EUDAT:
Data sharing and management in a collaborative data infrastructure

Rob Baxter, EPCC, University of Edinburgh

EUDAT

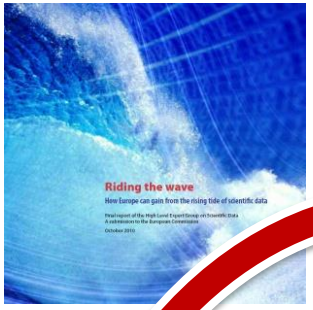
European Data – EUDAT

- Start Date: 1st October 2011
- Duration: 39 months
- Budget: 16.3 M€ (9.3 M€ EC)
- Call: INFRA-2011-1.2.2
- Consortium: 25+ partners from 13 countries
  - "Data scientists", data centres, technology providers
- Goals:
  - create a cost-effective, high-quality Collaborative Data Infrastructure (CDI)
  - …that meets users' needs in a flexible and sustainable way
  - …across geographical and discipline boundaries

EUDAT

EUDAT consortium:
data centres and data scientists

EUDAT



EUDAT vision – the CDI

Trust

Data Curation

Data Generators

Users

Community Support Services

Data Services

User-focused functionality, data capture & transfer, VREs

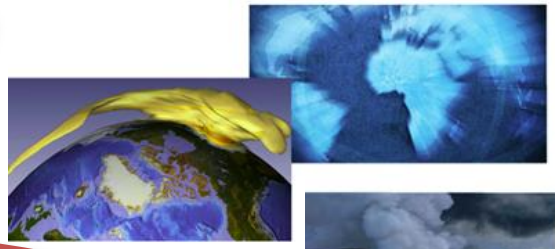
Data discovery & navigation, workflow creation, annotation, interpretability

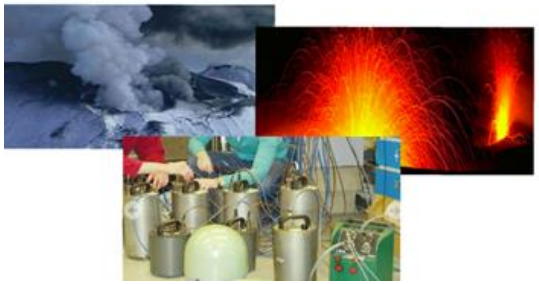
Persistent storage, identification, authenticity, workflow execution, mining

EUDAT

Five core research communities*

- **CLARIN**: Common Language Resources and Technology Infrastructure
- **LifeWatch**: Biodiversity Data and Observatories
- **EPOS**: European Plate Observatory System
- **ENES**: Service for Climate Modelling in
- **VPH**: The Virtual Physiological Human
- All share common challenges:
  - Reference models and architectures
  - Persistent data identifiers
  - Metadata management
  - Distributed data sources
  - Data interoperability
- EUDAT has to work bottom up to "refactor" cross-community common services

