On Trust

From an MPG and EUDAT Perspective

Raphael Ritz, RZG

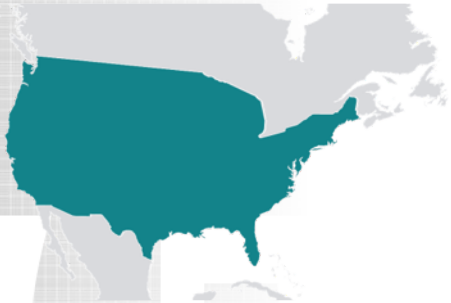
Stockholm, June 4, 2014

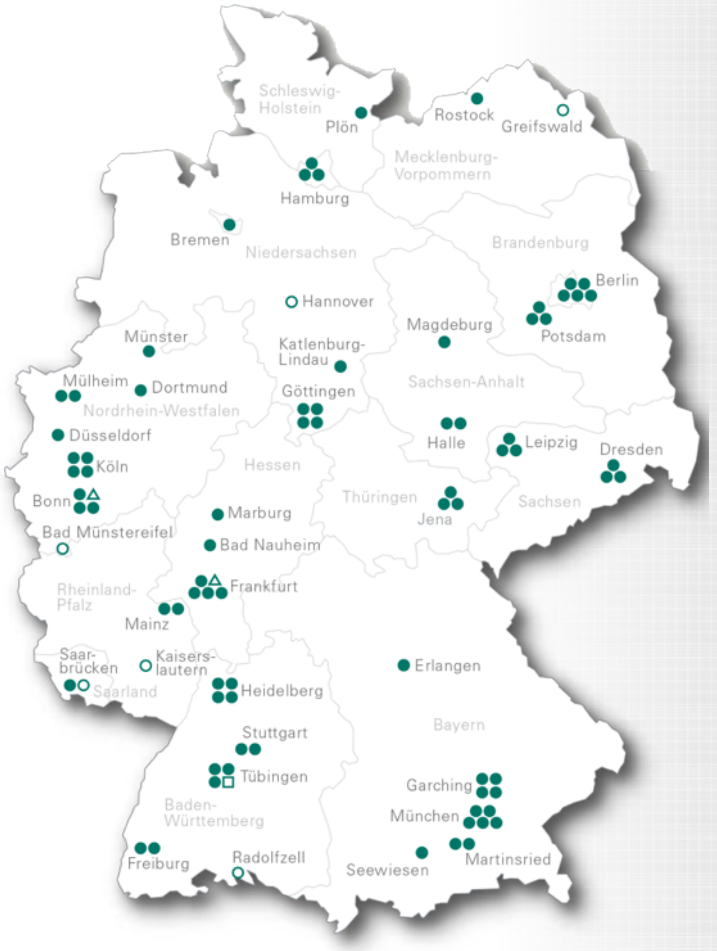
MPG: Mission and Guiding Principles

- Basic Research at cutting-edge, strictly curiosity-driven and quality oriented
- Autonomy, where scientists decide upon science
- „Harnack Principle“: People not programs
- Flexible, dynamic, interdisciplinary MPIs
- Long-term trust systems with significant core funding for high-risk projects
- Quality assurance by peers

Sites of Max Planck Research Institutes and Associated Institutes

RECHENZENTRUM GARCHING





MAX PLANCK INSTITUTES ABROAD

ITALY
Bibliotheca Hertziana, Rome
Art History Institute, Florence

THE NETHERLANDS
Max Planck Institute for Psycholinguistics, Nijmegen

USA, FLORIDA
Max Planck Florida, Jupiter

LUXEMBOURG
Max Planck Institute Luxembourg for International, European and Regulatory Procedural Law, Luxembourg

MPG: Facts & Figures

RECHENZENTRUM GARCHING

STAFF MEMBERS

» **21.405**

SCIENTISTS

In total, the workforce of the Max Planck Society consists of 21,405 employees, including 5,470 scientists as well as 4,487 guest scientists and grantholders.

