ELIXIR

Rafael C Jimenez

ELIXIR CTO

RDA Fourth Plenary Meeting, ELIXIR Bridging Force IG

2014, Tuesday 23 September

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

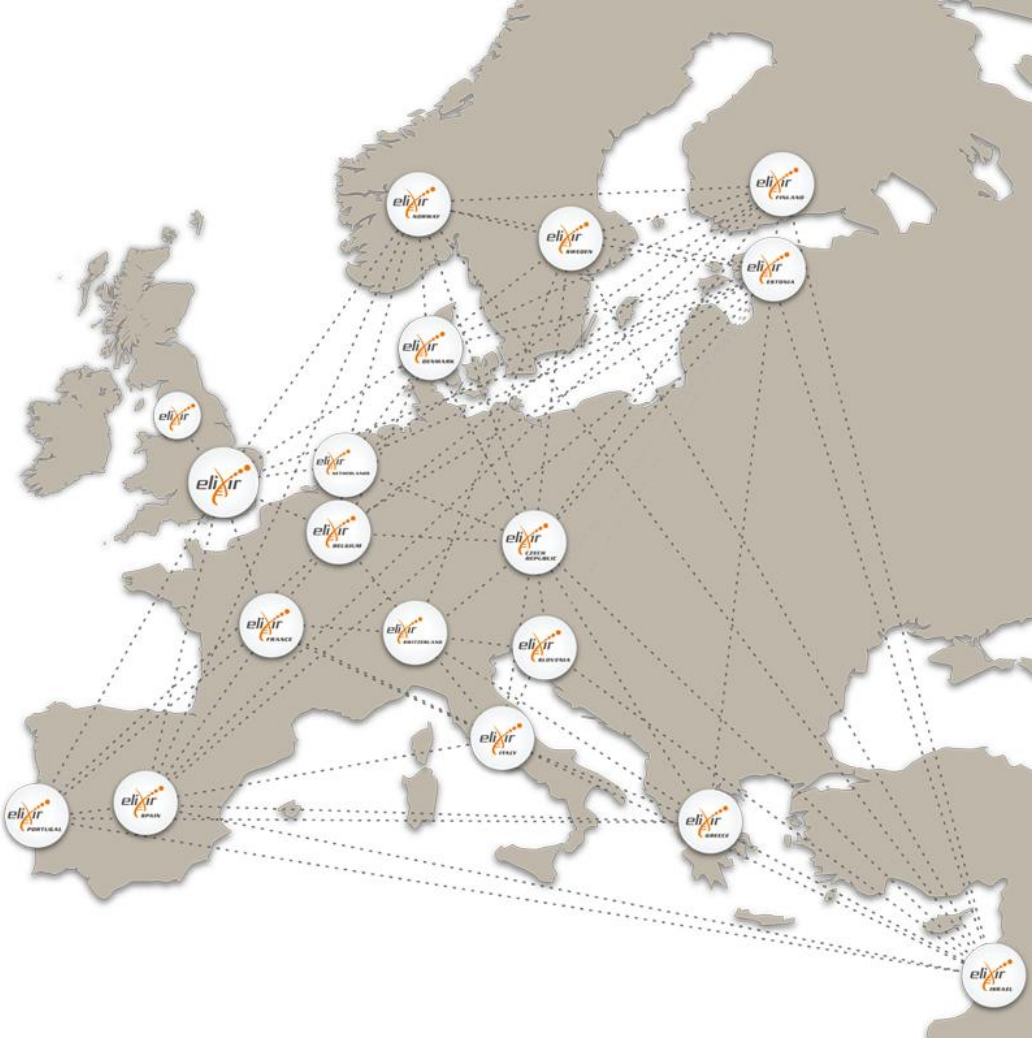
- What are the **partners**, who are the **service providers**, the **users**?

Challenges

- **Sustain** data and services
- Make data and service **interoperable**
  - Necessary to integrate data
  - Specially medical, clinical and research
- Data too big to **store**, **exchange** & **compute**? Forthcoming challenges …
  - Data production grows faster than storage
  - Cost of data production technologies declines faster than storage
  - It takes longer to transfer data than produce the data.
  - Privacy, security & access (AAI)
  - Training

ELIXIR

- European **life sciences** research infrastructure for **biological information** to **support all life science research**
- **Safeguard data** and build **sustainable data services**



ELIXIR Members

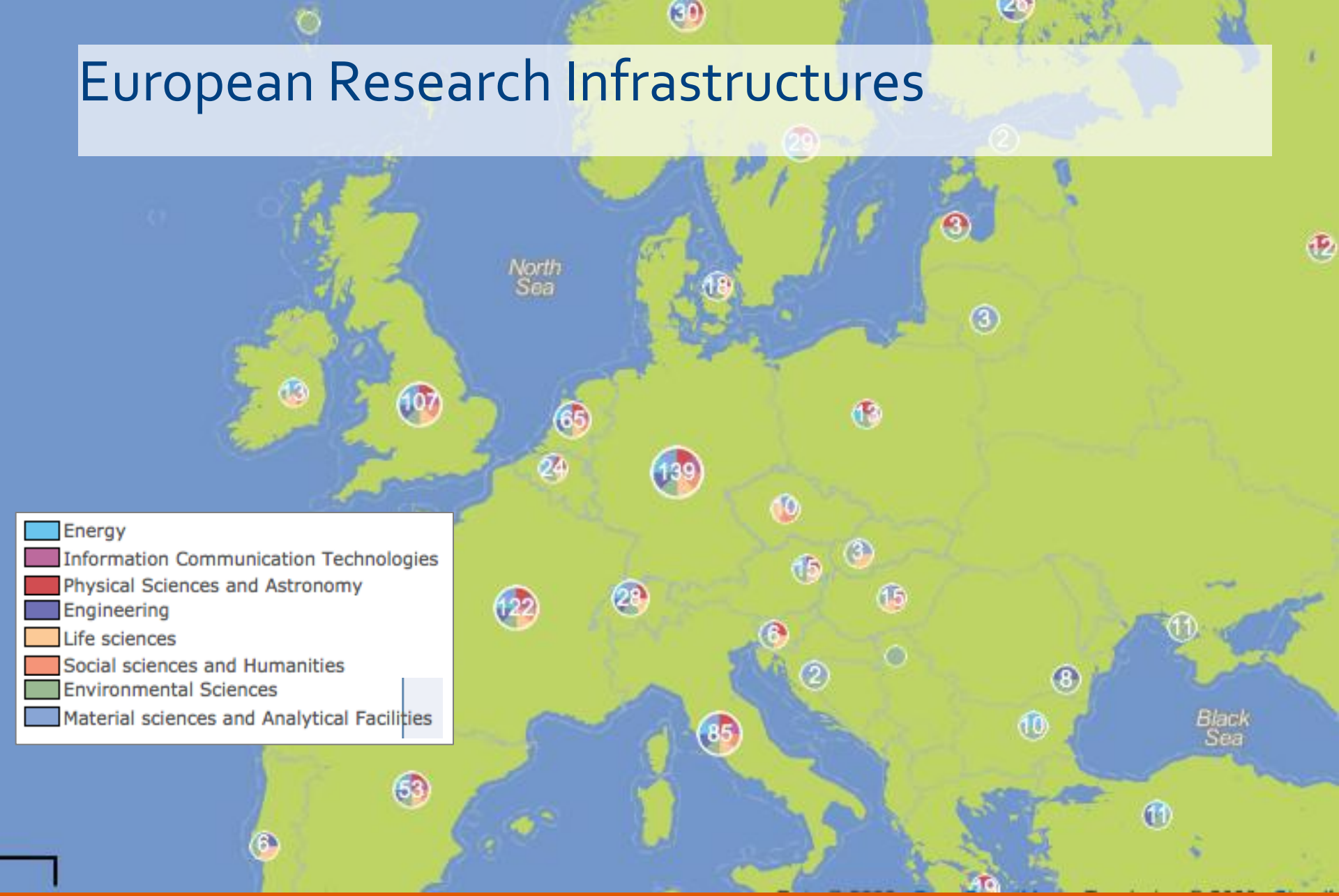
- Participated by major bioinformatics service providers (~130) and supported by **17 EU member states** and **EMBL**

- 11 EU states signed agreement
  - *Czech Republic, Estonia, Denmark, Finland,Israel, Netherlands, Norway,Portugal, Switzerland, Sweden, UK*
- 6 EU signed MoU
  - *Belgium, Greece, France, Italy, Slovenia, Spain*

ELIXIR node proposals

- **ELIXIR** deliver services through national **ELIXIR Nodes** building on national strengths and priorities

European Research Infrastructures

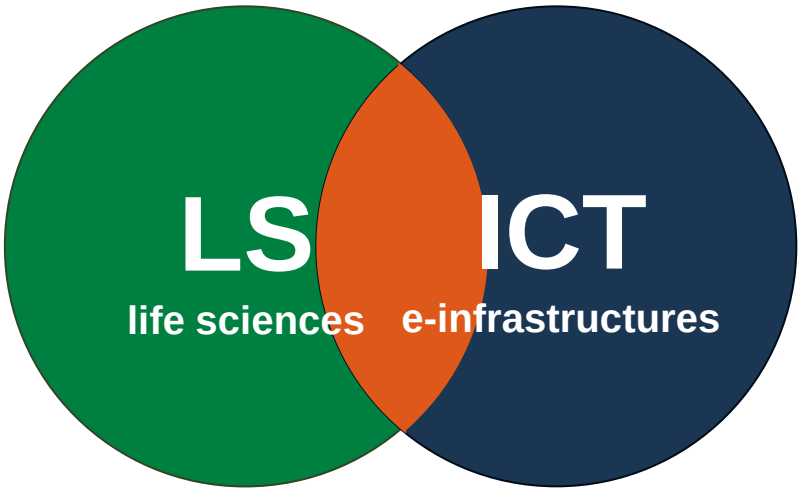


European Research Infrastructures

Research infrastructures

Facilitate research

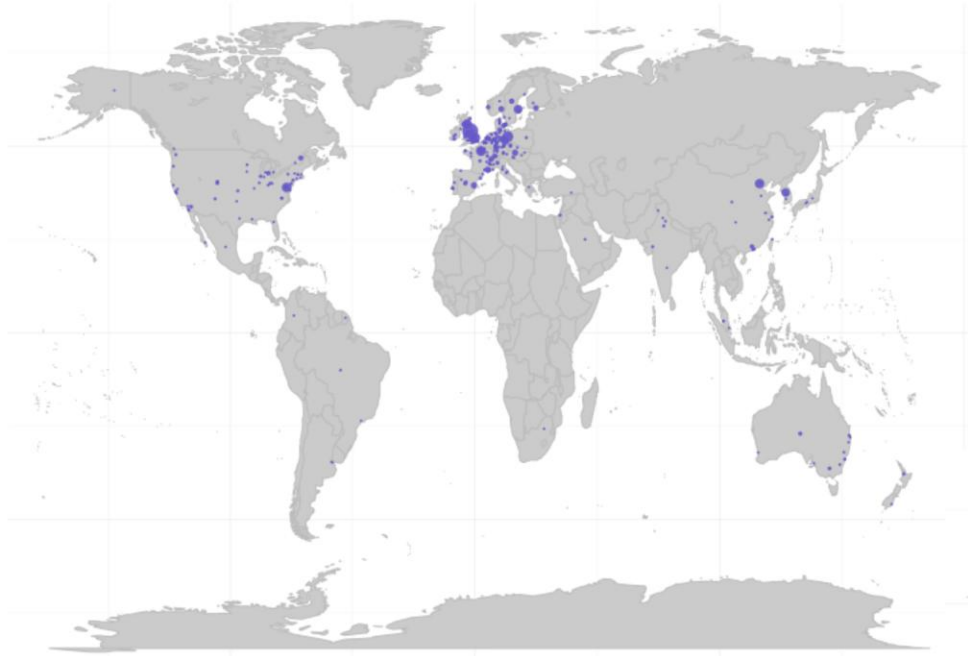
Physical facilities
Scientific information



