SCAI SuperComputing Applications and Innovation

Italian Scientific Big Data Initiative

Sanzio Bassini

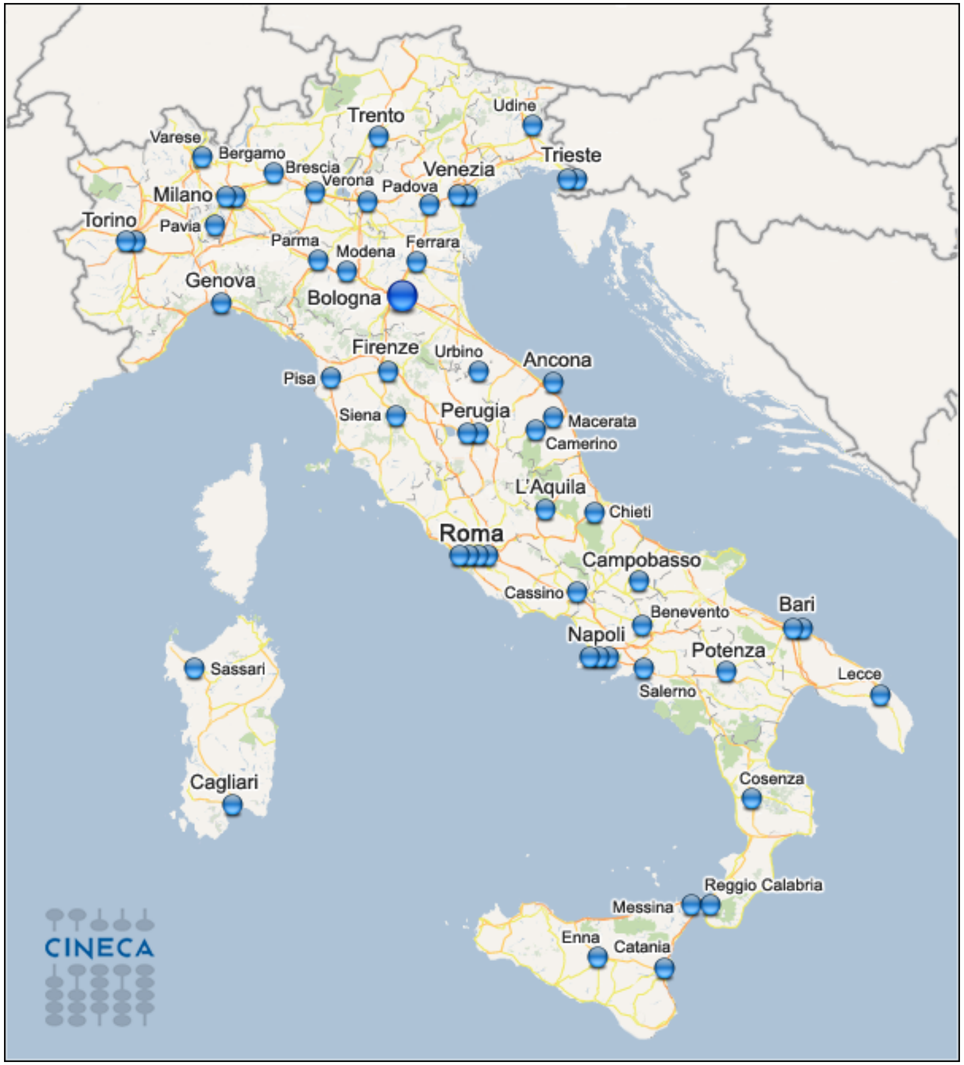
Director of Supercomputing Application & Innovation Department

S.Bassini@cineca.it

CINECA
Casalecchio di Reno (BO)
Via Magnanelli 6/3, 40033 Casalecchio di Reno | 051 6171411 | www.cineca.it

CINECA Interuniversity Consortium

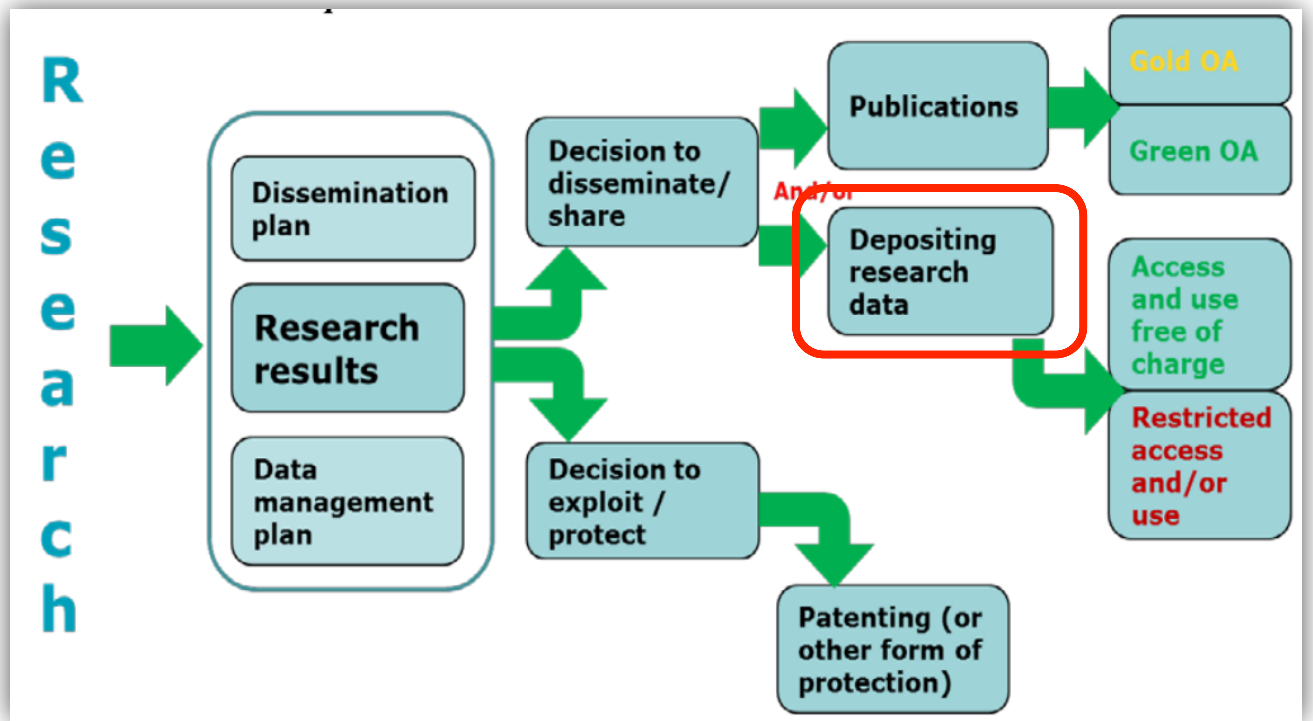
- 69 Italian Universities
- CNR, OGS, SISSA/ISAS Trieste
- Ministry of University and Research



Mission

- Private non for Profit Organization
- Founded in 1969 by Ministry of Public Education now under the control of Ministry Education, University and Research
- Main activities:
- Promote the use of the most advanced HPC systems to support **public** and **private** scientific and technological research
  - Services both for **Universities, National Research Institute, private enterprises** and **Ministry**
- Three operational site
  - Bologna (consortium headquarter), Milan, Rome
  - 711 employees
- Supercomputing, Application & Innovation Dpt.
  - Tier-0 and Tier-1 Supercomputing service
  - Data processing and data management for scientific big data
  - Enabling specialistic support for application development and exploitation