BUDGET

» **1.527,7 Mio. EUR**

MPG: Human Sciences Section

RECHENZENTRUM GARCHING

- CULTURAL STUDIES
- JURISPRUDENCE
- SOCIAL SCIENCES
- COGNITIVE RESEARCH
- LINGUISTICS

© Bibliotheca Hertziana / A. Jemolo

MPG: Biology and Medicine Section

RECHENZENTRUM GARCHING

- DEVELOPMENTAL AND EVOLUTIONARY BIOLOGY & GENETICS
- IMMUNOBIOLOGY AND INFECTION BIOLOGY & MEDICINE
- BEHAVIORAL SCIENCES
- MICROBIOLOGY & ECOLOGY
- NEUROSCIENCES
- PLANT RESEARCH
- STRUCTURAL AND CELL BIOLOGY
- PHYSIOLOGY

© MPI for Developmental Biology

MPG: Chemistry, Physics and Technology Section

RECHENZENTRUM GARCHING

- ASTRONOMY & ASTROPHYSICS
- CHEMISTRY
- SOLID STATE RESEARCH & MATERIAL SCIENCES
- EARTH SCIENCES AND CLIMATE RESEARCH
- PARTICLE, PLASMA AND QUANTUM PHYSICS
- COMPLEX SYSTEMS
- COMPUTER SCIENCE
- MATHEMATICS