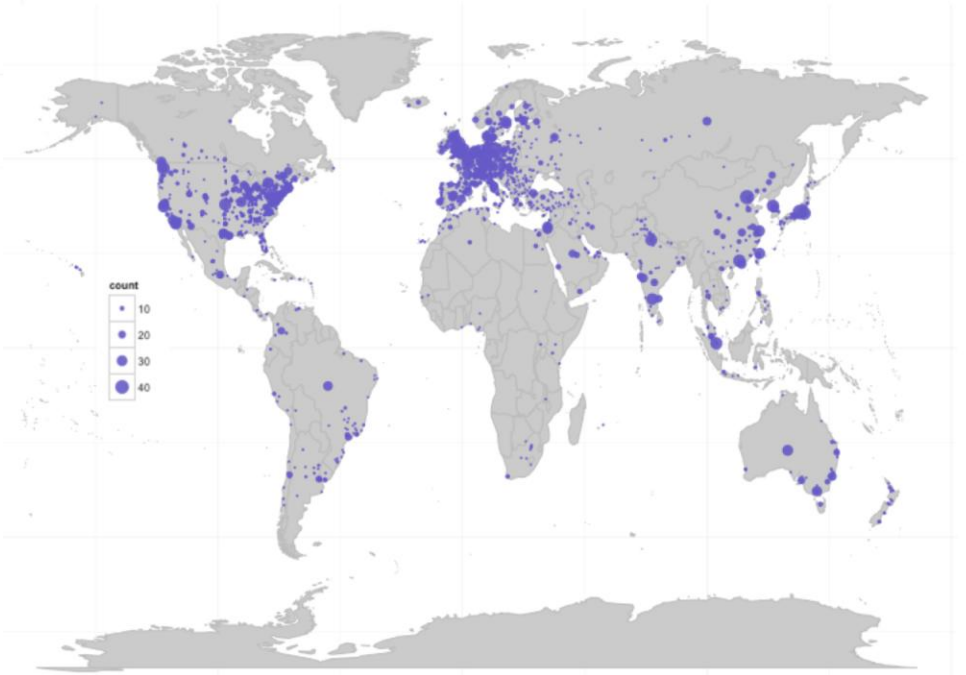
Transfer
Computation
Storage

Data

Dispersed science



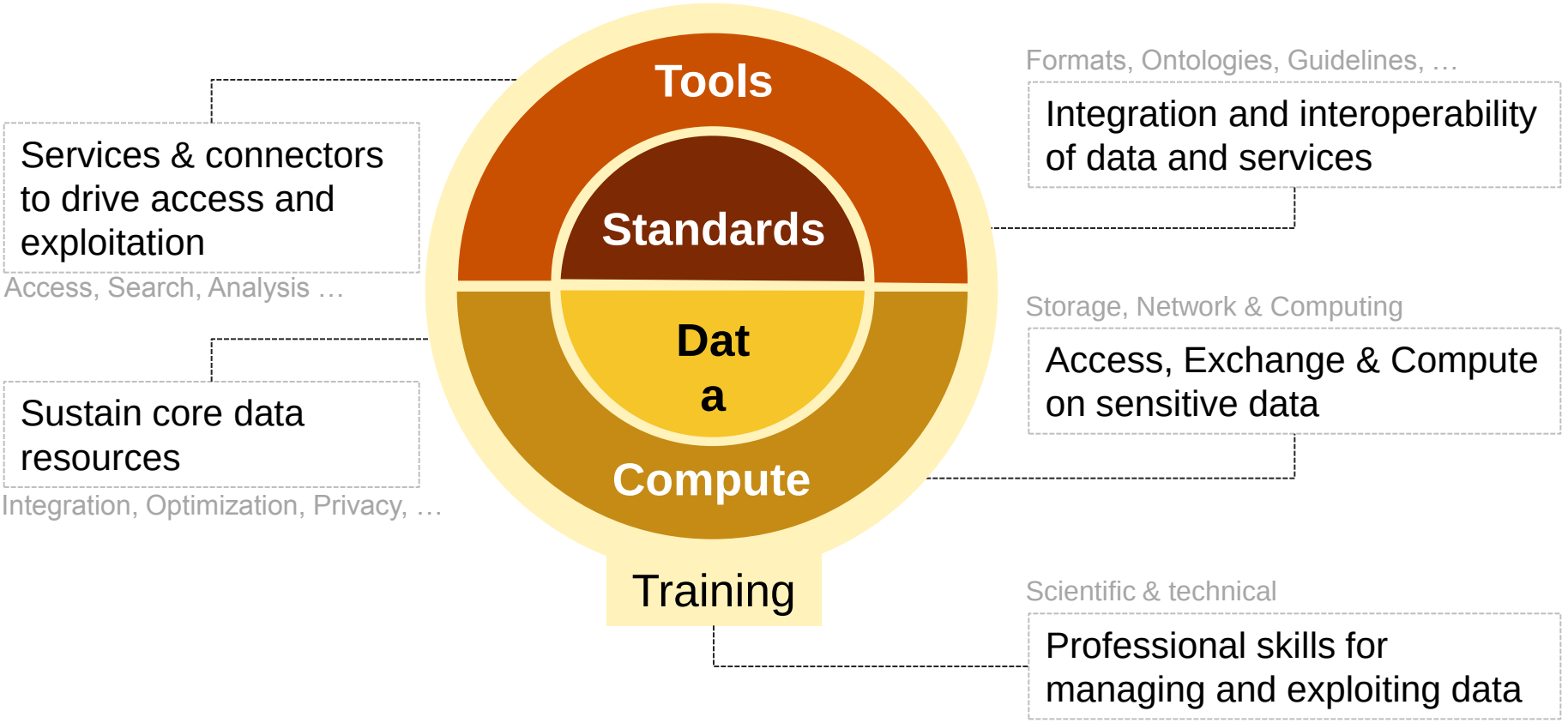
Data production



Data consumption

- What are the **resources/services** provided?

Infrastructure for Life Sciences



- What are the **agreements related to access to the RI** or the services it provides (e.g. machine time, software, data, support, etc.)?

- What are the current or planned legal structures and **funding models** in place, etc?

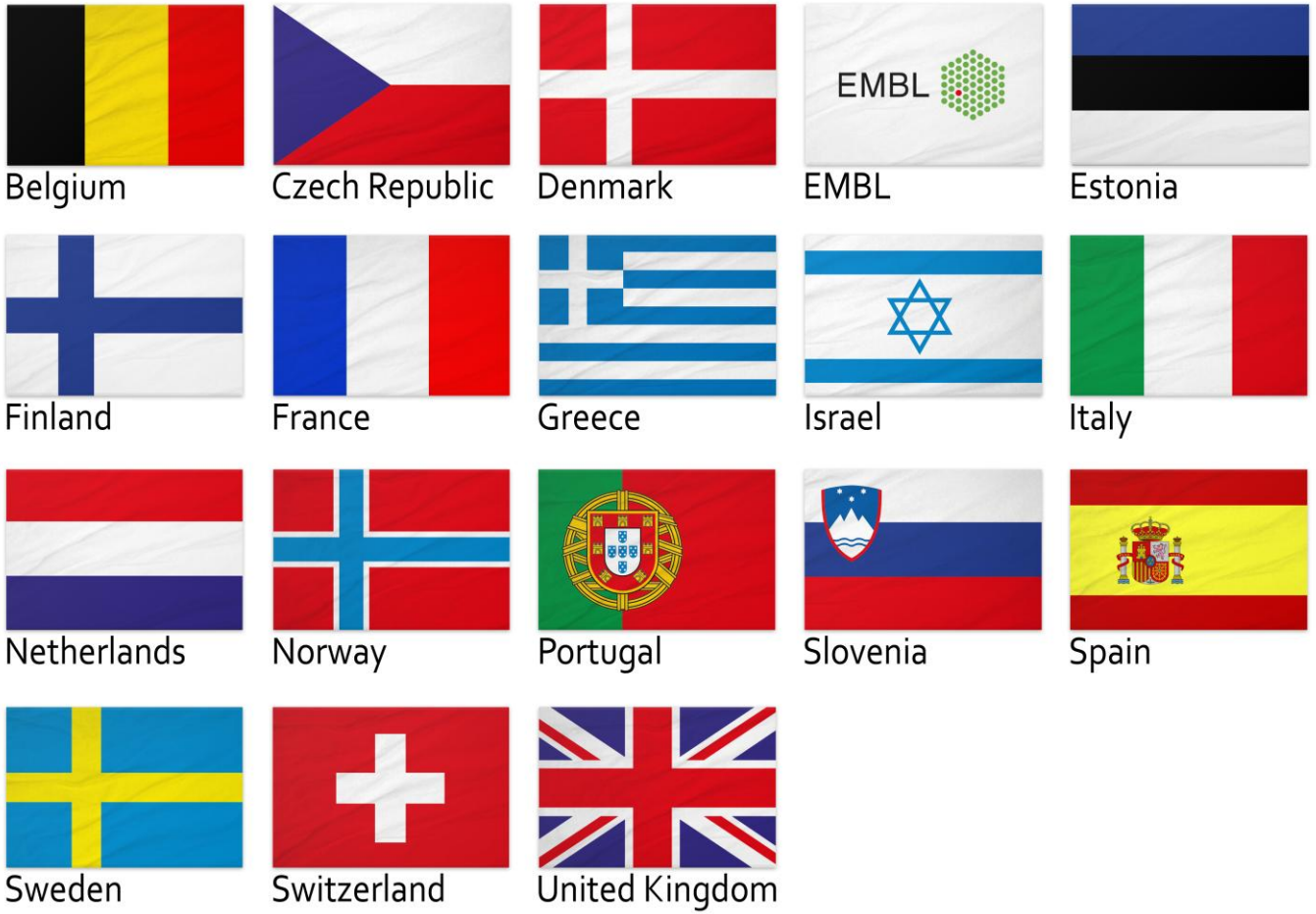
- According to the EUDAT CDI concept, what would be **the requirements from your RI to formally join** and integrate components of this RI as part of the CDI?

- EUDAT aims to provide storage resources and other related services to the widest numbers of researchers. These resources have a **cost** and their access should be regulated. In your opinion, what would be the **best model for accessing these resources**:
  - **Quality-based:** researchers apply for resources which are allocated on the basis of scientific excellence, originality, quality, and feasibility of the applications
  - **Quota-based:** access is based on quotas determined by e.g. the financial contributions from the CDI partners, or the research programmes agreed with pre-defined users)
  - **Market-based:** access is granted to anyone against a fee

- Suppose you are looking for a place to store and take care part of your scientific data for at least 10 years. What **conditions/requirements should EUDAT meet to be seen as the best place to put your data**?
  - Affordable, trustworthy, robust, persistent, and easy to use. Easy to replicate and compute. Integration with RI and e-infrastructures.