The focus



Research

Dissemination plan

Research results

Data management plan

Decision to disseminate/ share

Decision to exploit / protect

Publications

And/or

Depositing research data

Patenting (or other form of protection)

Gold OA

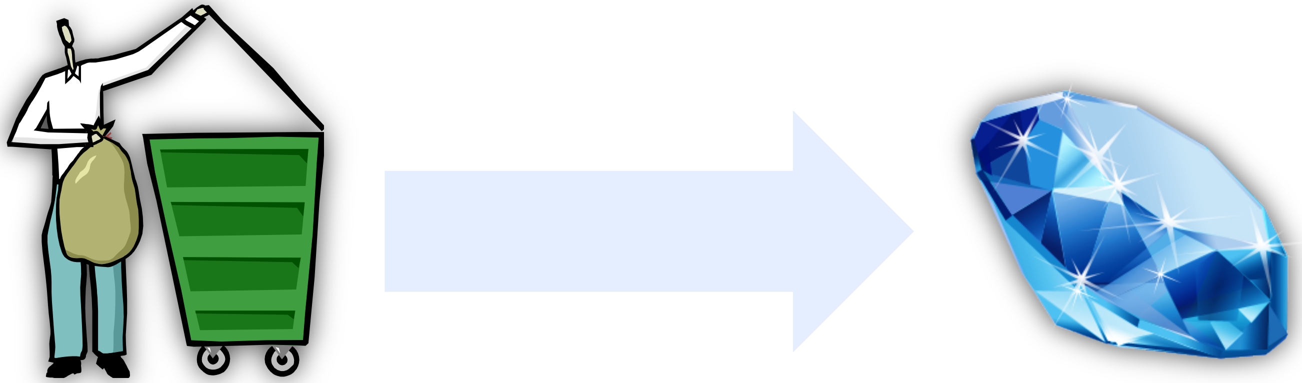
Green OA

Access and use free of charge

Restricted access and/or use

Data as an asset

“One of the most significant changes of the past decade has been the widespread recognition of data as an asset rather than the refuse of research.”



G8 + O5 White Paper
5 Principles for an Open Data Infrastructure

- **1. Discoverability**
  - Implementation of appropriate **persistent identifier** frameworks, adoption of descriptive metadata standards, and the use of appropriate data formats and taxonomies.
- **2. Accessibility**
  - **Publicly funded scientific research data should be made openly available** with as few restrictions as possible. Ethical, legal or commercial constraints may be imposed on the use of research data to ensure that the research process is not damaged by inappropriate release of data.
- **3. Understandability**
  - Scientific data sets must be understandable in order to be effectively used. A set of numbers, texts, pictures or even videos alone cannot be understandable without **additional context, semantics, data analysis tools, and algorithms**.
- **4. Manageability**
  - Data management policies and plans must make it clear **who is responsible for maintaining the availability of data and how the associated costs** are to be met including issues associated with curation, storage and services.
- **5. People**
  - A global approach to research data infrastructure requires a **highly skilled and adaptable workforce** and culture that is able to capture the available data and make it available to those that are able to use it appropriately.

G8+O5 Data working group and GSO Report Recommendations 9 & 10

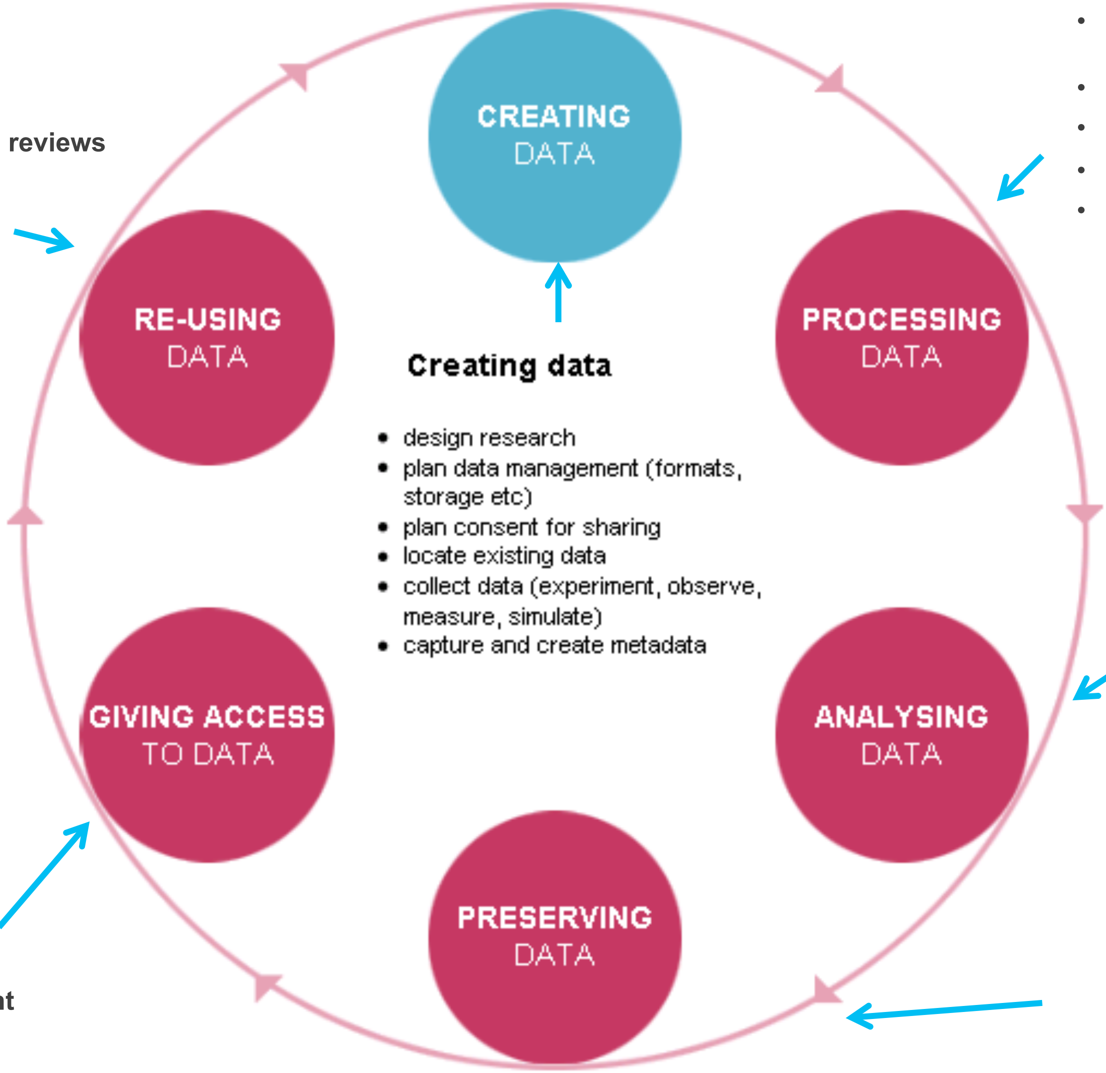
Recommendation 9 - e-infrastructure

Global research infrastructure initiatives should recognize the utility of the **integrated use of advanced e-infrastructures**, services for accessing and processing, and curating data, as well as remote participation (interaction) and access to scientific experiments.

Recommendation 10 - Data exchange

Global scientific data infrastructure providers and users should recognise the utility of **data exchange and interoperability** of data **across disciplines and national boundaries** as a means to broadening the scientific reach of individual data sets.