© MPI for Radio Astronomy

RZG: Compute and Data Intensive Sciences in the MPG

RECHENZENTRUM GARCHING

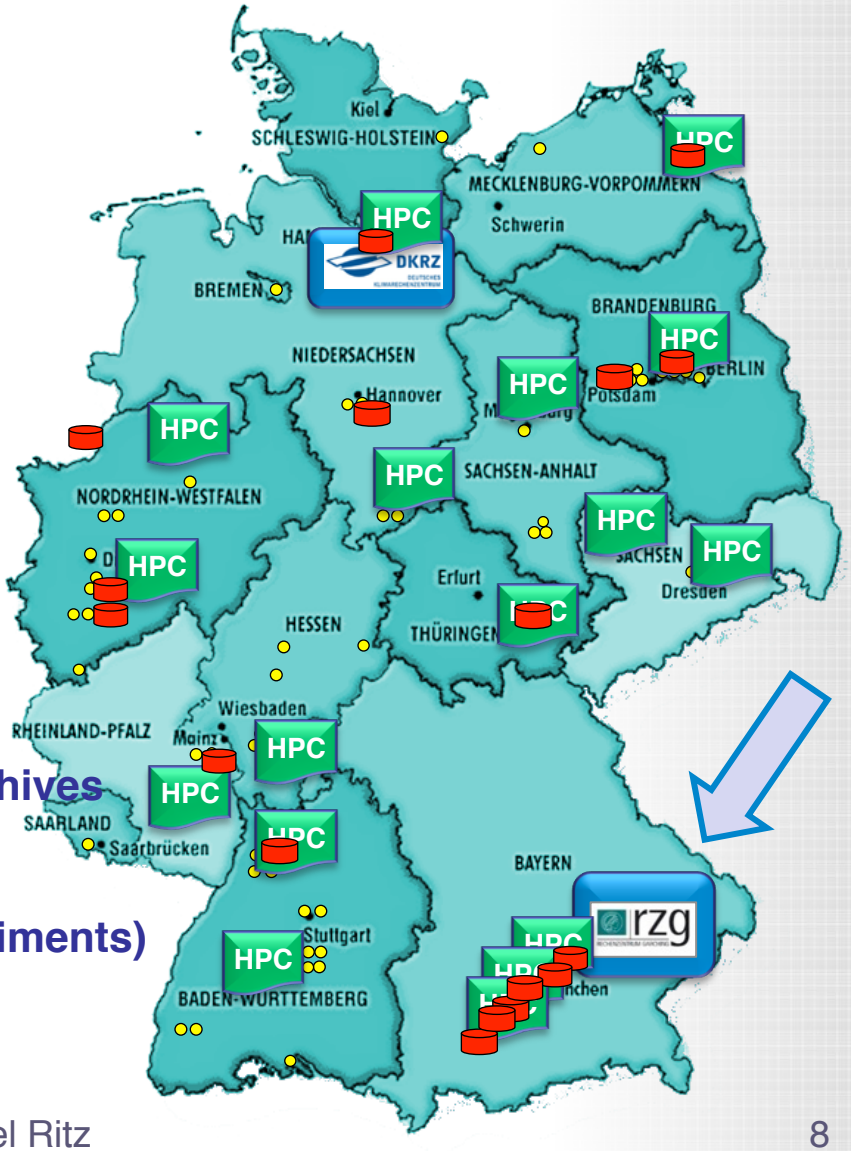
HPC meets Data Intensive Computing

- Material Sciences
- Astrophysics, Cosmology
- Earth System Sciences
- Fusion, Plasmaphysics
- Life Sciences, Biomedicine
- High Energy Physics