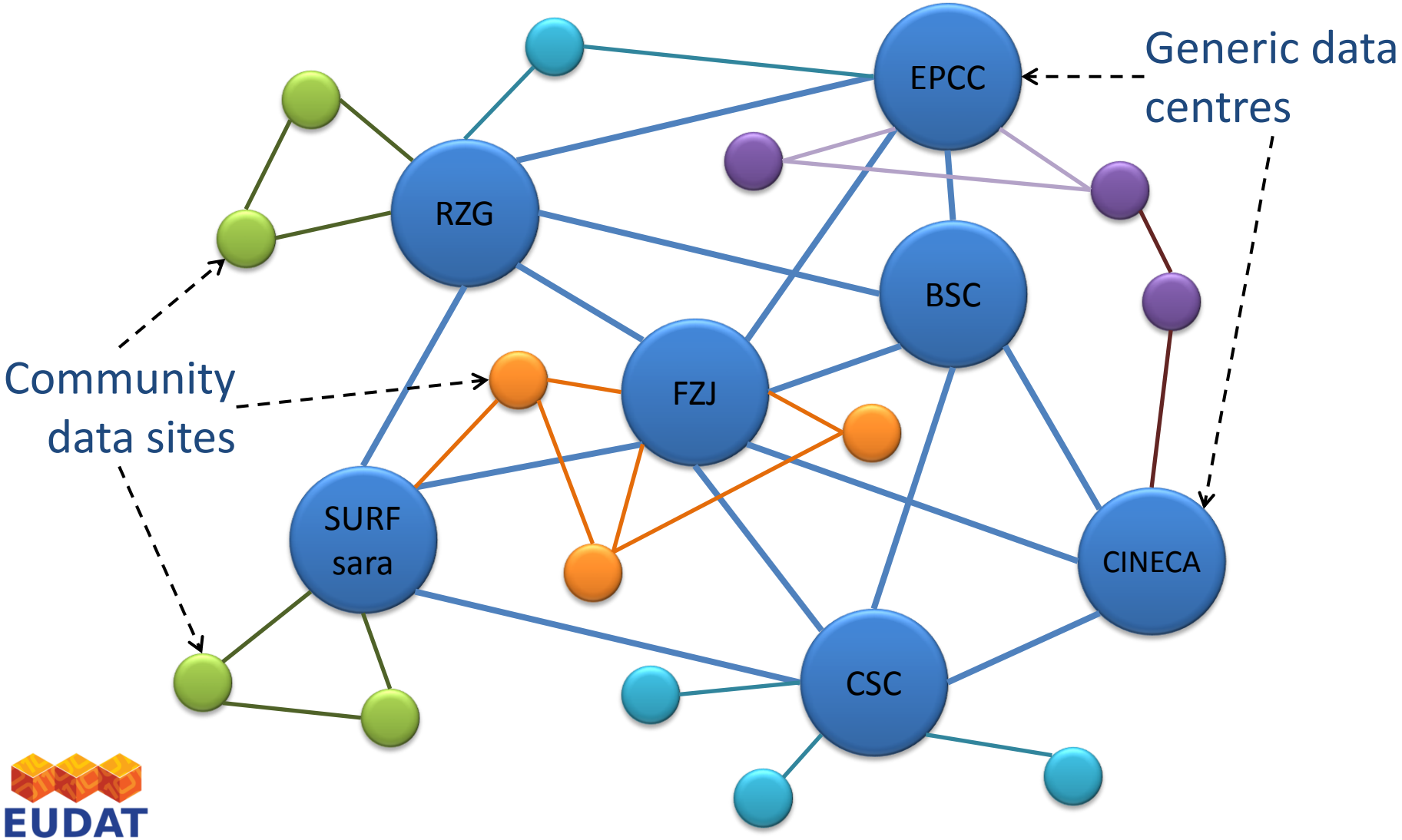




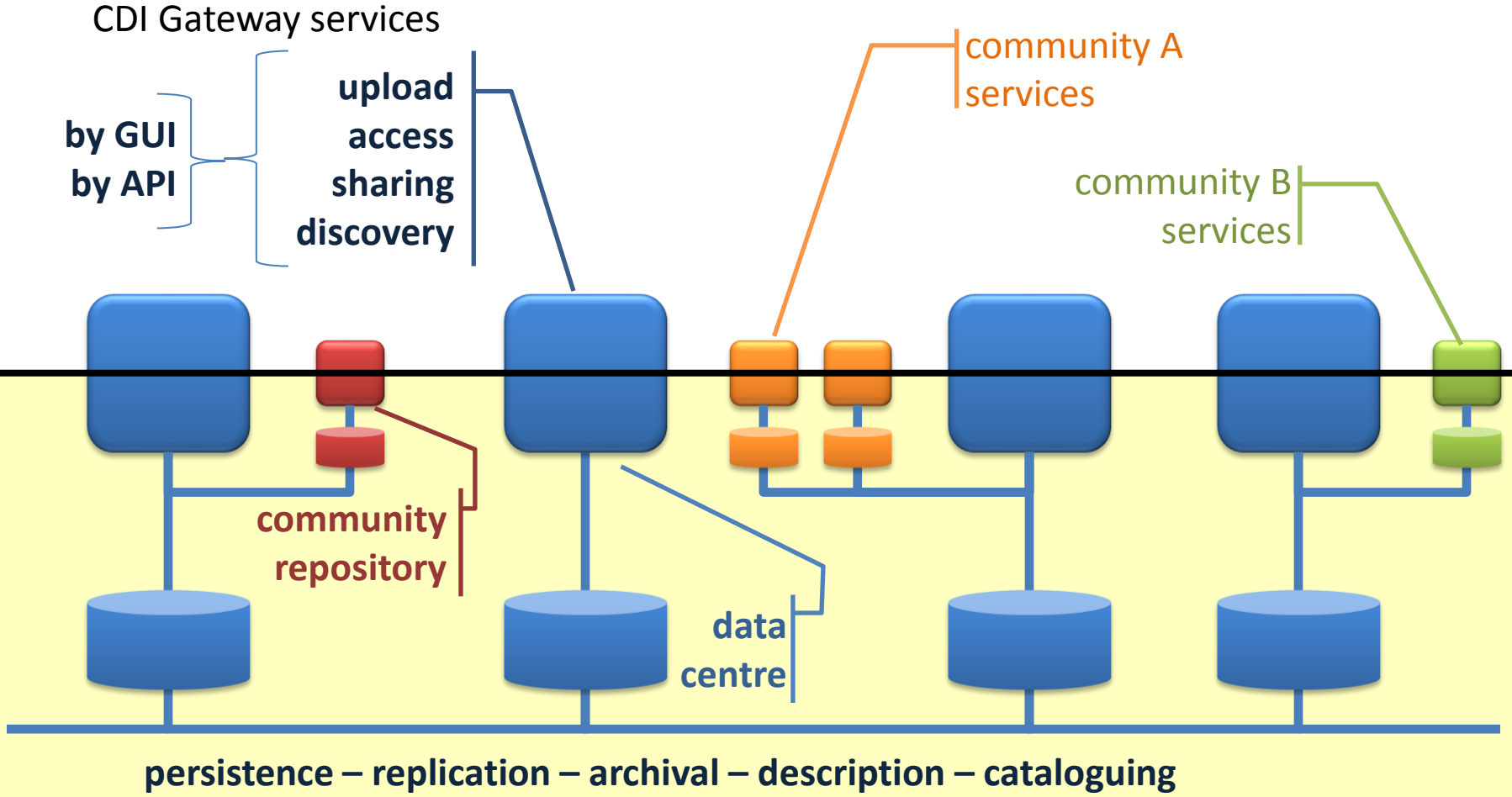
EUDAT

* and even more associate communities!