Research model data lifecycle



Creating data

- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

- Enter data, digitize, transcribe, translate
- Check, validate, clean data
- Anonymise data where necessary
- Describe data
- Manage and store data

- Interpret data
- Derive data
- Produce research outputs
- Author publications
- Prepare data for preservation

- Migrate data to best format
- Migrate data to suitable medium
- Backup and store data
- Create metadata and documentation
- Archive data

- Distribute data
- Share data
- Control access
- Establish copyright
- Promote data

- Follow-up research
- New research
- Undertake research reviews
- Scrutinize findings
- Teach and learn

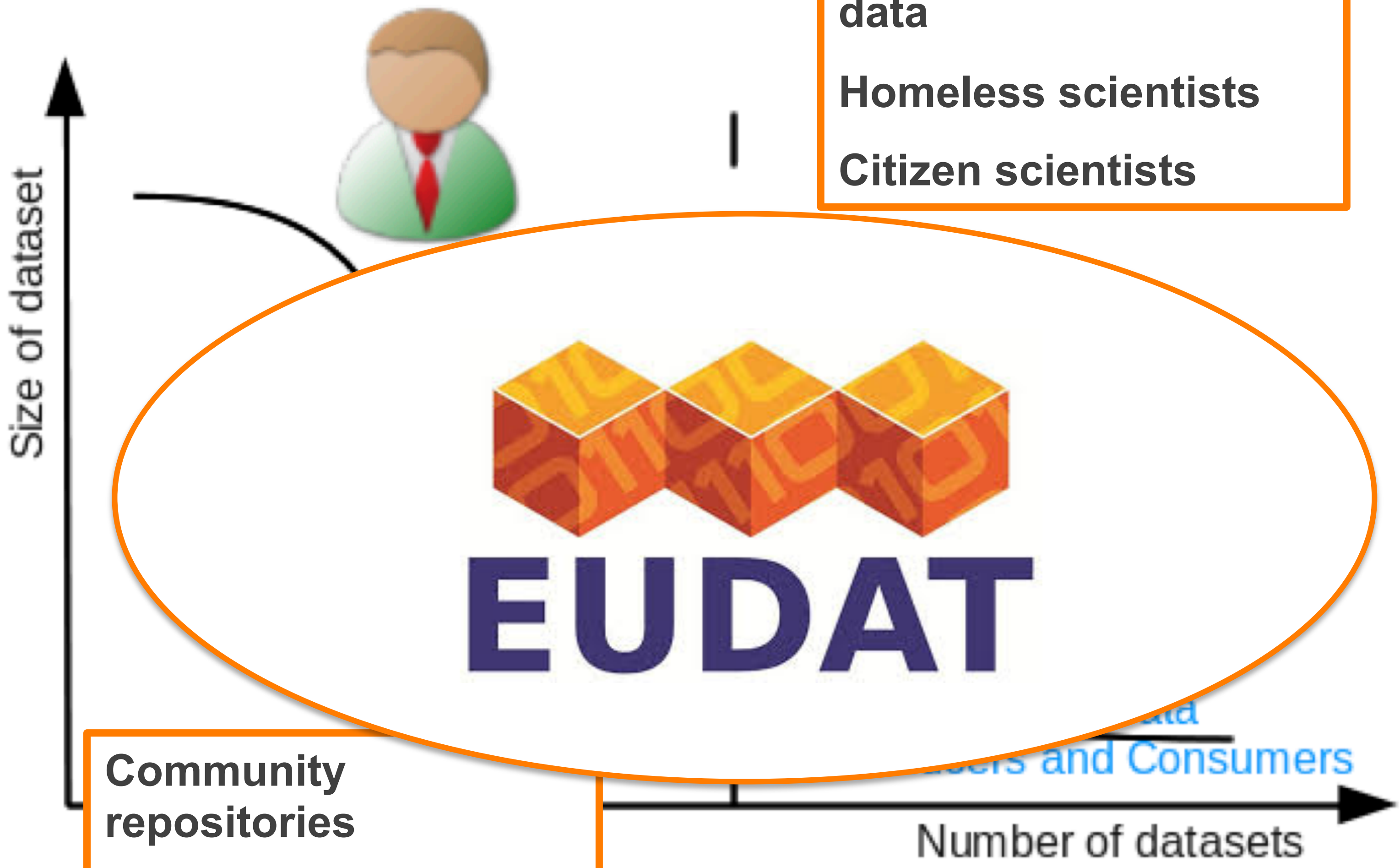
http://www.data-archive.ac.uk/create-manage/life-cycle