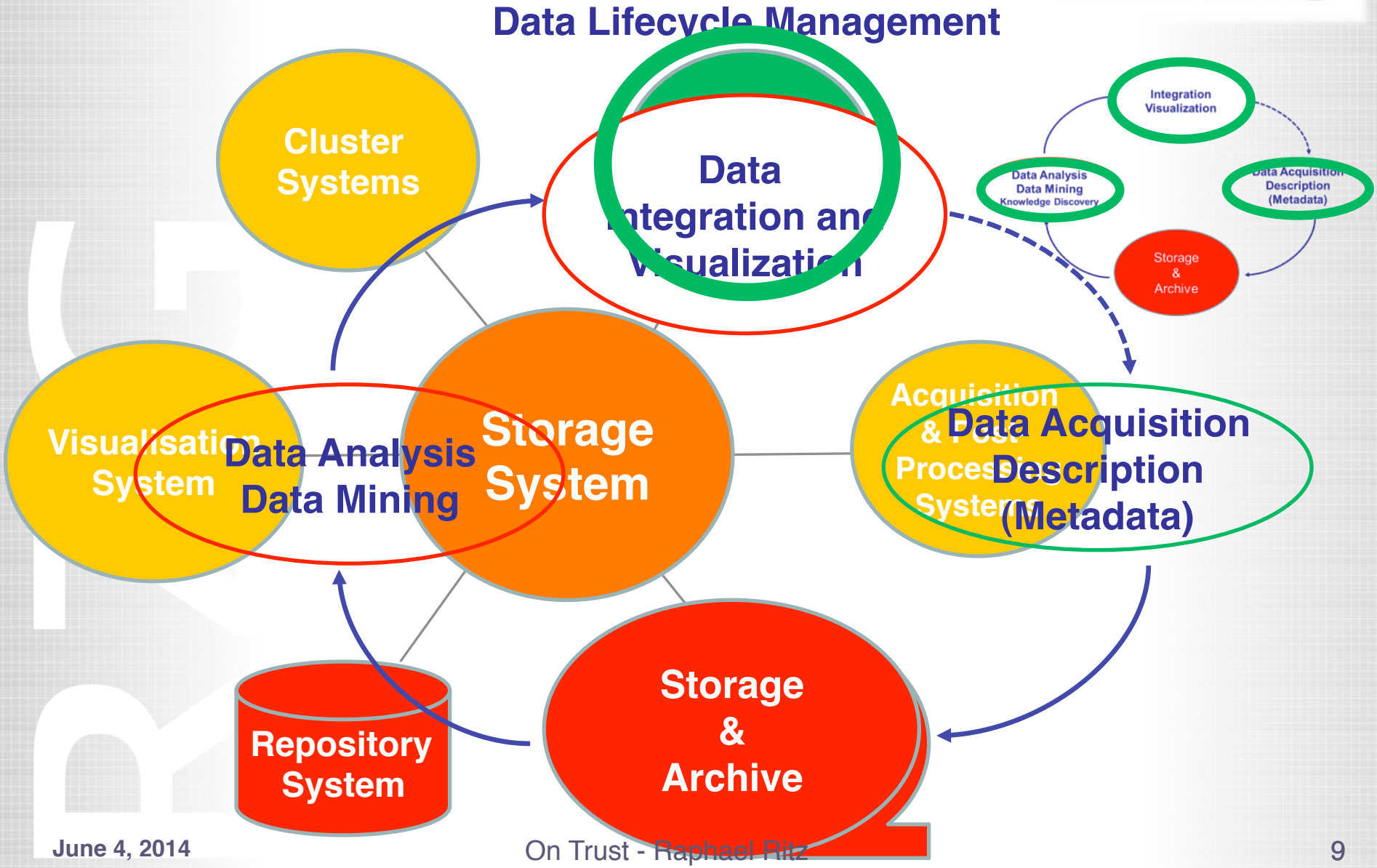
Simulation ↔ Experiment

Garching Computing Centre – RZG

- HPC Facilities and Applications Support
- Data Management, Backups, Long-Term Archives
- File Systems Technologies
- Data Processing and Visualisation
- Data Acquisition Support (e.g. Fusion Experiments)
- Data Science Services



RZG: Integrated Environment



RZG: Data Services

File Systems and Data Bases: **Provisioning, Expertise**

RECHENZENTRUM GARCHING

Filesystems:

- AFS — Andrew File System
- GPFS — General Parallel File system
- GHI — GPFS with policy-based migration functionality
- HDFS — Hadoop Distributed File System

Main Data Base technologies:

- Oracle, DB2
- Objectivity, noSQL DBs (Hbase, MongoDB)
- Postgres, MySQL, MS SQL

Data Services

RECHENZENTRUM GARCHING

Tools for Data Management, Organisation, Acquisition, Access and Analysis

Grid-Middleware
- glite, Globus, Unicore

Data Management
- GHI & HPSS, TSM, dCache, iRODS

Applications Development and Web-Services
- E.g. Tomcat, JBOSS, Django

Global Identifier Service (Persistent Identifiers)
- Handle Service at RZG (EPIC enabled)

Site and Service Registry Service
- EGEE/EGI GOCDB
- Service provided for EUDAT and explored for PRACE

General Data Services

- Long-Term Archiving
  Bit-preservation for 50 years at least

- Meta Data Support
  Discipline- and project-specific support

- “Data Enabling” projects
  Structuring/Classification of processed data,
  Provisioning e.g. via Data Bases
  Access via Web Interfaces

RZG: General Data Services

Example 1: Long-Term Archiving since 30 years

- Experiment data from IPP
  since 1980
- Satellite data from MPI for extraterrestrial Physics
  in collaboration with the NASA
  since 1991
- Telescope data of MPI for Physics (Magic Project)
  since 2003

RZG: General Data Services

Example 2: Long-Term Archiving (for > 50 years)

Arts and humanities institutes

- Video/Audio documents (*MPI for Psycholinguistics)*
- Picture collections from the *Biblioteca Hertziana, Rom*
- Picture collections from „*Deutsches Kunsthistorisches Institut“, Florenz*
- Human-Ethology movie archive of the *MPG, Andechs*

RZG: Long History of Automatic Migration Software

And the Provisioning of Mass Storage Systems

RECHENZENTRUM GARCHING

1971 - 1989 AMOS

1981 - 1996 HADES

1993 - 1995 AMOS 2

1993 - 1999 DMF

1994 - 2008 mr-afs

since 1998 TSM-HSM

since 2008 afs-osd

since 2011 HPSS

Data Growth

from 15 TB to 15 PB
in 11 years
(from 4/2001 to 4/2012)

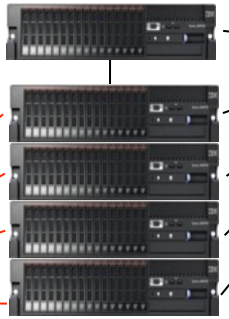
RZG: Data Infrastructure

RECHENZENTRUM GARCHING

2x SL8500 Libraries

scalable data movers

HPSS core + movers
with large disk cache



FC switches

10 Gb/s switch

GPFS-HPSS-Interface

4 GHI nodes

GPFS

remote GPFS mount

HPC

remote site (LRZ)

TS3500 Library

HPSS remote
mover

IRODS UMSS

Linux AFS server

Where is Trust in RZG Arising From?

