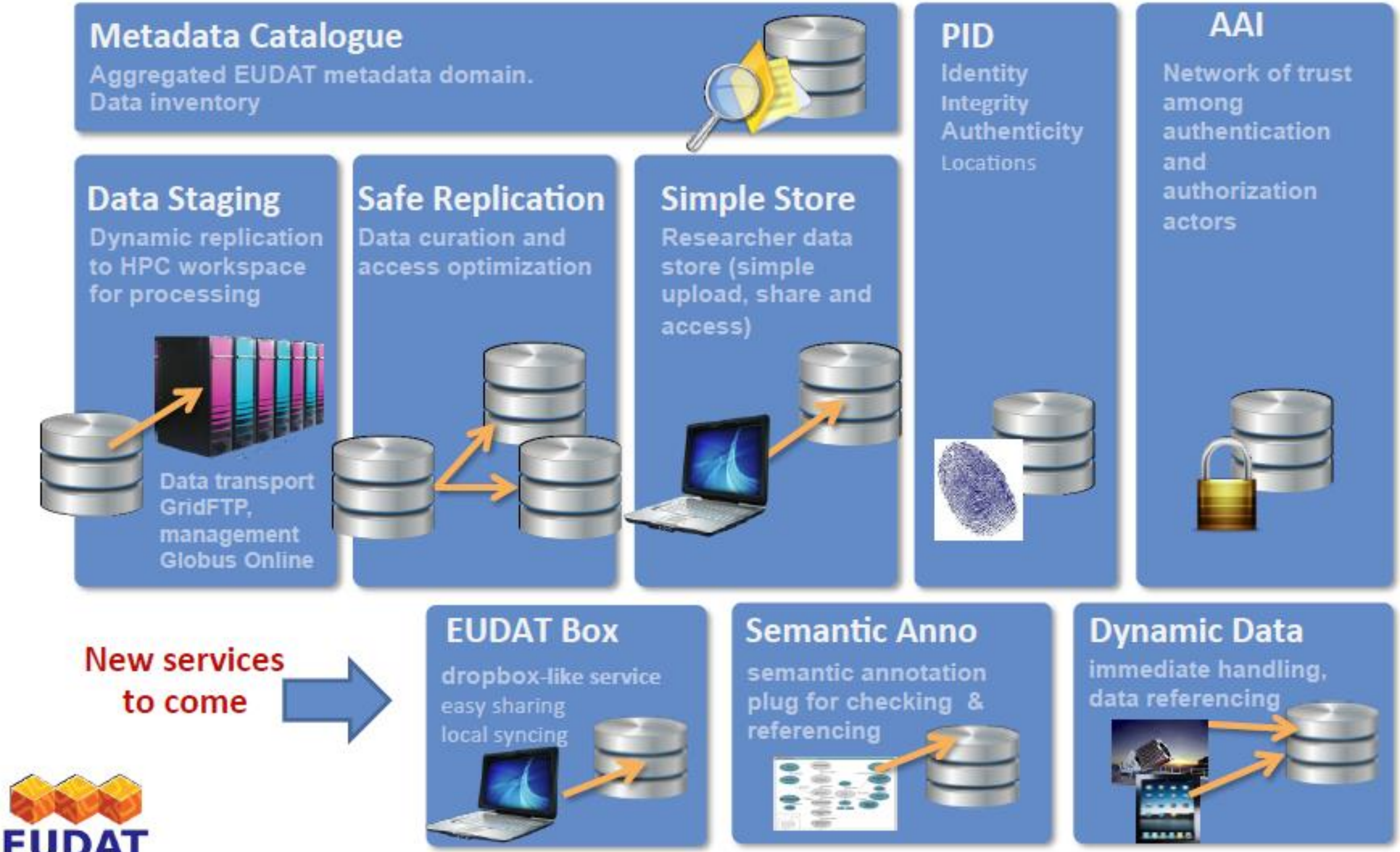


🟦 general data centres

🟥 community centres
representing all the associated community data centres

EUDAT

Selected Services



Metadata Catalogue

Aggregated EUDAT metadata domain.
Data inventory

PID

Identity
Integrity
Authenticity
Locations

AAI

Network of trust among authentication and authorization actors

Data Staging

Dynamic replication to HPC workspace for processing

Data transport GridFTP, management Globus Online

Safe Replication

Data curation and access optimization

Simple Store

Researcher data store (simple upload, share and access)

New services to come

EUDAT Box

dropbox-like service
easy sharing
local syncing

Semantic Anno

semantic annotation plug for checking & referencing

Dynamic Data

immediate handling, data referencing

Experiences

- Communities demand for
  - Storage, analysis and archival facilities
  - Sharing data with other groups
  - Data safety and preservation
  - ‘Consulting’

- Needs are driven by
  - Increasing amount of data
  - Cooperation between groups
  - Policies
  - Open access/data
  - Long-term preservation

- Communities differ in
  - Previous knowledge
  - Level of specification of the data life cycle
  - Tools and services already used

- Lessons learned
  - Interoperable AAI crucial
  - Data privacy very challenging, both legally and technically
  - Communities need evolution, not revolution
  - Needs can be very specific