Data Domain

Scientists personal data

Homeless scientists

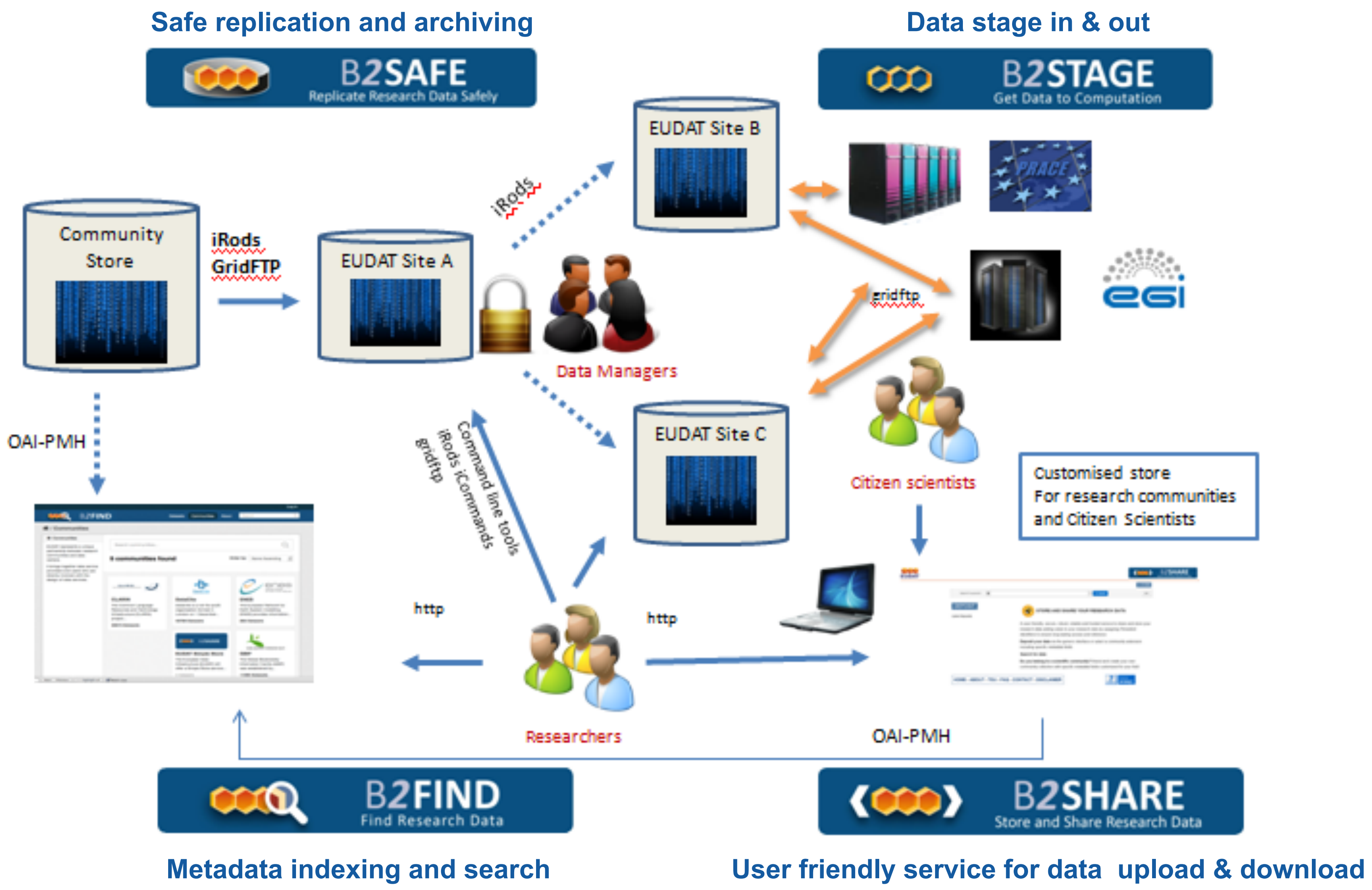
Citizen scientists



Community repositories

Institute repositories

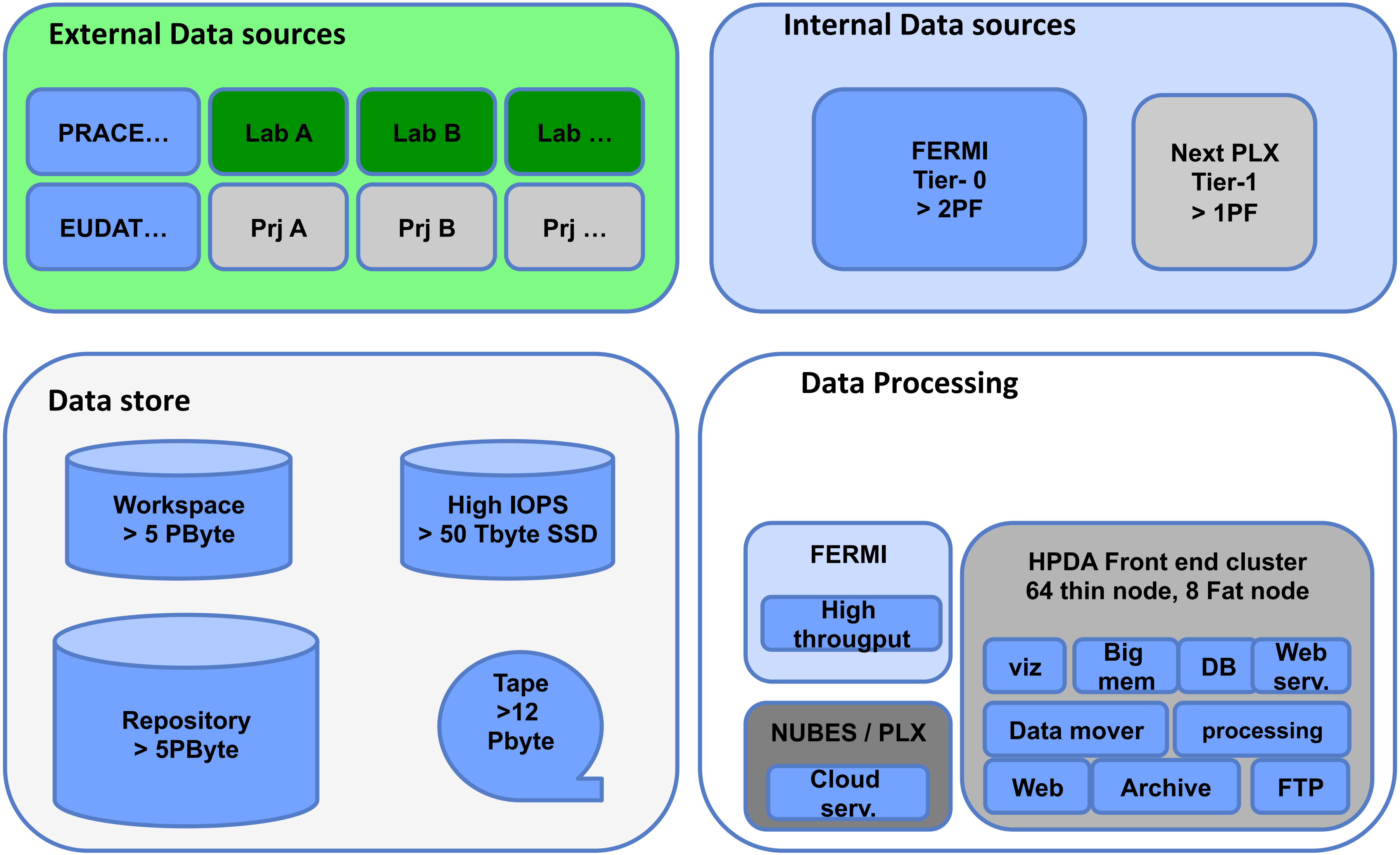
EUDAT available services



National Scientific Data Repository

- Aim of the project the creation of the national scientific data repository
- **Objectives**:
  - Service provision for national and EU funded project:
    - EUDAT (European Data Infrastructure)
    - V-MUST (Virtual Museum Transnational Network)
    - Connettomics – ICON Foundation,
    - LENS – Human Brain Project
    - Fluid dynamic – COST Action MP0806 Particles in turbulence
    - EuHIT - facilities for turbulence research
    - Astrophysics – Mission PLANK, Mission GAIA
    - Geophysics – INGV RI EPOS Project
    - NFFA / Sincrotrone - Nanoscience Foundries and Fine Analysis
    - Elixir – Italian National Infrastructure
    - Epigen – National Epigenetic infrastructure
    - Nextdata – National Environment data infrastructure
- New “business” marketplace
- Improve the processing of data (include the data scarcely structured)

CINECA Current infrastructure



EUDAT-EPOS success story

- The whole archive of the INGV's data center in Rome is replicated to CINECA, the EUDAT's data center in Bologna (years:1990-2013)
- Granularity is at single file level (the collection of sesimic waveforms related to 1 sensor for 1 day)
- Metadata are also replicated, as files (dataless).
- The replicas are stored, synchronized and accessed by a single user, the INGV community manager.