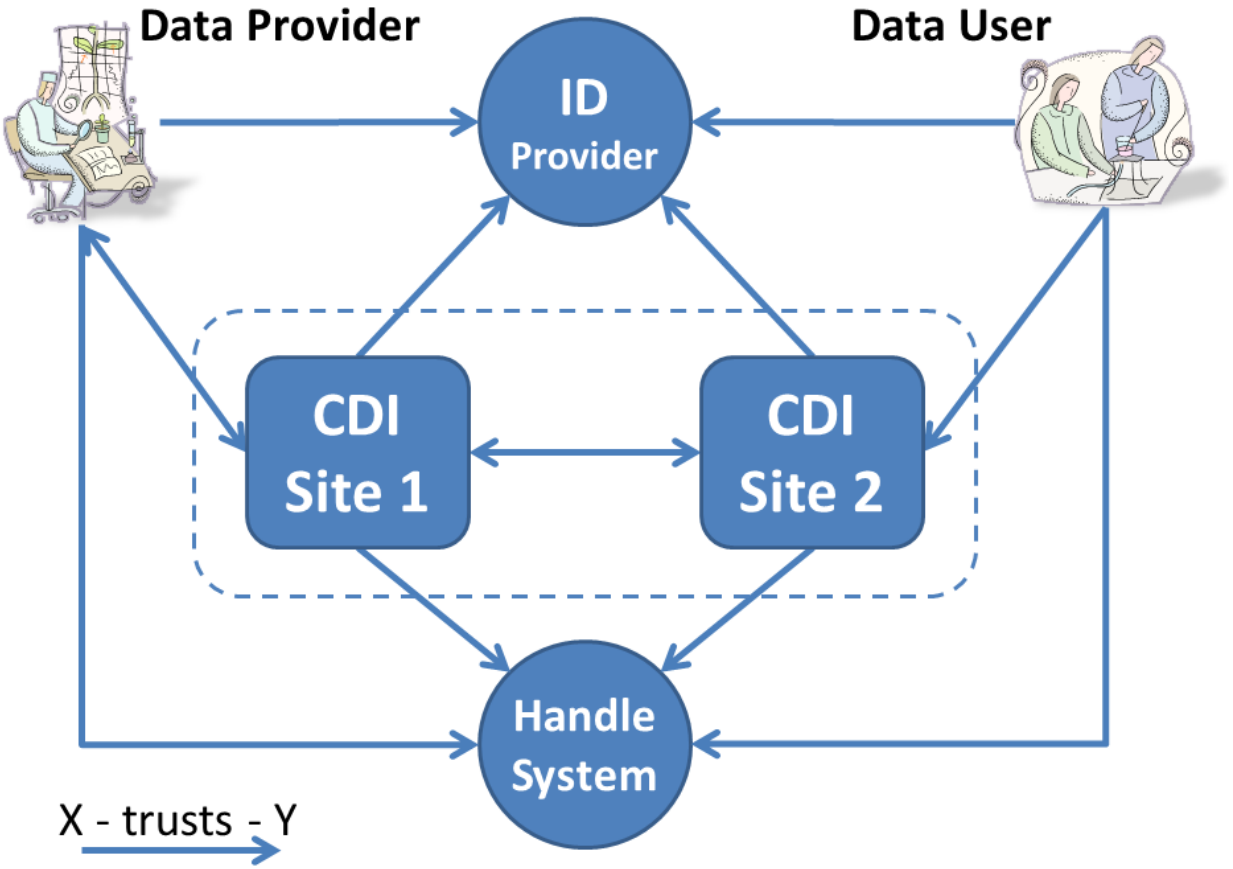
RECHENZENTRUM GARCHING

- RZG has been around for a while
- RZG has strong institutional backing (incl. funding)
- RZG has strong personal relationships
- RZG has shown that it can manage and scale
- RZG has strong partners (e.g., LRZ, DKRZ)

How can this be translated into a distributed environment?