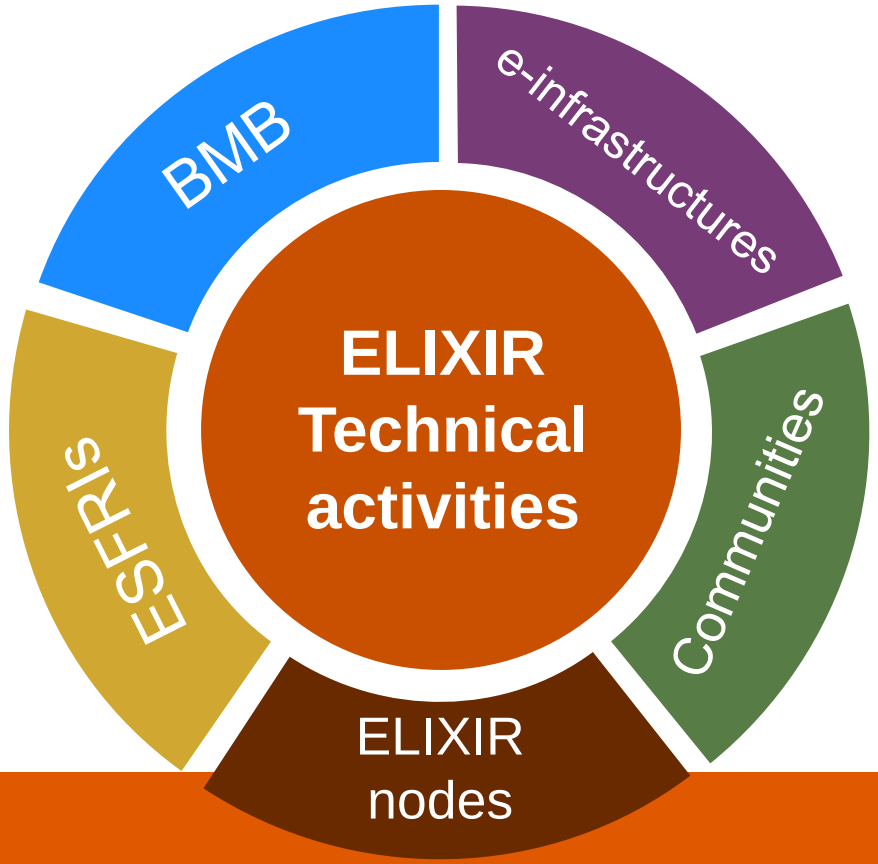
- EUDAT is currently a network of independent centers working within a common framework to develop and propose services. At present, contractual agreements (e.g. SLAs) can only be backed by centers as individual legal entities. In your opinion, **should EUDAT move towards a single legal entity**?

Thank you


Belgium
Czech Republic
Denmark
EMBL
Estonia
Finland
France
Greece
Israel
Italy
Netherlands
Norway
Portugal
Slovenia
Spain
Sweden
Switzerland
United Kingdom

ELIXIR technical activities

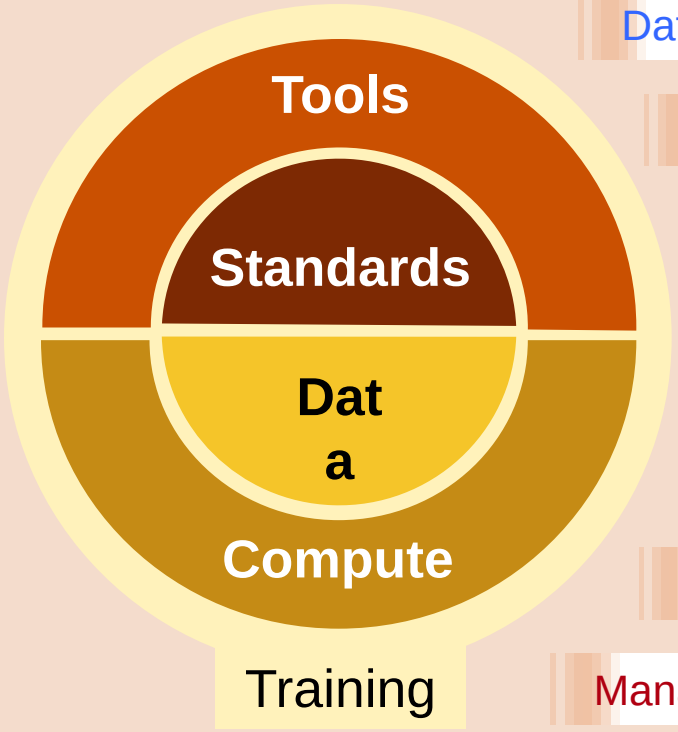
- ELIXIR Node activities, Task forces, Pilots
- Technical activities among different interest groups …



ELIXIR strategic drivers

1. Establish a distributed infrastructure to scale with the challenge of **data growth**
2. Secure and deliver the **core data resources** underpinning life science research
3. Provide **discoverable** tools, services and connectors to drive data access and exploitation
4. Provide robust **technical platforms** and clouds for secure data access, data exchange and compute
5. Develop and maintain **standards** for data management, reuse and integration
6. Drive **partnerships** with user communities and other organisations to ensure high impact
7. Close the computational biology skills gap through a comprehensive **training** programme for professionals
8. Support innovation in big data biology

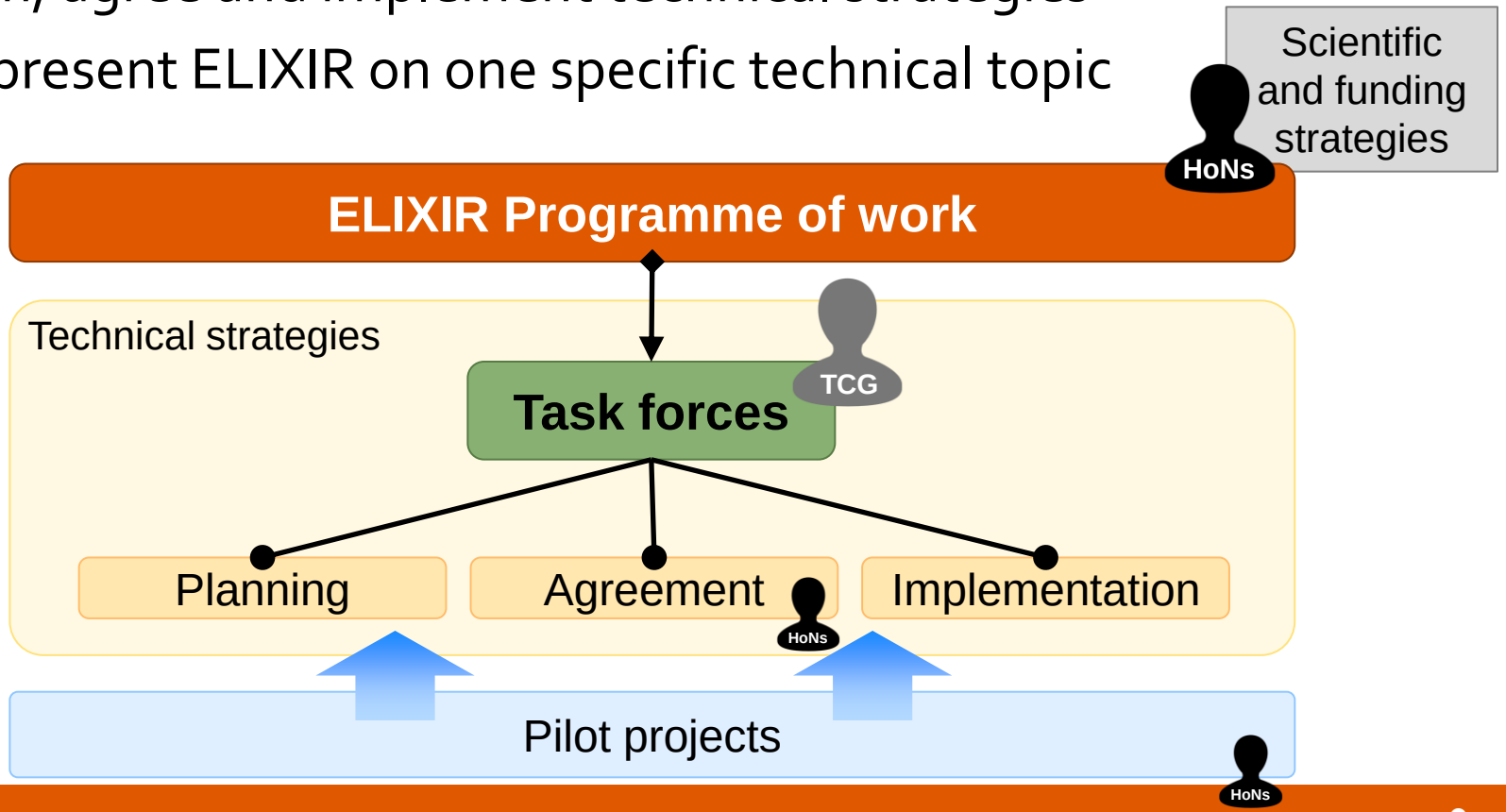
Areas of interest and programme of work



- Data interoperability, vocabulary and ontology services
- Tools interoperability and Service Registry
- Data resources & services
- Technical Services
- Domain specific services
- Training
- Management and operations

Task forces

- Working groups to coordinate the ELIXIR technical strategy
- Driven by national technical leads
- Plan, agree and implement technical strategies
- Represent ELIXIR on one specific technical topic



Task forces

Data interoperability, vocabulary and ontology services

- Interoperability

Tools interoperability and Service Registry

- Service registry

Data resources & services

- Metrics, monitoring & quality control

Technical Services

- Cloud
- Storage
- AAI

Training

- Training portal
- E-learning

Management and operations

- Communications
- Website

ELIXIR Pilot Projects

1. ELIXIR Facing **Cloud Support and Virtual Machines** - with SIB
2. ELIXIR Data IO to pilot the **continuous transfer of major archive resources** to a remote European location - with CSC, Finland
3. Establishing EGA **Distributed authentication** - with CSC, Finland
4. Establishing **EGA** as joint venture – with CRG, Spain
5. **Improving links** between Human Proteome Atlas (HPA) and EMBL-EBI resources
6. BILS-ProteomeXchange integration using EUDAT resources
7. Interoperable controlled-access big data transfer technology for ELIXIR - application to EGA EBI / CRG ELIXIR collaboration and beyond
8. Harmonising Marine Metagenomics pipelines

Affinity with RDA groups

Life science

- The BioSharing Registry: connecting data policies, standards & databases in life sciences
- Wheat Data Interoperability WG
- Agriculture Data Interest Group (IGAD)
- Marine Data Harmonization IG
- Metabolomics
- Structural Biology IG
- Toxicogenomics Interoperability IG

Affinity with RDA groups

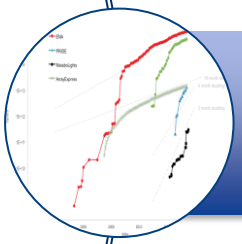
- Data Description Registry Interoperability
- Big Data Analytics IG
- Domain Repositories Interest Group
- Education and Training on handling of research data
- Federated Identity Management
- Metadata IG
- Preservation e-Infrastructure IG
- Service Management IG
- Active Data Management Plans
- Sustainability of eResearch / Cyberinfrastructure

Affinity with RDA groups

- Data Citation WG
- Data Foundation and Terminology WG
- Metadata Standards Directory Working Group
- PID Information Types WG
- RDA/WDS Publishing Data Bibliometrics WG
- RDA/WDS Publishing Data Services WG
- RDA/WDS Publishing Data Workflows WG
- Repository Audit and Certification DSA–WDS Partnership WG
- Long tail of research data IG
- PID Interest Group
- RDA/WDS Publishing Data IG
- Research Data Provenance

ELIXIR 2015 Objectives

Build the ELIXIR community.