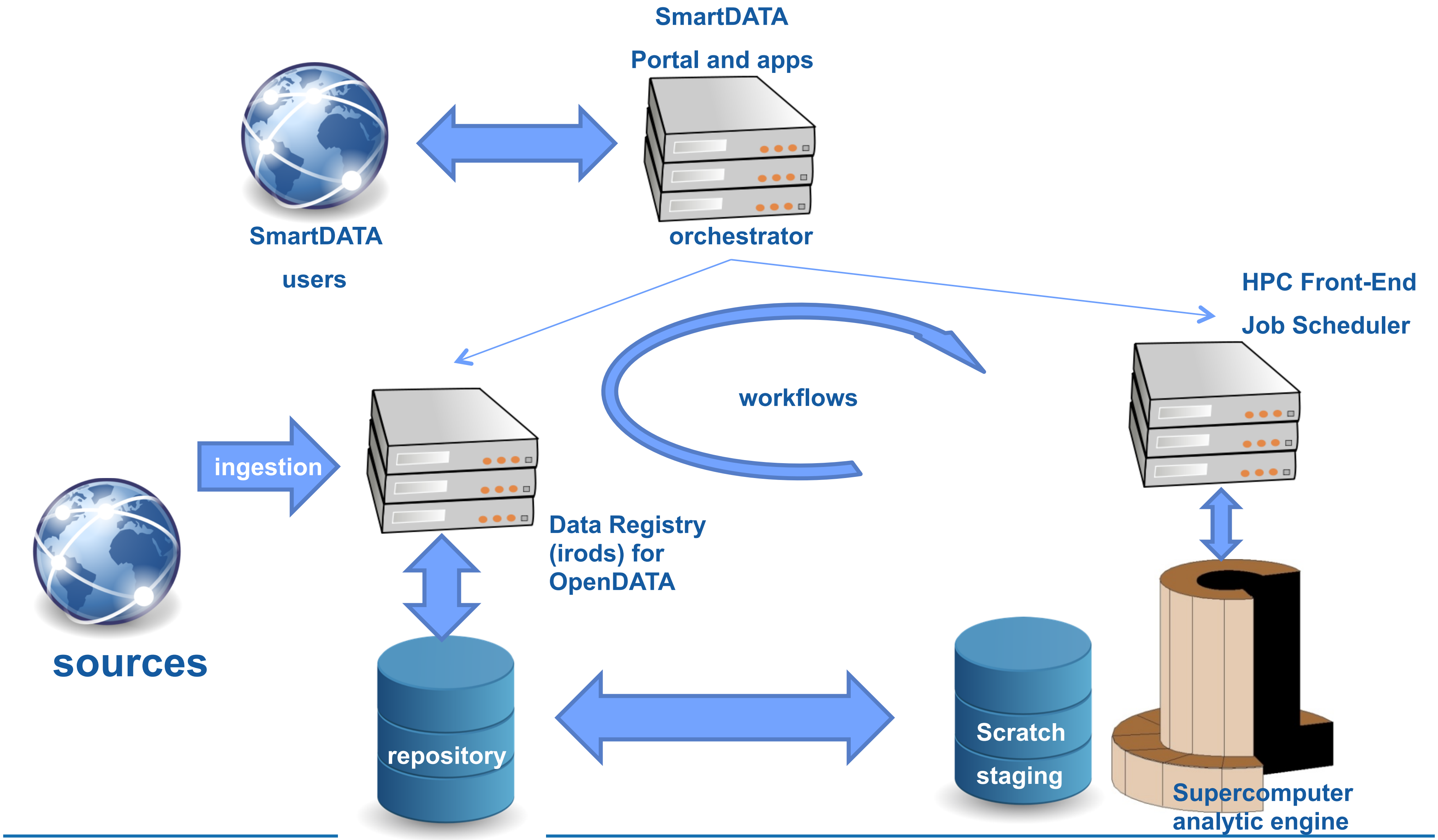
SmartDATA project

CINECA’s project to deal with “BigData” and HPC: a **Big Data Analytics engine for OpenDATA catalogues – EUDAT choerent**

- Initial pre-conditions:
  - **Data are archived, catalogued, searchable and accessible**:
    - Robust and reliable infrastructure
    - Interoperability and adoption of standards
- SmartData satisfies these requirements and besides:
  - make available an **HPC environment with analysis tools** which foster the re-use of such data.
  - ease the information exchange among experts belonging to different disciplines, from both the point of view of the methods and of the data.

→ Hence the **data sharing creates new knowledge**

SmartDATA project



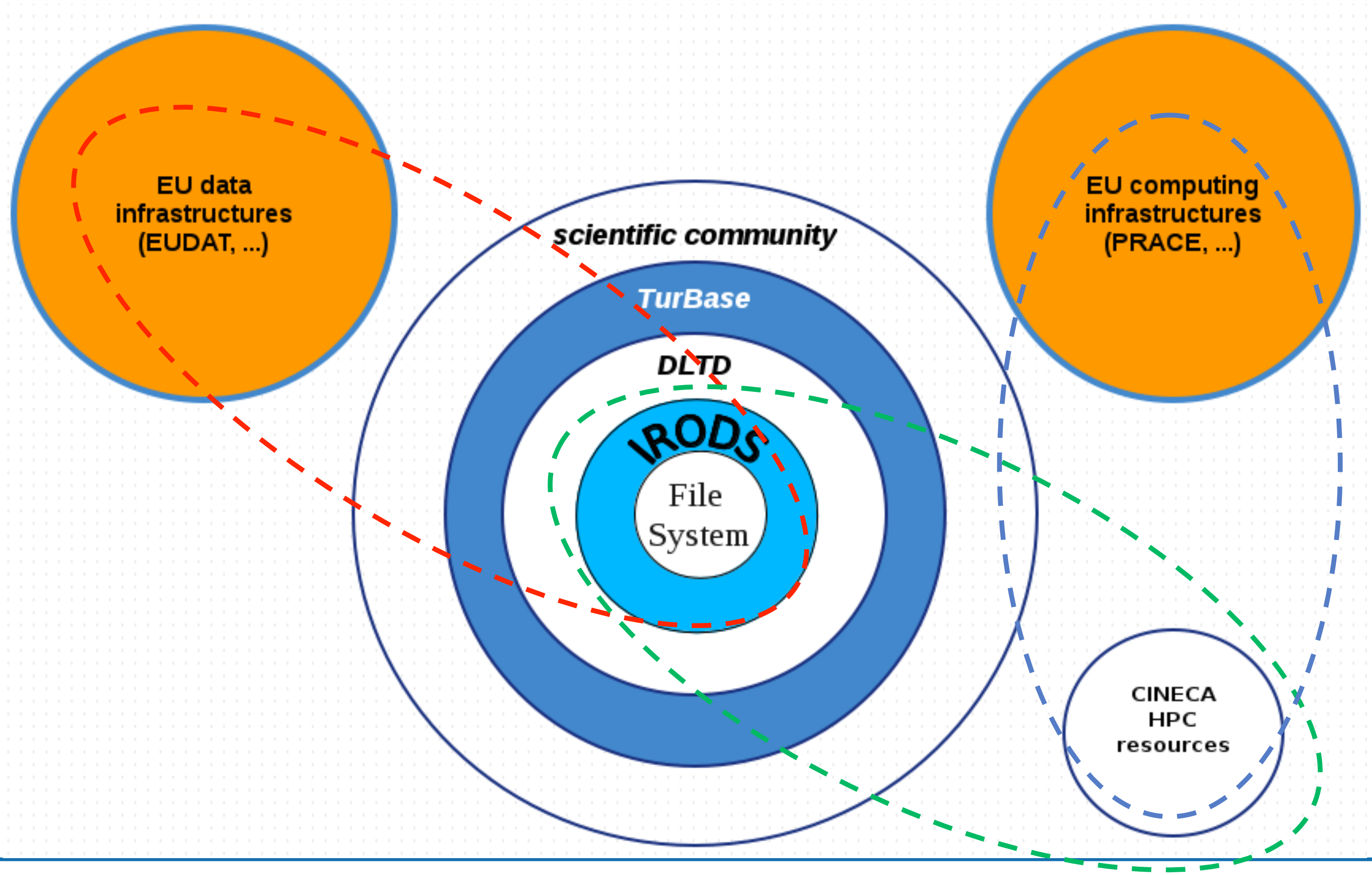
EuHIT project

- EuHIT is a consortium that aims at integrating cutting-edge European facilities for turbulence research across national boundaries (2013-2017).
- It includes
  - 25 research institutes and 2 industrial partners from 10 European countries.
  - A total of 14 cutting-edge turbulence research infrastructures

- CINECA will provide the technical and hardware infrastructure for the turbulence data digital library service (DLTD), offering:
  - A service to **manage data** coming from numerical simulations and experiments in the field of fluid dynamics.
  - A **storage space** of 150 TB online.
  - A **helpdesk** and **specialized support**.