What else is needed?

Chains of Trust in the CDI Network



CDI: Common Data Infrastructure

Leverage Existing Islands of Trust

- Community Centers often already have preferred Data Centers, often within the same country
- Trust often results from years of collaboration more than through formal agreements
- It is crucial for EUDAT to have strong partners across many countries

Summary
Aspects of Trust

- cultural, organizational, legal and ethical considerations based on experiences,
- long-term persistence of service offers and proper stewardship of data without creating dependencies on one institution or company,
- proper management of data objects and collections including their identities and properties stored in persistent identifier and metadata records,
- quality of data content and of repository procedures and applying appropriate assessment procedures.

EUDAT CDI Challenges

- offering common data services on a distributed network of established data centers, which are based on reliable organizational, national and/or European funding statements,
- utilizing existing relationships but nevertheless offering the strength of redundancy through a network,
- requesting to base its services on data objects and collections that have persistent unique identifiers (PIDs) and metadata associated with these,
- turning its operations that are often hidden in executable code stepwise to explicit policy rules,
- requesting all centers to participate in data quality assessments according to recommended procedures.

RZG: MPG Institutes with large Data Projects

(active or in preparation)

RECHENZENTRUM GARCHING

MPI for Meteorology (Hamburg)

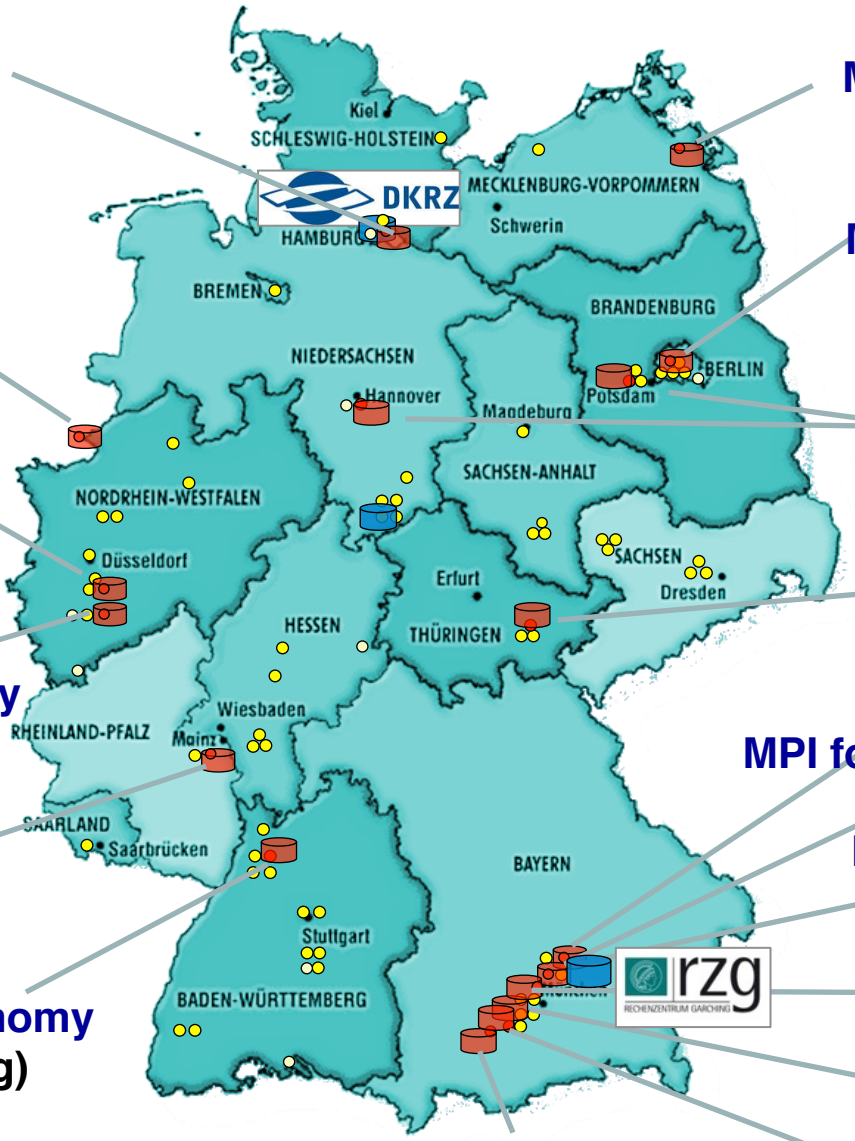
MPI for Psycholinguistics (Nijmegen)