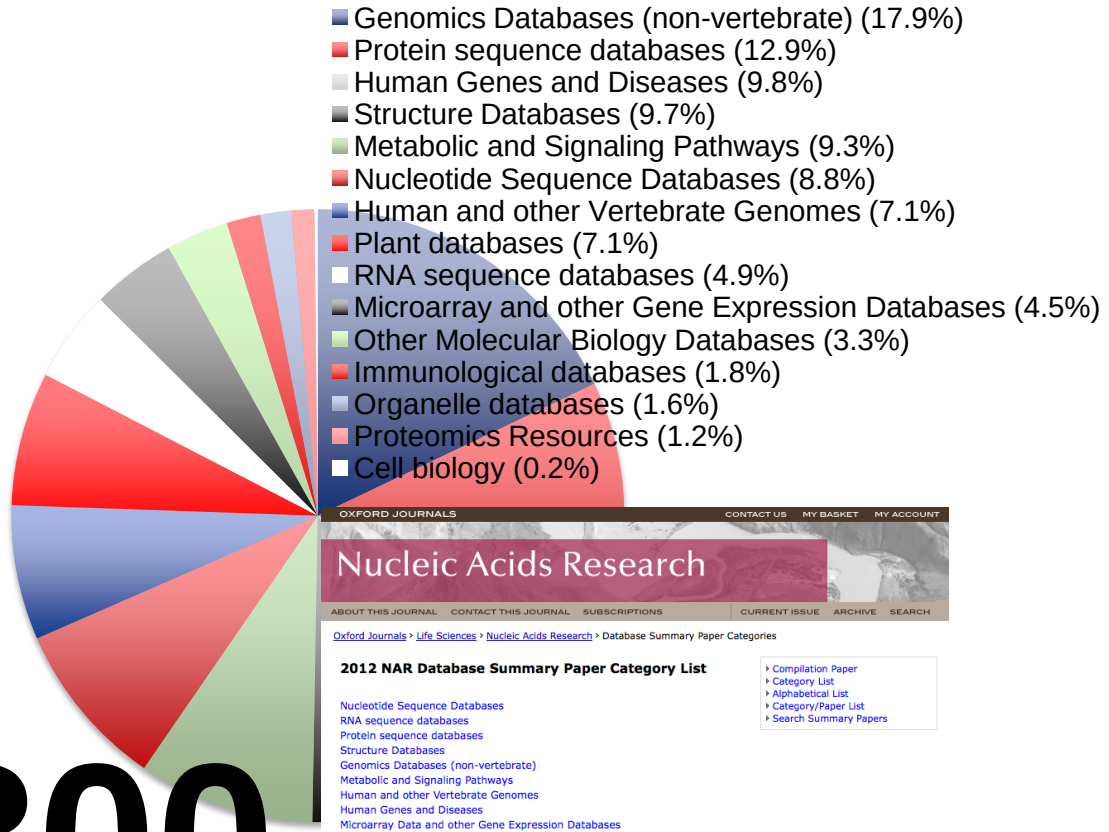


Lay the foundation for long-term sustainability of core resources.

Define and deliver visible and useful ELIXIR services.

Data resources in life science

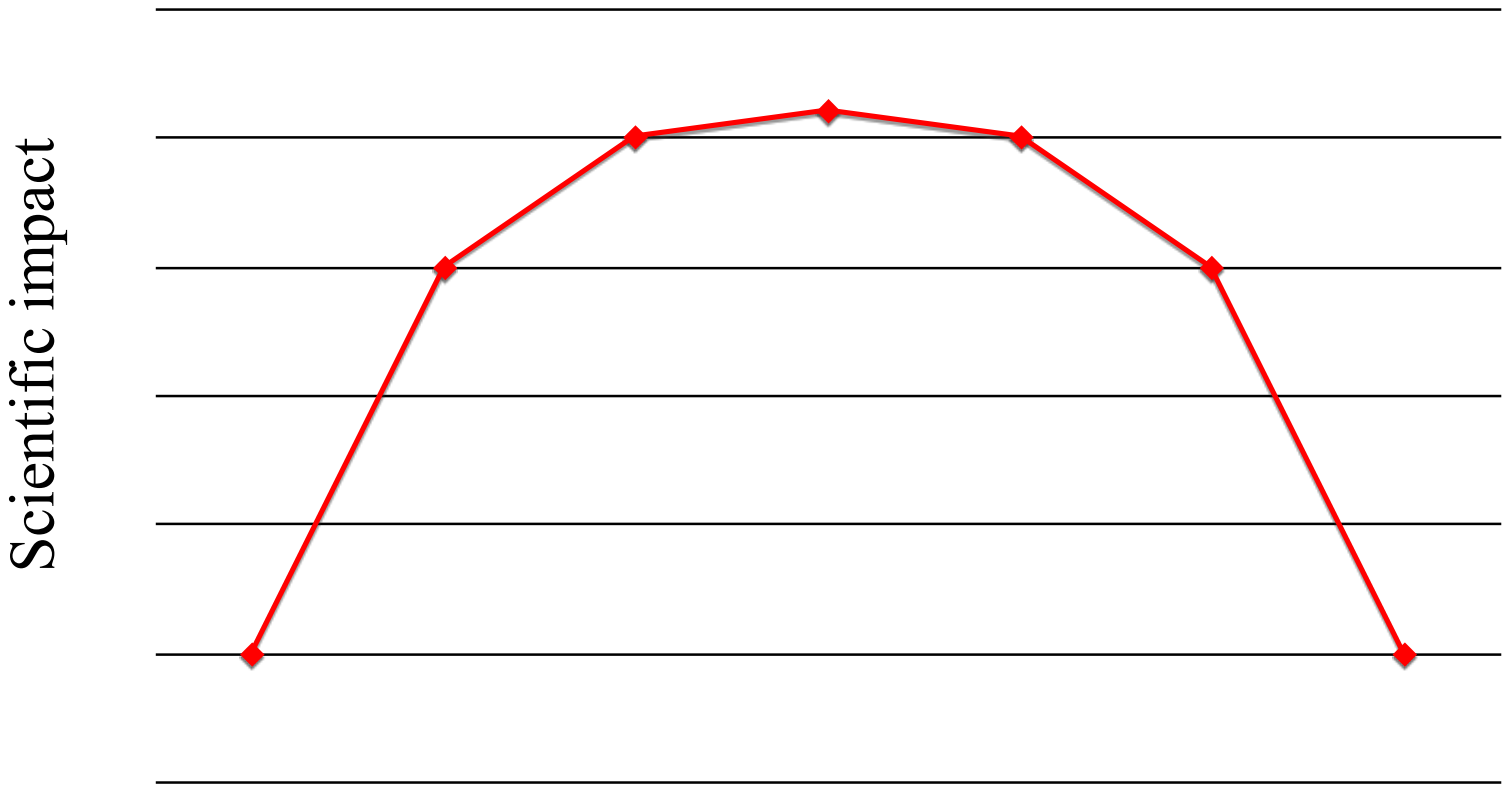
- Many
- Diverse
- Disperse



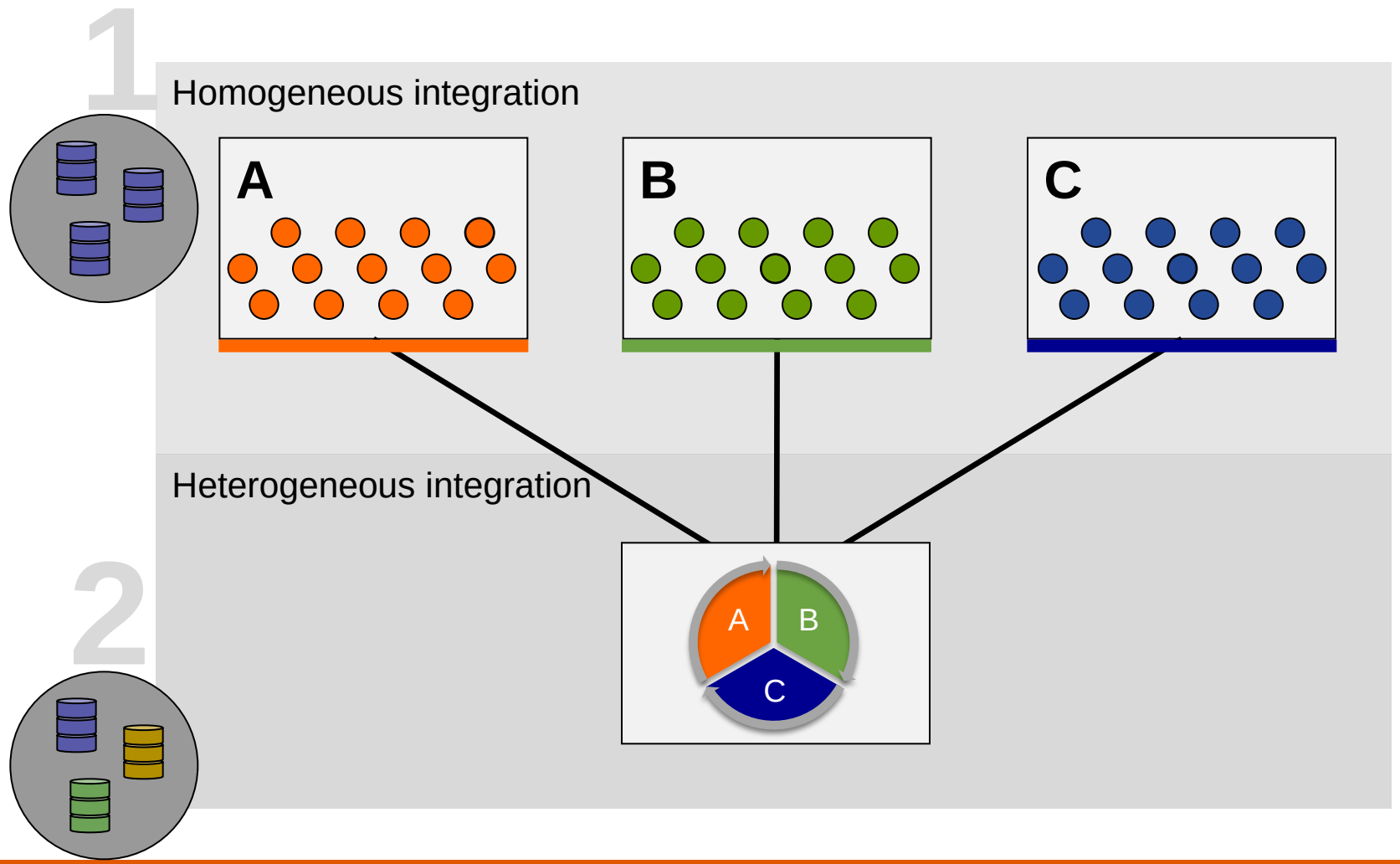
~1800
molecular biology data resources

Nucleic Acids Research annual Database Issue
and the NAR online Molecular Biology Database Collection in 2012.
MY Galperin, GR Cochrane – Nucleic Acids Research, 2011