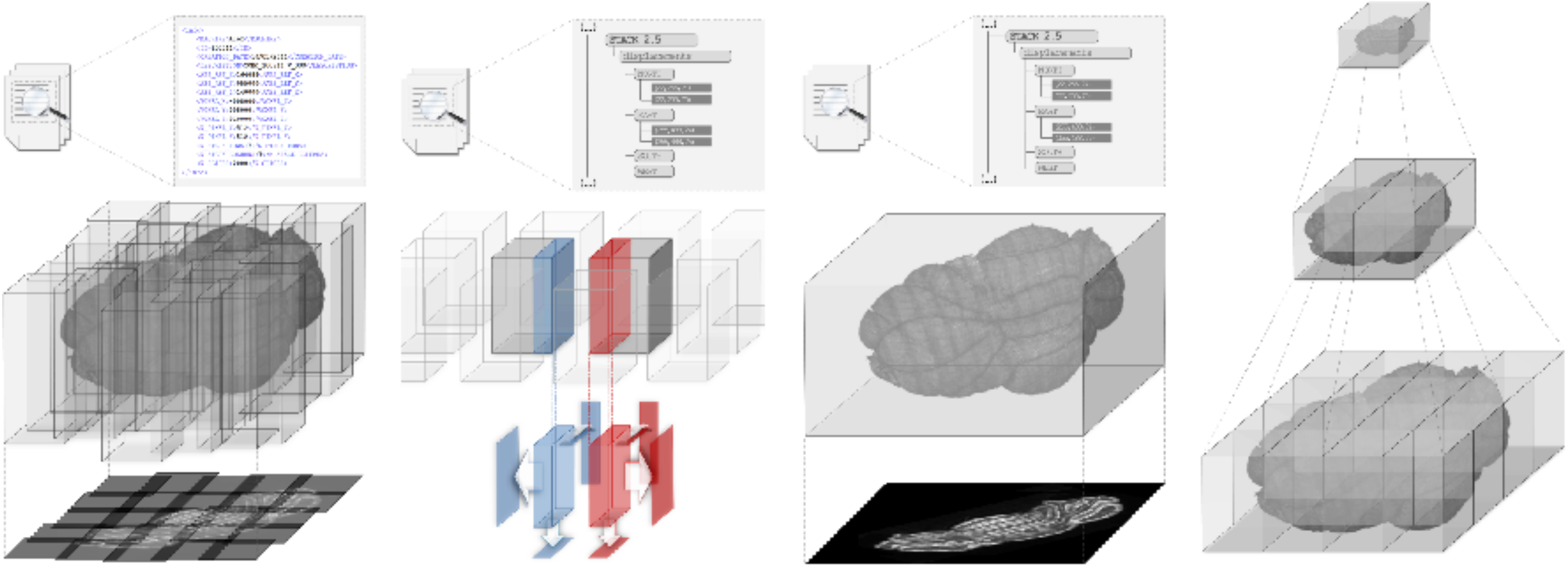
EUHIT project

TurBase, a freely accessible “living” knowledgebase for high quality turbulence data, will rely on the CINECA’s DLTD service, which will be implemented using the same technology adopted by EUDAT to foster interoperability and accessibility


EU data infrastructures (EUDAT, ...)
EU computing infrastructures (PRACE, ...)
scientific community
TurBase
DLTD
iRODS
File System
CINECA HPC resources

Human Brain project

- The Human Brain Project (HBP) is a large scale European research project which aims to simulate the world's first and most exact human brain in a supercomputer.
- Working together with 86 different European institutions the project is planned to last ten years (2013-2023).



To image an entire brain many parallel stacks of images are acquired. They are afterwards merged with a custom-made algorithm suited to work with very large data sets (~ 1 TB)

Human Brain project

- WP7.5 "*High Performance Computing Platform: integration and operations*"
- *CINECA is responsible for* Task 7.5.4 The HBP supercomputer for massive data analytics:

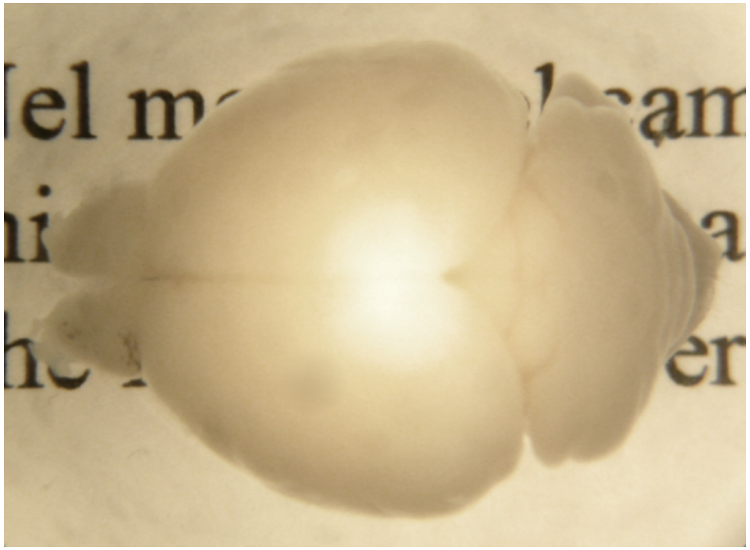
- Implement and operate a data-centric HPC facility providing efficient storage, processing and management of large volumes of data generated by the HBP.

- The *HBP Massive Data Analytics Supercomputer* in CINECA will provide Tier-0 and Tier-1 service, integrated with a mass storage facility of more than 5 Petabytes of working space with an aggregated file system bandwidth of about 70 GB/sec in a full production environment.

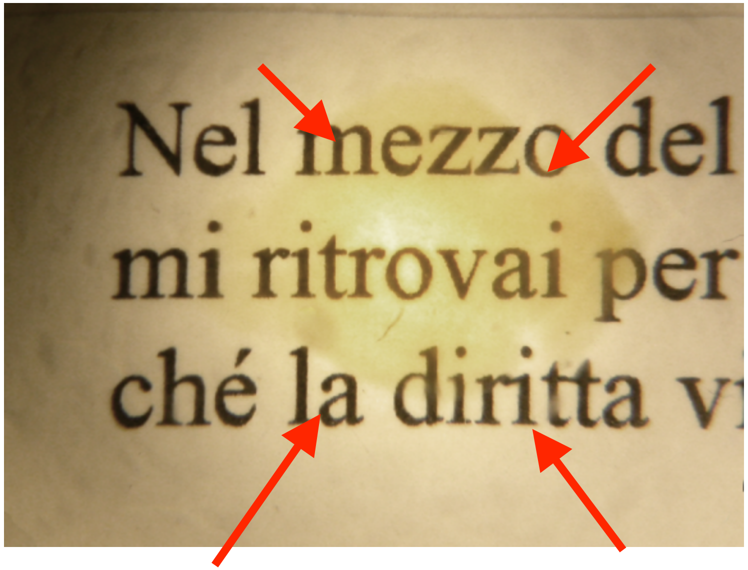
- The system will be integrated with a Data Facility of more than 5 Petabytes for on-line disk storage repository and more 10 Petabytes for long term data archiving, providing efficient data life-cycle management of structured and un-structured data generated by the HBP.