MPI f. Plant Breeding Research (Köln)

MPI for Radio Astronomy (Bonn)

MPI for Chemistry (Mainz)

MPI for Astronomy (Heidelberg)



MPI for Plasmaphysics (Greifswald)

MPI for Molecular Genetics (Berlin)

MPI for Gravitational Physics (Potsdam, Hannover)

MPI for Biogeochemistry (Jena)

MPI for Plasmaphysics (Garching)

MPI for Extraterrestrial Physics (Garching)

MPI for Astrophysics (Garching)

MPI for Physics (München)

MPI for Neurobiology (Martinsried)

MPI for Biochemistry (Martinsried)

MPI for Ornithology (Seewiesen)

MPI for Neurobiology: Dr. Helmstaedter

in preparation

TECHNOLOGY FEATURE

CHARTING THE BRAIN'S NETWORKS

The field of connectomics is pulling neuroscience into a speedy, high-throughput lane that is generating vast amounts of data.

ALLEN INST. BRAIN SCI.

Massive stores of brain-tissue slides are providing a resource for scientists working on mapping neural networks.

BY VIVIEN MARX

Researchers seeking to understand the brain want big data. And they are getting them. Just as geneticists have moved from genes to genomes to the interacting network of factors that regulate and modify the genome, neuroscientists are going from studying single neurons to tracing how vast neuronal networks connect and interact. "I think this is a really exciting field," says neuroscientist Moritz Helmstaedter at the Max Planck Institute for Neurobiology in Martinsried, Germany, who is working to obtain a cell-level overview of the neuronal connections — the connectome — of the mammalian cortex. "Many people are pretty ambitious about breaking the next barrier in understanding how the brain works by using this new field of connectomics."

Sceptics argue that current methods lack the power to map the massively interconnected web of around 100 billion neurons in the human brain. Even if technology can rise to the challenge, they say, it is impossible to decipher so much data.

Clay Reid, a neuroscientist at Harvard Medical School in Boston, Massachusetts, and recently appointed as a senior investigator at the Allen Institute for Brain Science in Seattle, Washington, counters detractors by pointing to recent progress in neuroscience. A few years ago, it was nearly impossible ▸

11 OCTOBER 2012 | VOL 490 | NATURE | 293

- Electron microscopy of tissue sections
- Data acquisition at Martinsried,
  many parallel data streams
  Data Staging via 10 Gbit-link into migrating file system at RZG.
- Data volume: 10 – 100PB
- Network expanded to 100Gbits/s
  in cooperation with the Münchner Hochschulnetz