EUDAT CDI network

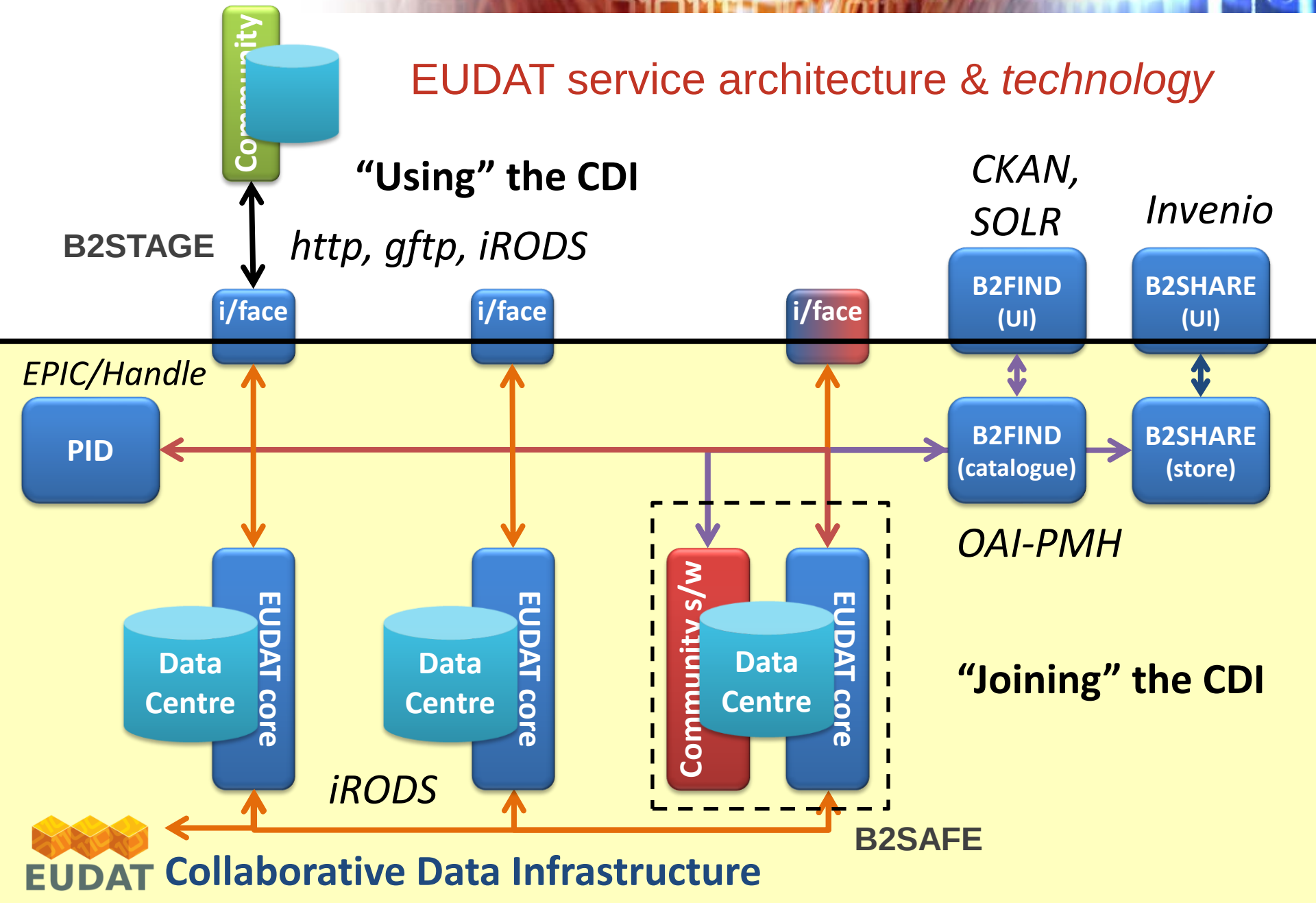


EUDAT



EUDAT CDI capabilities

EUDAT service architecture & *technology*



EUDAT Collaborative Data Infrastructure

Heterogeneity and Homogeneity

- The generic data centres* are quite homogeneous
  - (* actually, we're more generic HPC centres)
  - Big disks; big filesystems (Lustre, GPFS); TSM, DMF
- The community sites… aren't
  - Anything from a small research group to DKRZ
- Challenge is to build (distributed) foundations on the DCs that are easily usable by the CSs

- Reclaim the Web!

EUDAT

EUDAT Guiding Policy Principles

1. Data deposited with the EUDAT CDI will be preserved long-term
2. Data are best curated in their own communities.
3. Access to data in the EUDAT CDI is free at the point of use
4. For an EUDAT community repository to be designated a Trustworthy Digital Repository (TDR), it follows that EUDAT services and infrastructure must be a suitable target for “TDR outsourcing”
5. EUDAT will not assert ownership of any data it holds

EUDAT

Open Access Principles

- Two further principles on open access:
  - all data in the CDI should, in time, become full open access. Open access is the norm for CDI data;
  - embargo periods for original producers are fully supported, on condition that such data become openly accessible when the embargo period expires.
- These imply:
  - policy harmonisation
  - a common licensing scheme

EUDAT

EUDAT Licensing

- Comparatively easy at one level…
  - We will (almost certainly) recommend CC 4.0 BY/SA for open data
- Rather difficult at another…
  - Persuading all members of the network to sign up!
  - Maybe OK if site owns data copyright
  - But some sites hold third-party copyrighted data
- Will need to be pragmatic
  - Follow DANS's maxim: "*Open if possible, restricted if necessary*"

EUDAT

Policy Harmonisation

- Aim for 2014: a roadmap across EUDAT
  - Ideally, comprehensive study with yes/no answers
  - Create a “heatmap” of the policy landscape across sites
  - Identify areas of harmony, areas needing further work
- Adopt a taxonomy/set of headings
  - From APARSEN’s *Exemplar Good Governance Structures and Data Policies*
  - From Data Seal of Approval
  - From Open Access principles

Top Three Challenges: #1

Policy automation

- Policies → requirements and constraints
  - "ensure at least 3 copies of this object are extant"
  - "ensure no copy of this object leaves the UK"
- Need automatic means to propagate rules deep into the infrastructure
  - Will need tagging, annotation of data objects
- *In progress*

EUDAT

Top Three Challenges: #2

Distributed authorisation

- Propagating authz is a special case of policy
- Needed especially for replicas managed in different administrative domains
- Need fine-grain control for sensitive data

- *Currently no common solution*

EUDAT

Top Three Challenges: #3

Designing for change

- Needs to be straightforward to join or leave the CDI
- Need to connect at common level (with sufficient richness to make it worthwhile) while not disrupting site-by-site operations
- Need connection-oriented, protocol-oriented approach
- *In place; arguably needs revised for “CDI 2.0”*

EUDAT

Conclusions

- Heterogeneity is EUDAT's biggest challenge
  - And its reason for being
- Adopting open access principles may help
  - Possibilities in streamlining policy, licensing
- Open questions right now:
  - Propagation of authorisation requirements
  - Identification of a useful core metadata set
  - Streamlining the connections that new nodes need
  - Legal entity or not?

EUDAT