Utility of databases

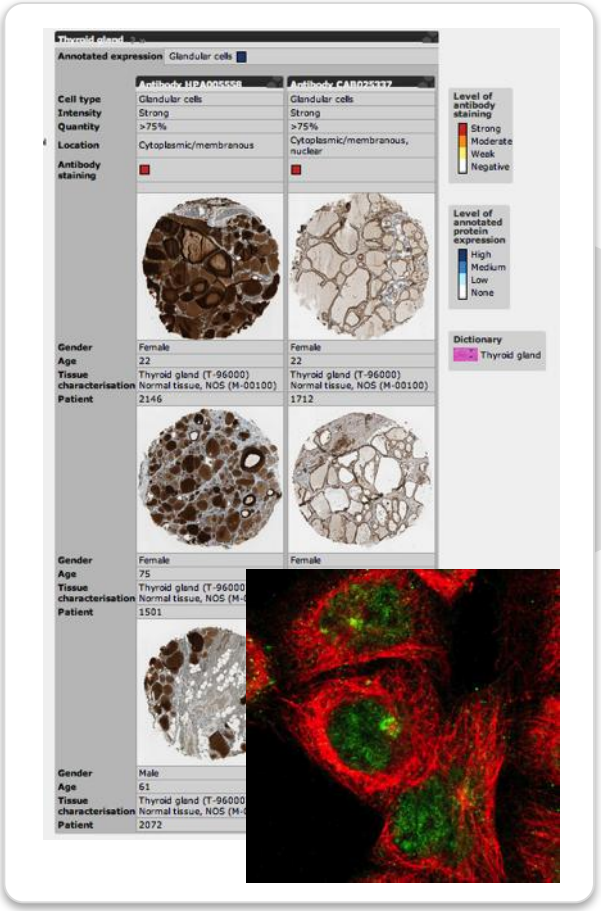


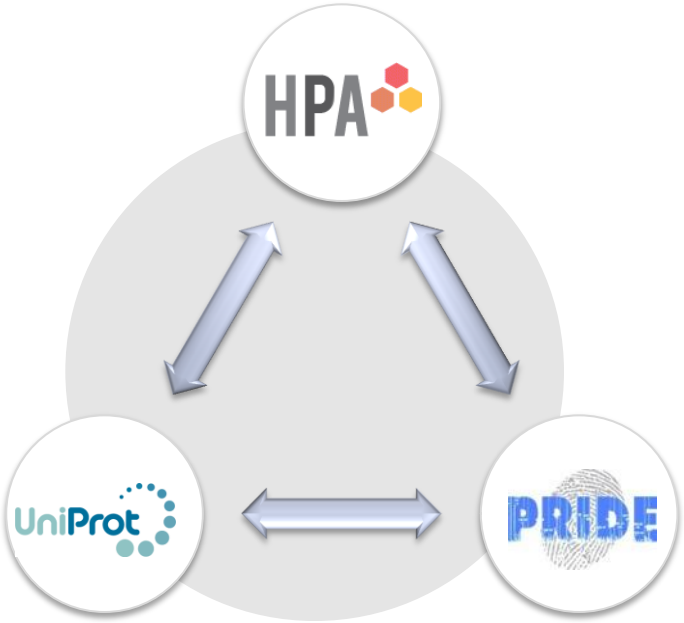
Data interoperability



Improving Links Between distributed European resources

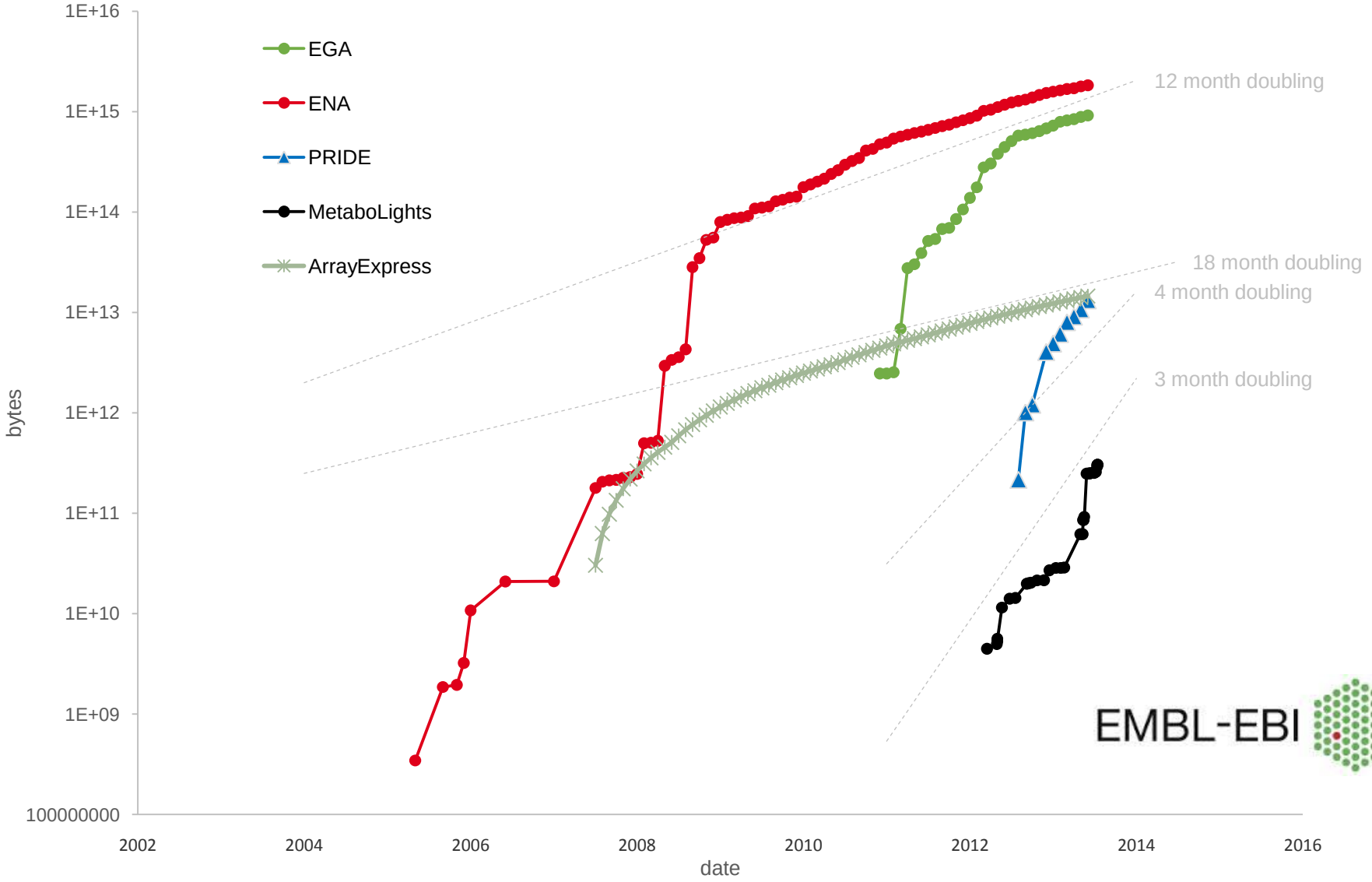
ELIXIR pilot: Interoperability of protein expressions resources





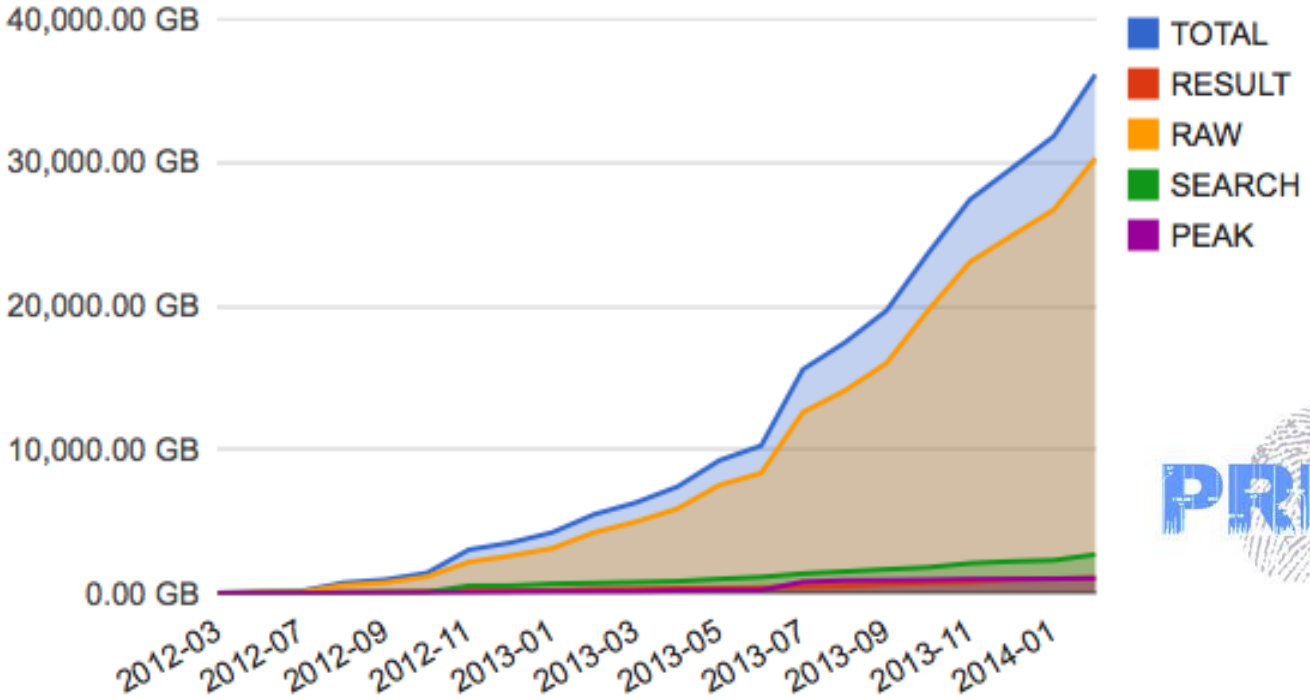
The Human Protein Atlas portal is a publicly available database with millions of high-resolution images showing the spatial distribution of proteins in 46 different normal human tissues and 20 different cancer types, as well as 47 different human cell lines.

Growing data

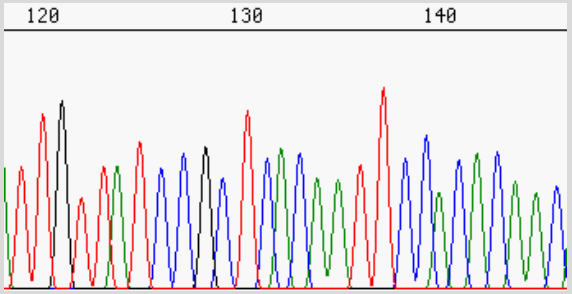
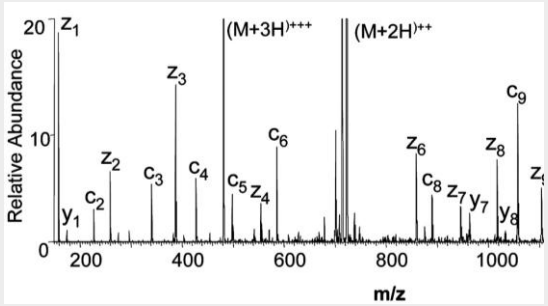