Sample preparation

1. Transcardiac mouse perfusion with PBS 1X followed by Paraformaldehyde
2. Extraction of the brain from the skull & post-fix in PFA 4% overnight @ 4 °C.
3. (optional) excision of the desired part of the brain.
4. Embedding of the sample in a 0.5 % w/w agarose gel.
5. Dehydration in graded ethanol series.
6. (for whole brains) additional dehydration in hexane 100% (1h).
7. Incubation in clearing solution (1:2 Benzyl Alcohol / Benzyl Benzoate).



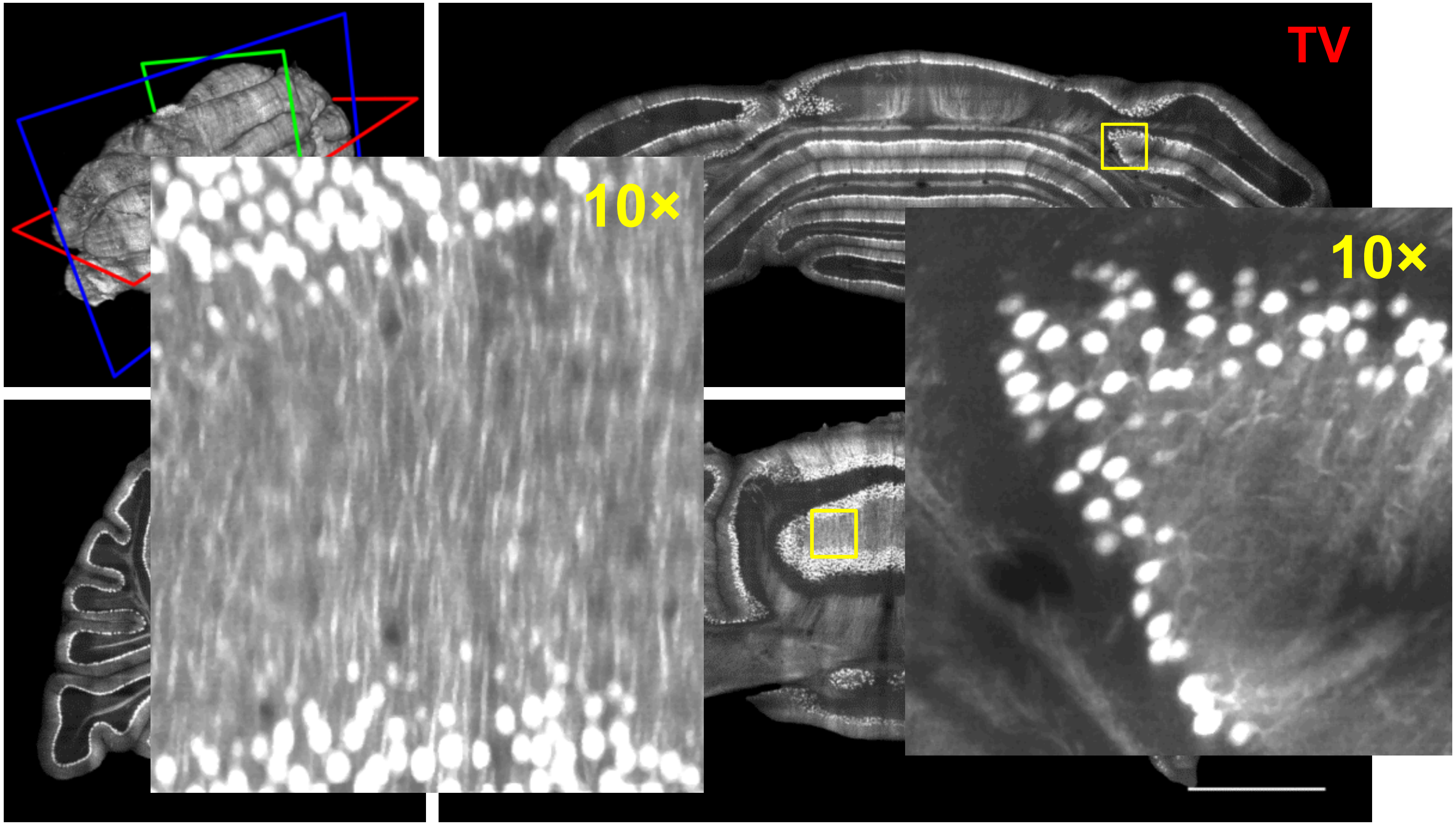
Uncleared mouse brain



Cleared mouse brain

Whole Brain Imaging

Scale bars: 1 mm

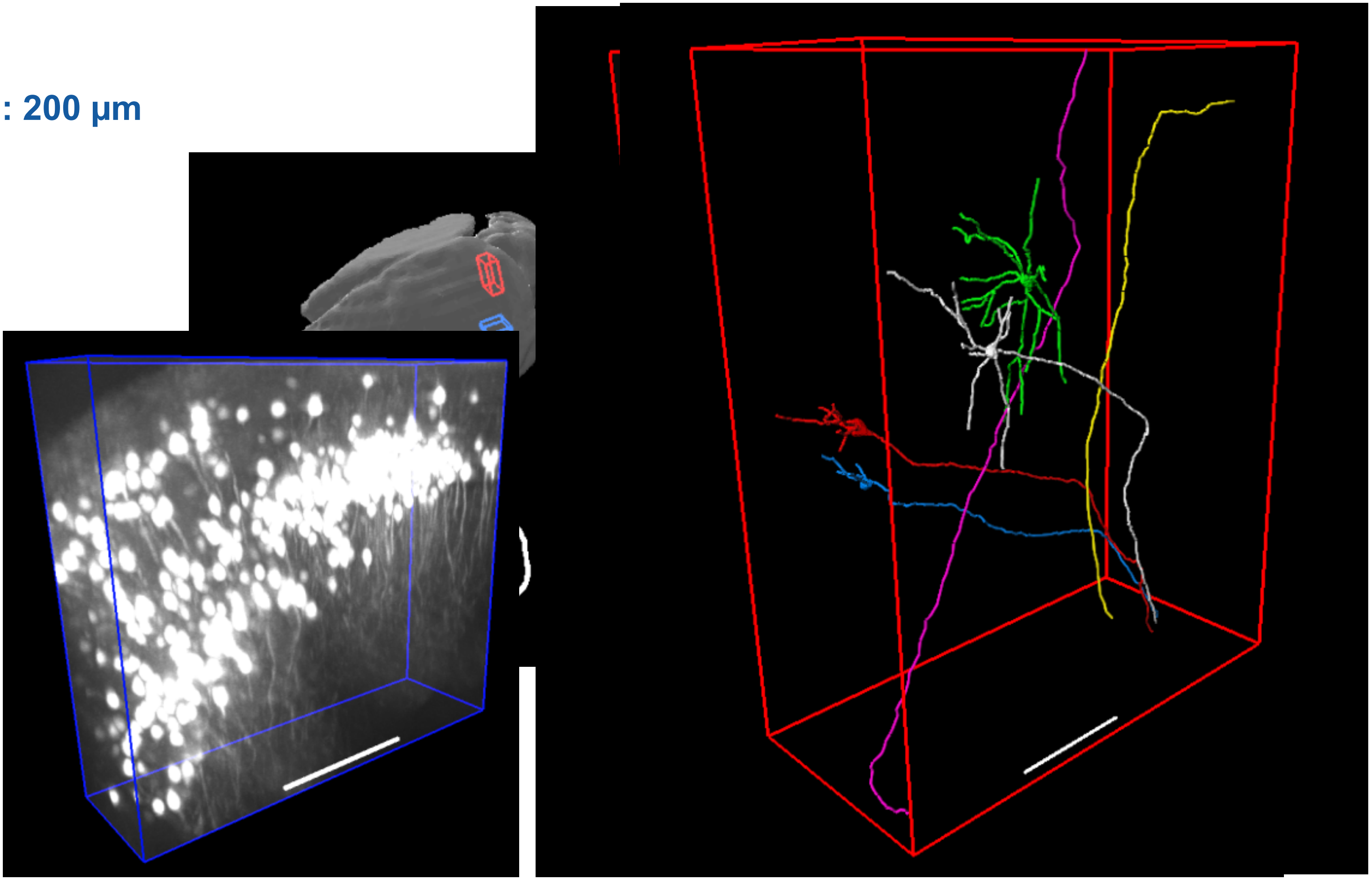


Cerebellum from a P10 L7-GFP mouse

Total volume 73 mm^3, voxel size 0.8×0.8×1 µm^3, acquisition time ≈ 24 h (1.3 MegaVoxels/s)

Whole Brain Imaging

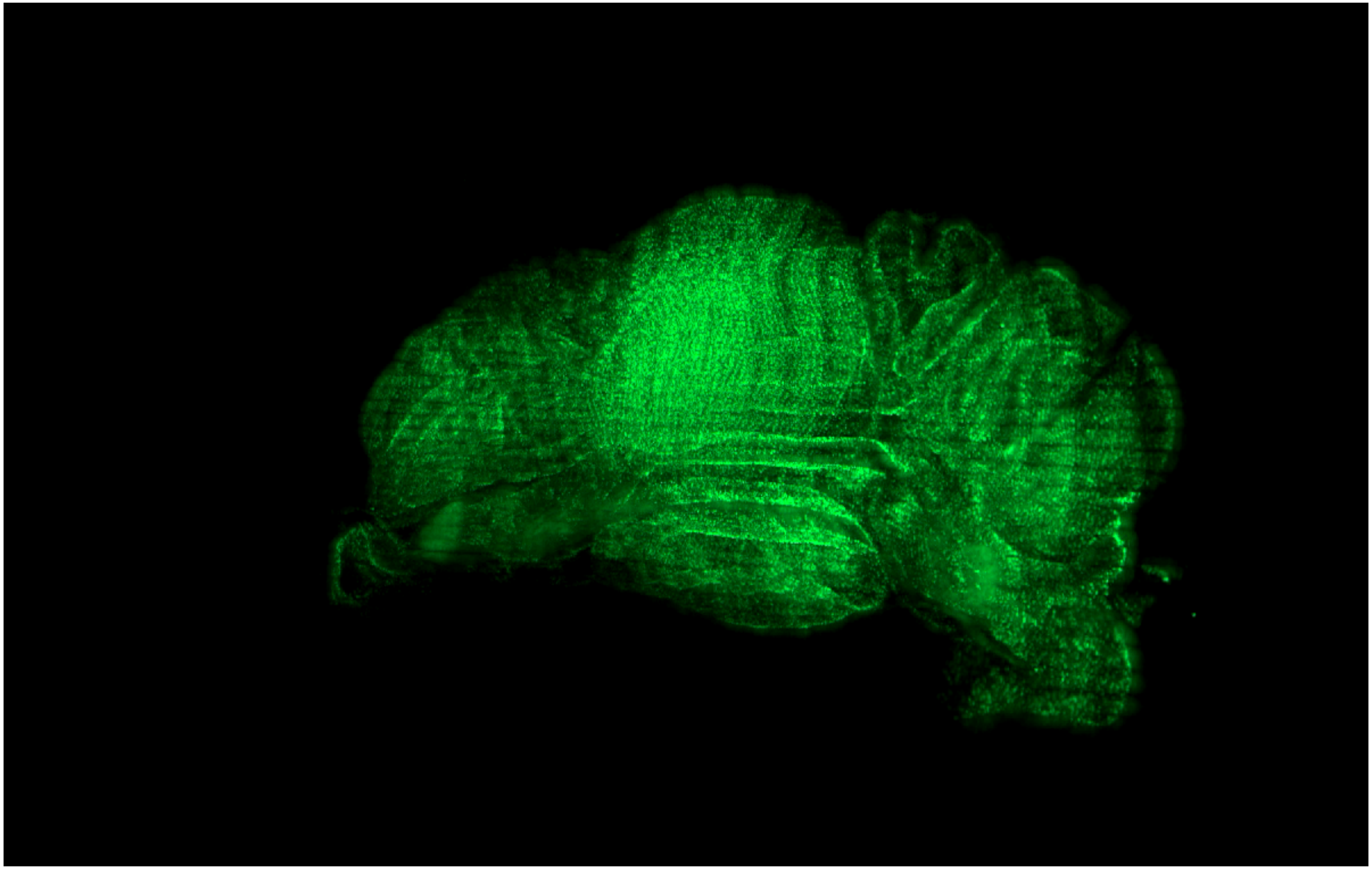
Scale bars: 200 μm



Whole brain from P15 thy1-GFP-M mouse

Portions of hippocampus and superior colliculus are magnified

Total volume 223 mm^3, voxel size 0.8×0.8×1 μm^3, acquisition time ≈ 3 days (1.3 MegaVoxels/s)



Initiatives (HPC – HPDA)

HPC

Tier-0 (FERMI) >2PF; Tier-1 > 1PF;

- Open Access
- Peer Review
- PRACE, ISCRA, LISA

HPDA

Nubes, cloud, remote visualization, structured and unstructured data processing