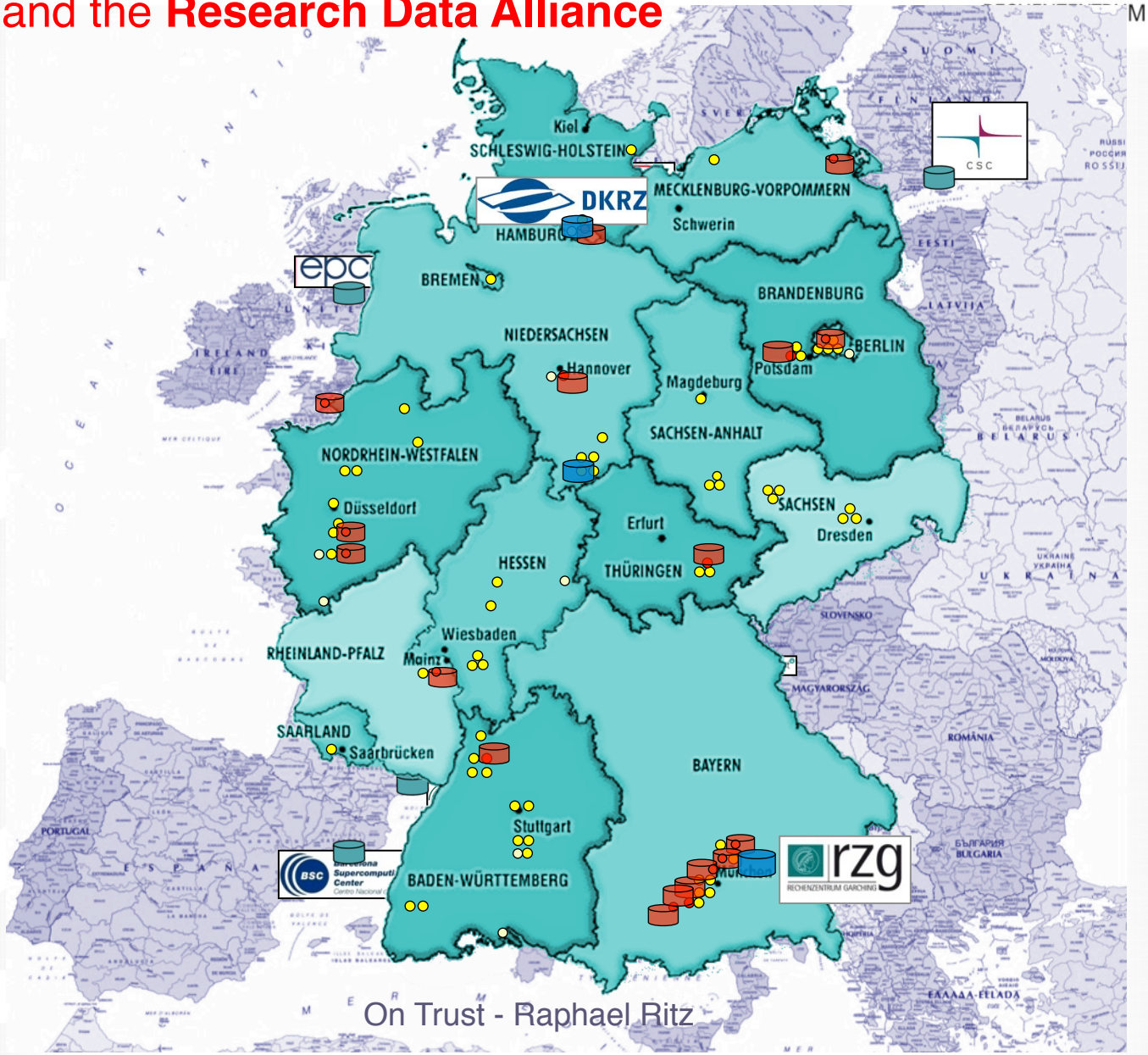
In preparation:

- GHI file system (HPSS)
- Data analysis on GPU clusters at RZG with mounted GHI file system
- Policy-based archiving of raw data

11 OCTOBER 2012 | VOL 490 | NATURE | 293

Embedded in a European context:

EUDAT and the **Research Data Alliance**



International Data Projects
with involvement of RZG

RECHENZENTRUM GARCHING

EUDAT
http://www.eudat.eu

Research Data Alliance Europe (iCORDI), RDA
http://europe.rd-alliance.org
http://www.rd-alliance.org

RESEARCH DATA ALLIANCE
Research Data Sharing without barriers

irods Consortium
http://irods-consortium.org
http://irods-consortium.org/announcement-of-new-partnership-between-the-irods-consortium-and-the-dice-group/