Proteomics data in PRIDE

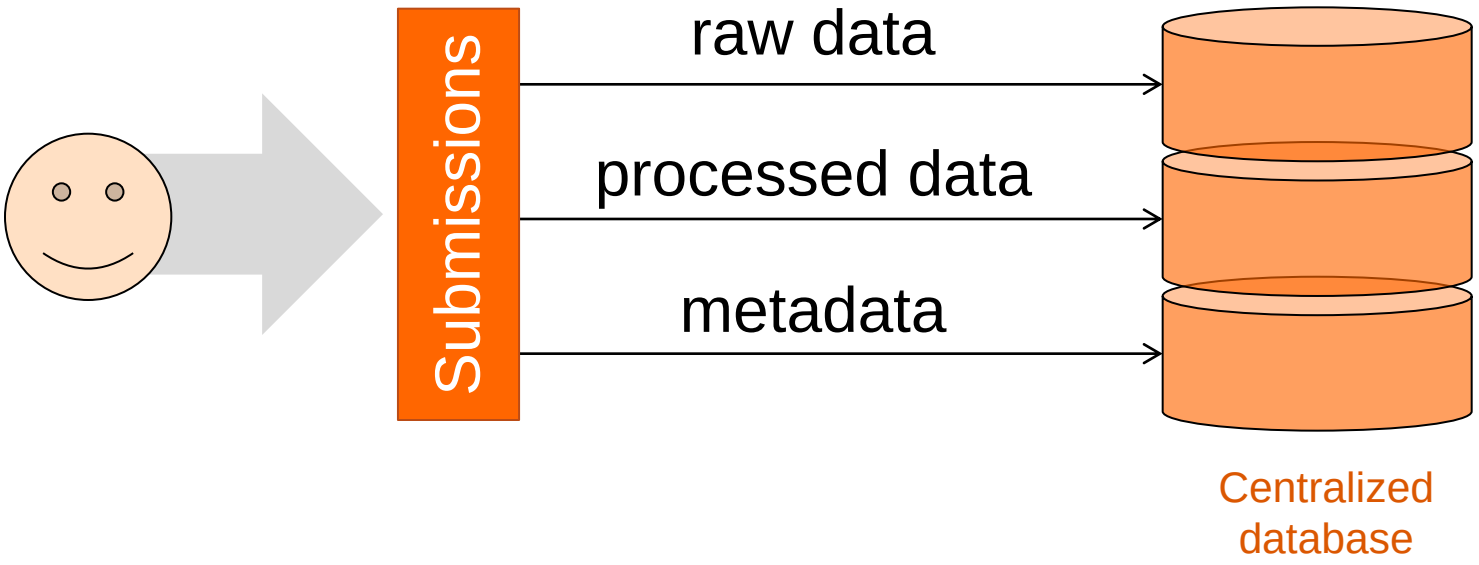
~85% raw data



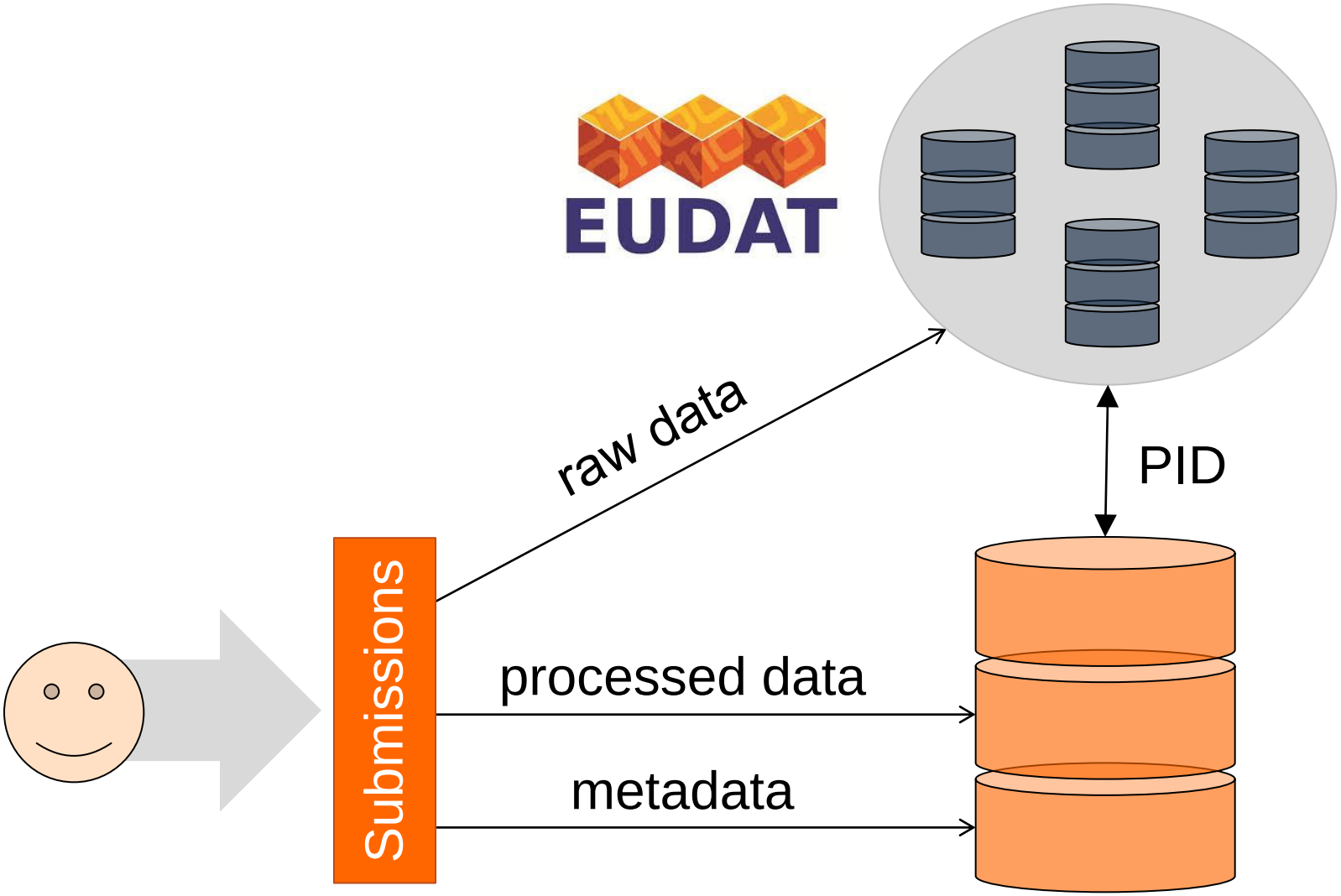
Data types examples

Raw data	Process data	Metadata
	TTGTTATCCG…	***DNA*** *Human* *Liver* *Mitochondria* *W. Smith* …
	LPISASHSSK…	***Peptide*** *Mouse* *Heart* *Nucleus* *J. Heinz* …
…	…	…