80 servers node Linux cluster

Repository

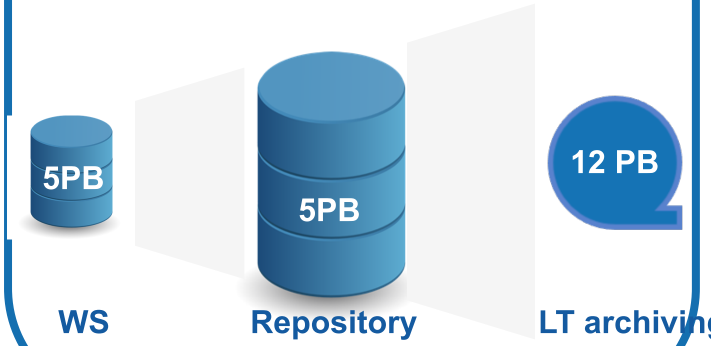
Access

Reference

Curation

Replica

Preservation



Business model for (Cineca) RI

- Research infrastructure supported by Minister, National and European (structured funds)
- Scientific community contribution for marginal and operational costs
- Standard service provision best practices are commonly in use by Cineca which ISO fully certified
- Partnership and scientific collaboration
- Vertical disciplinary project should (have to) define in its budget the contribution for the persistency of the (Cineca) RI on base of its programmatic access
- Pay for service model (applicable for Cineca being a private organization) does imply costs for VAT and couldn't be widely applicable
- Open tender for service procurement may be very difficult to apply because of the difficult to define a suitable terms of reference documentation for in progress scientific activities
- The skills and competence profiles need a continuously training programme part also on the job
- Technology transfer process also in collaboration with scientific communities for added value service towards private organizations and industries.

Many thanks for your attention!

Bologna

Milan

Rome