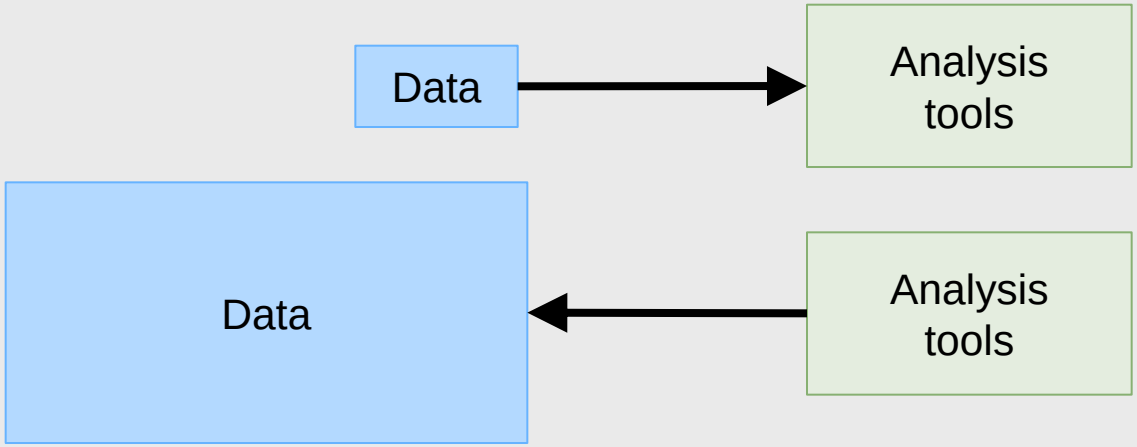
Data submission

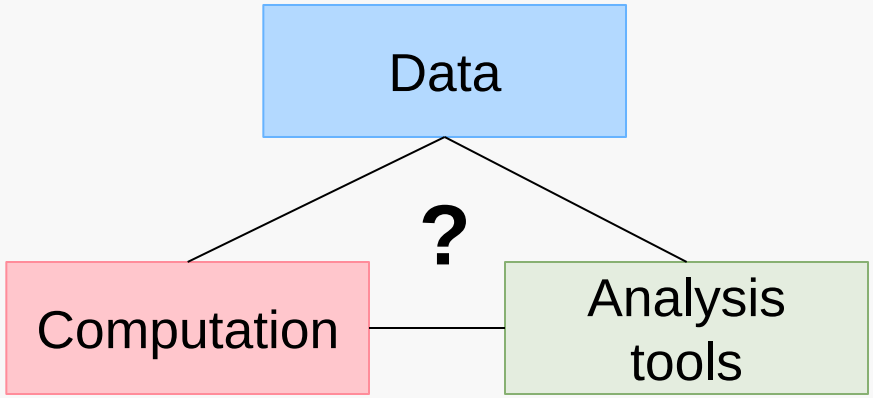


Data submission - pilot



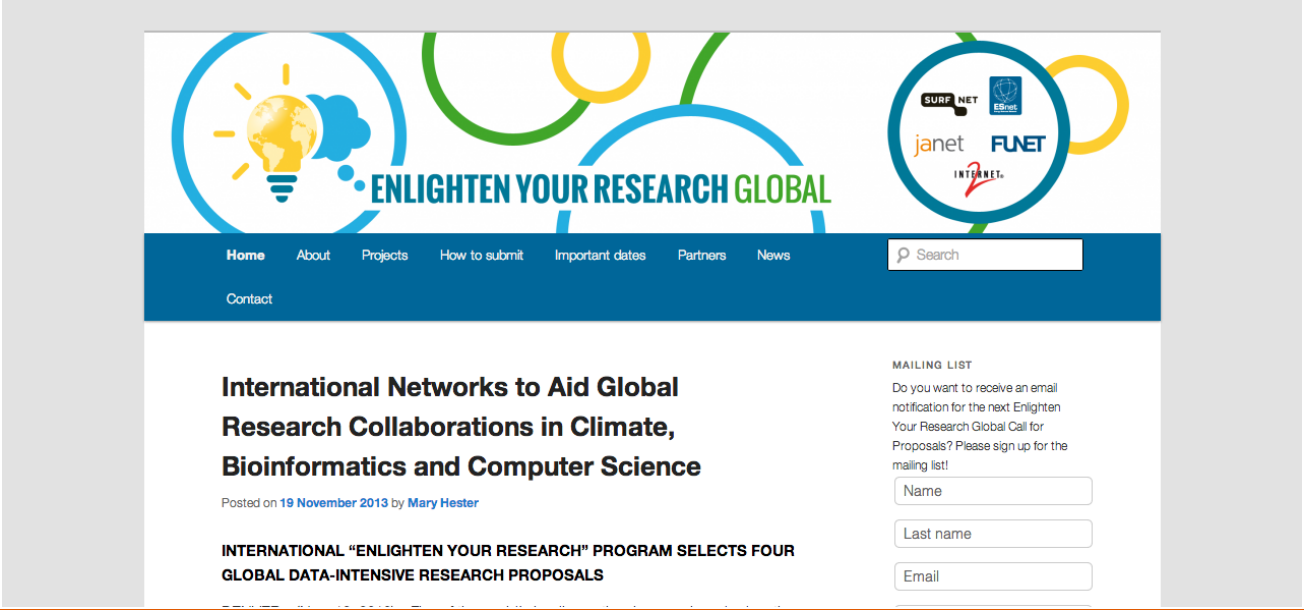
Data analysis



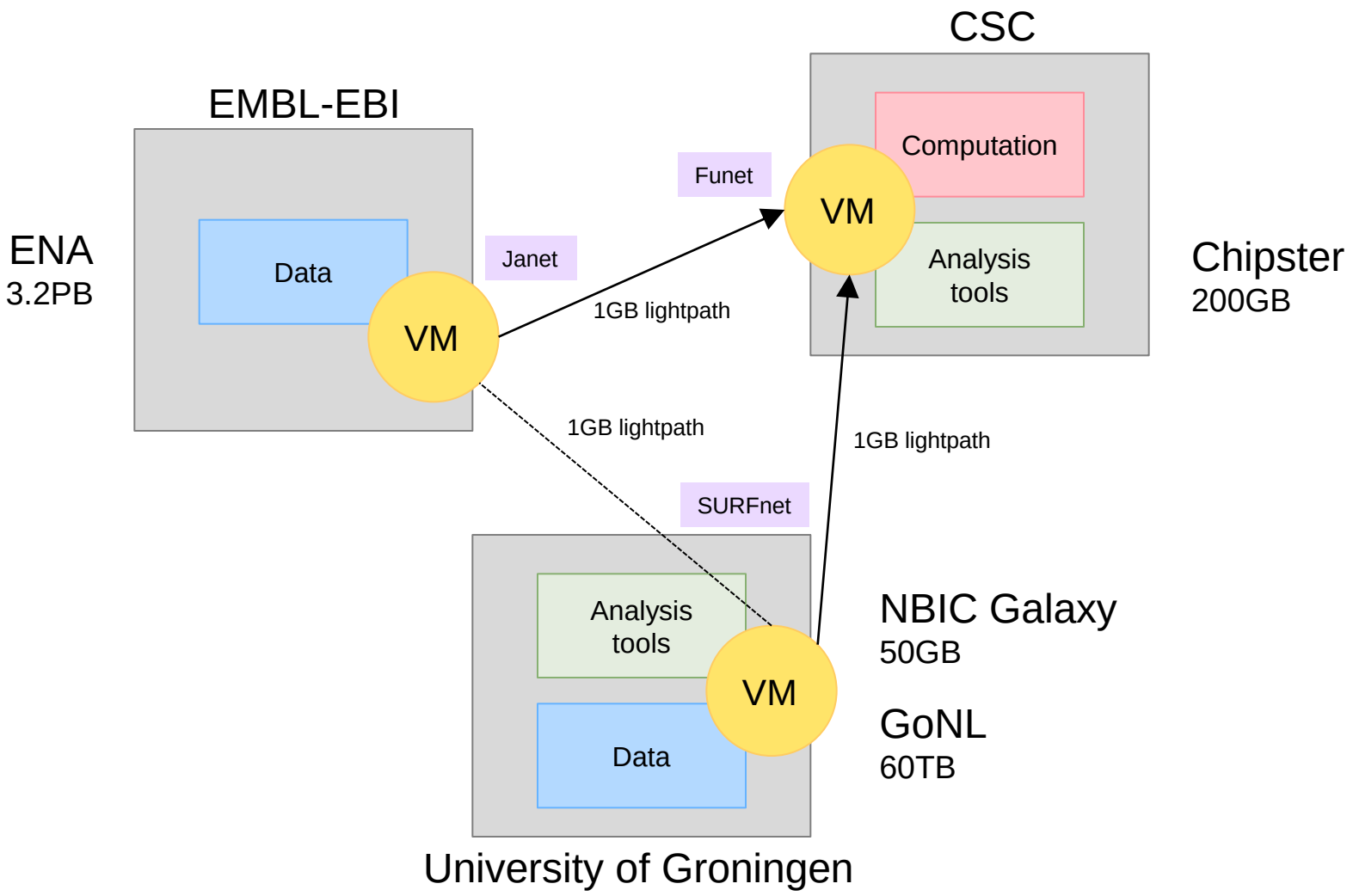


Cross-site VM Operation - pilot

- Perform analysis via cloud infrastructures and VMs
- Transfer VMs between computing centers to allow researchers to perform analyses that they could not otherwise do locally
- Supported by 5 NRENs and in collaboration with

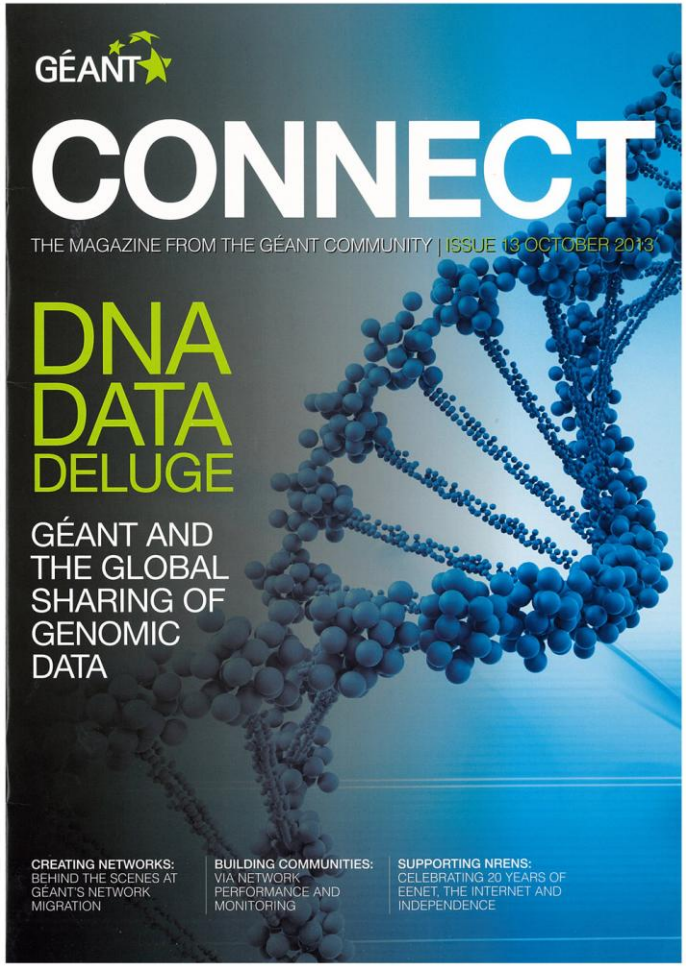


Cross-site VM Operation



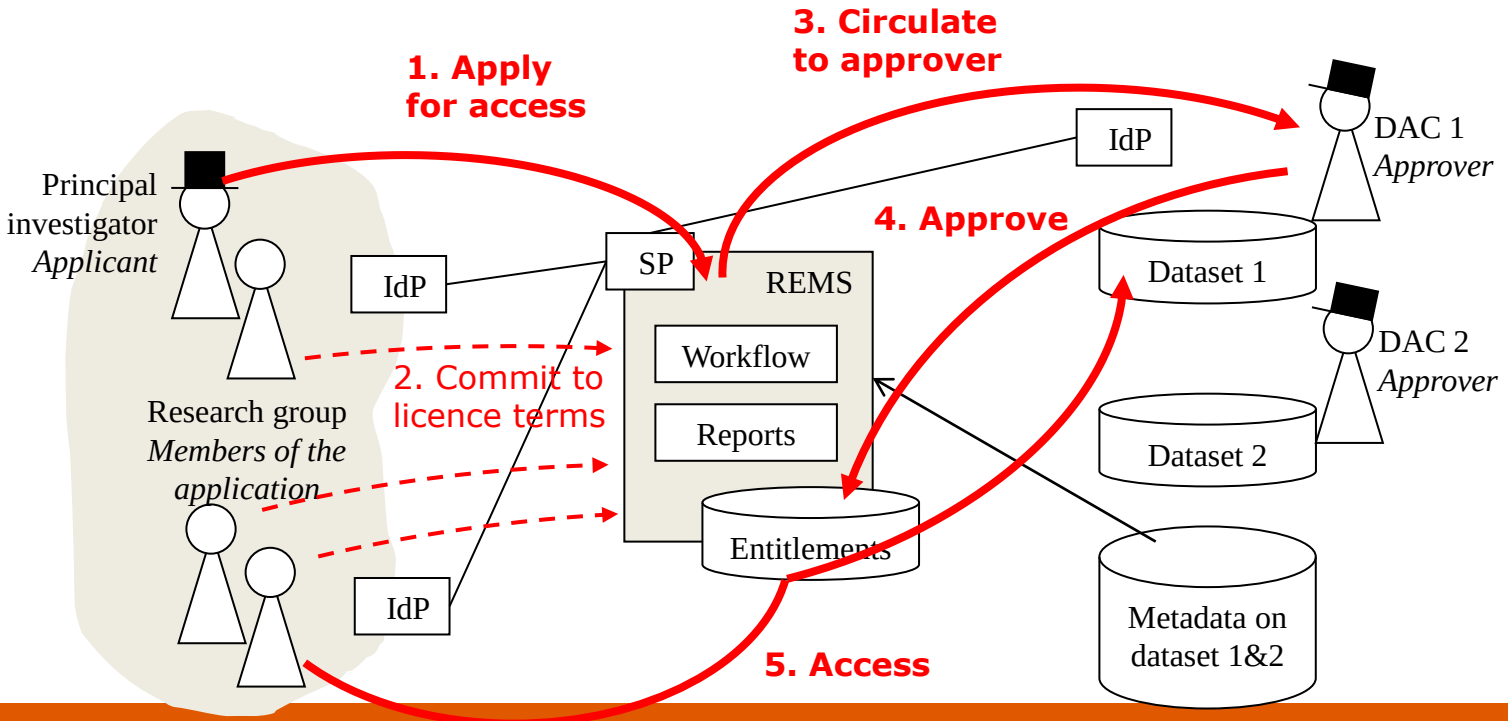
European ELIXIR Data - "LightPath" (EBI / CSC)

- Aim
  - To explore the replication of large scale (Petabyte scale) archives to remote sites
  - To create a separate source of data files for challenging DataIO projects
- Update:
  - Selection of pilot data transfer technology between EBI and CSC
  - Established a dedicated light path between datacenters in London and Kajaani
  - Development of model for future IO needs in the lifesciences in Europe



REMS - Resource Entitlement Management System

- Access to sensitive data (genomics) granted by a Data Access Committee
- In collaboration with eduGAIN
- Agreements to be applied to other domains: FI-CLARIN & FI-CESSDA



https://tnc2013.terena.org/getfile/421

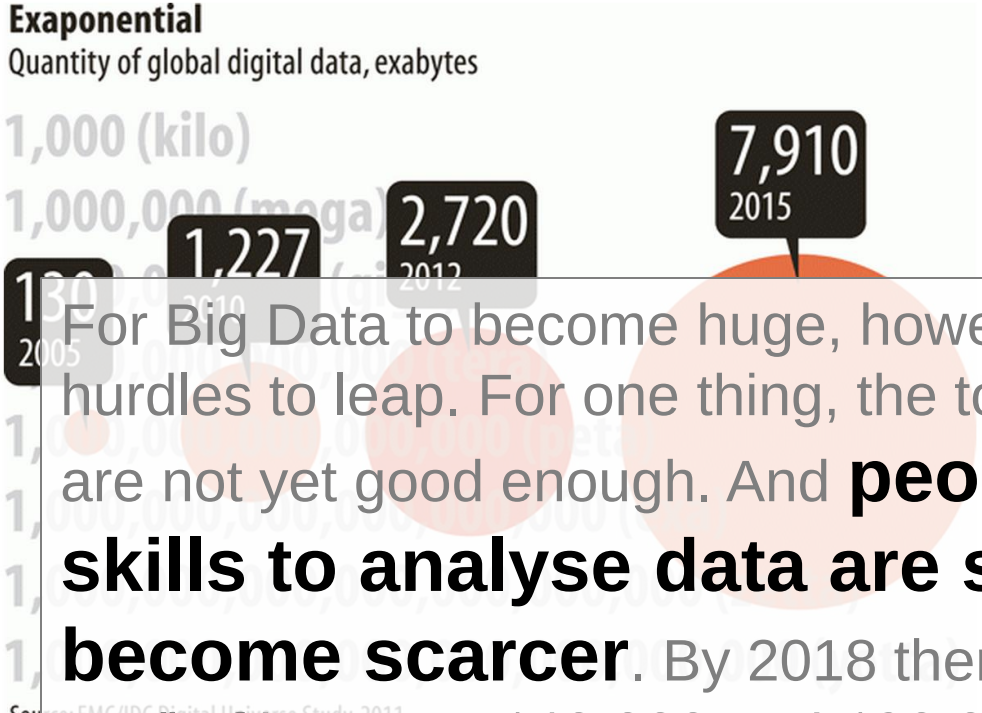
Starting point: https://remsdemo.csc.fi/

Business

Welcome to the yotta world

Big Data will flood the planet

Nov 17th 2011 | From The World In 2012 print edition



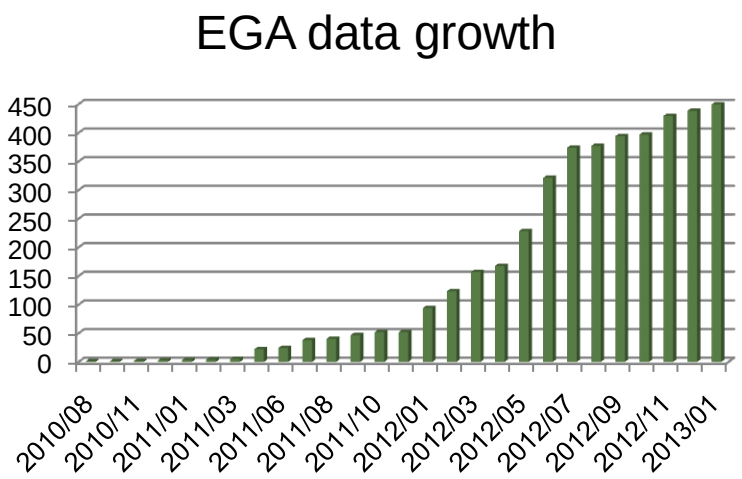
Source: EMC/IDC Digital Universe Study, 2011

Even if you still have to think twice about the meaning of "giga" and "tera" in computer-speak, you'd better get ready for "peta", "exa" and "zetta". These binary prefixes, which

For Big Data to become huge, however, there are still hurdles to leap. For one thing, the tools to analyse data are not yet good enough. And **people with the skills to analyse data are scarce and will become scarcer**. By 2018 there will be a "talent gap" of between 140,000 and 190,000 people, …

European Genome-phenome archive

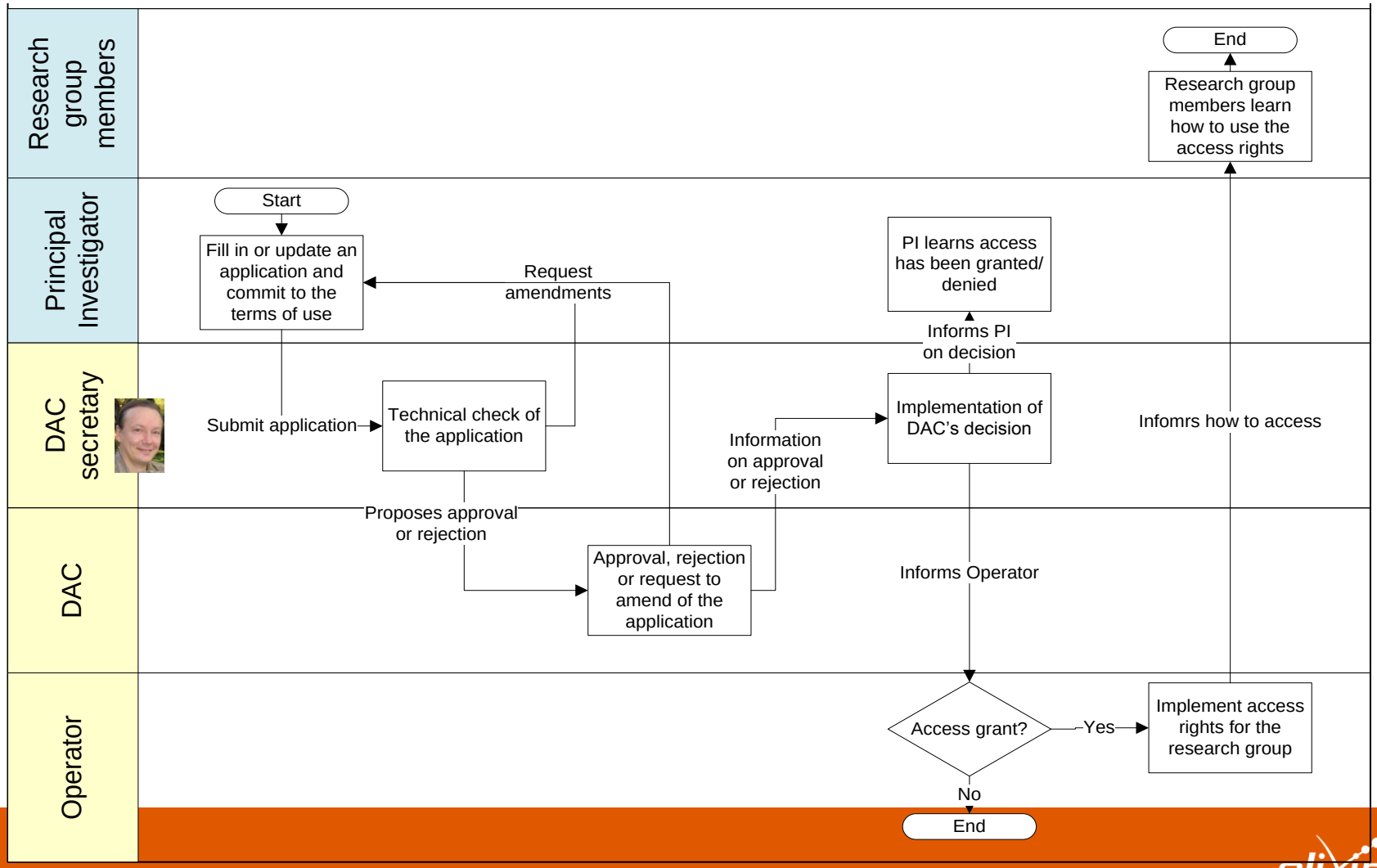
- EGA is to be **distributed effort** with **archive**, **submission**, and **data distribution** capacity at both the EBI and CRG



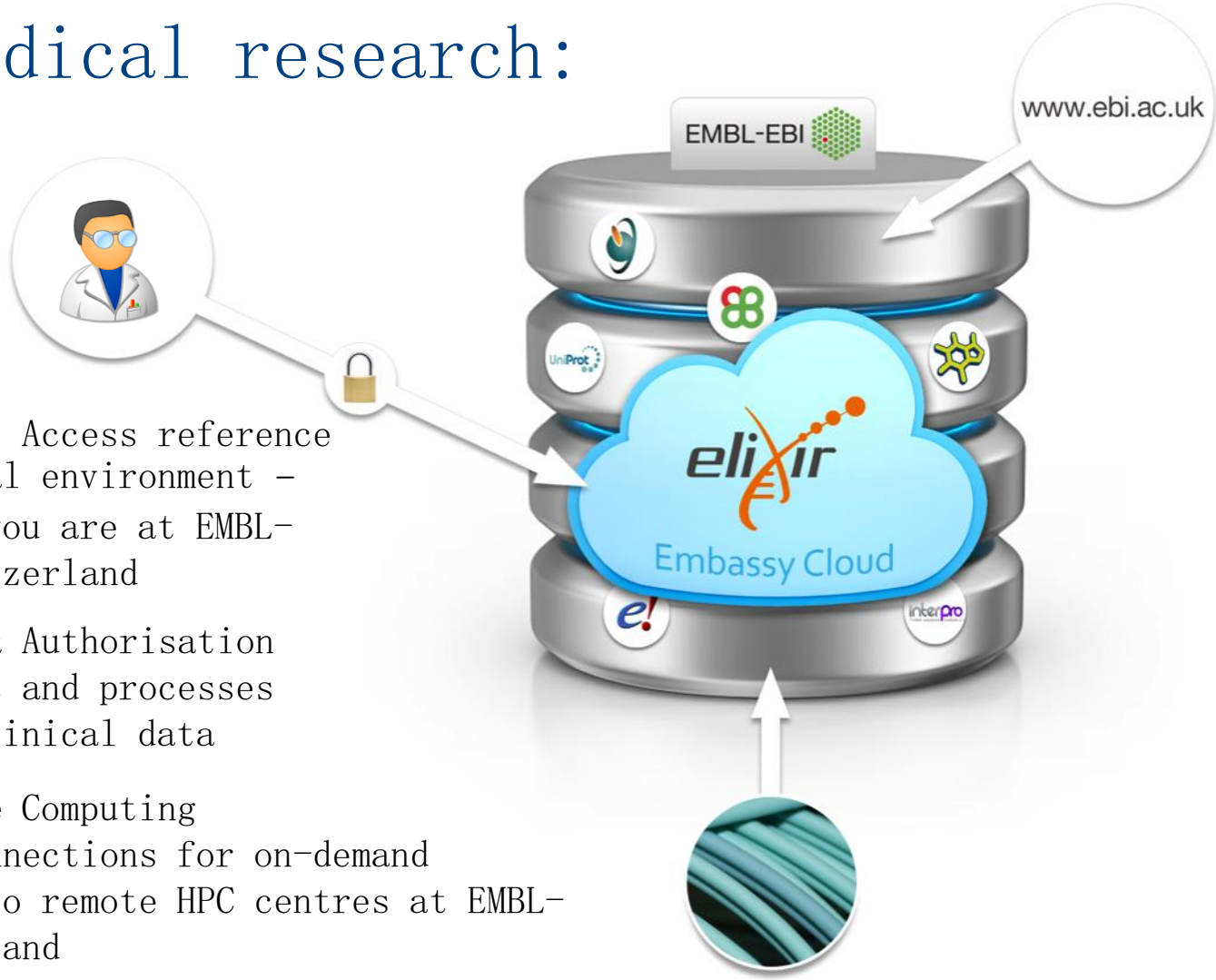
- From the users point of view, EGA remains one **integrated** Archive of **secure human biomedical research data**.

- **Search** of datasets at either website is "**global**" across the EGA.

CASE: process for applying access to the Nordic Control Database



ELIXIR pilots to address key challenges in biomedical research:

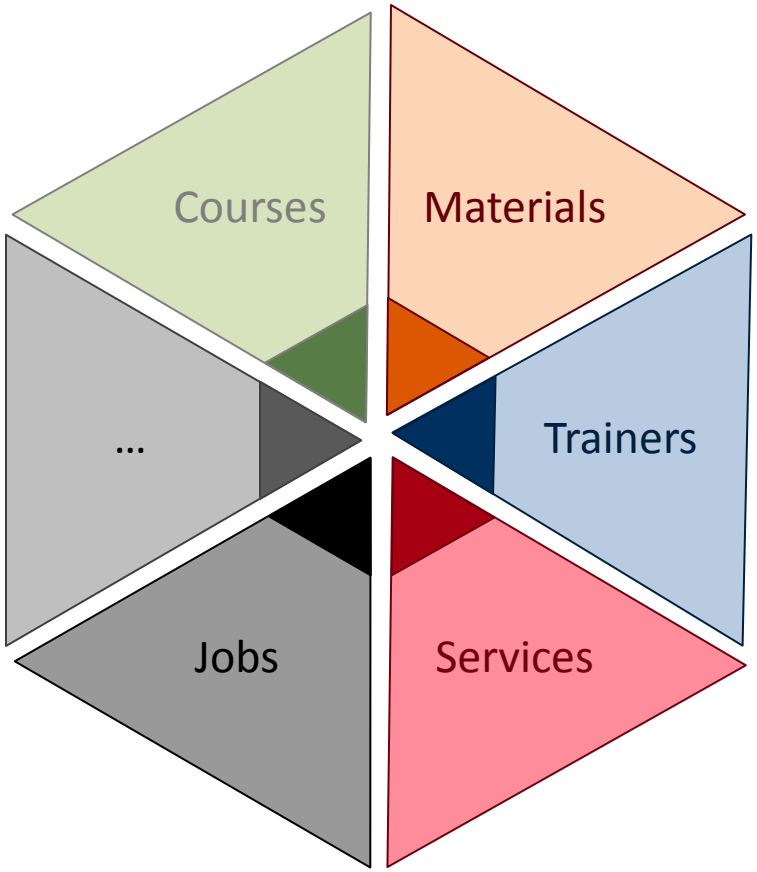


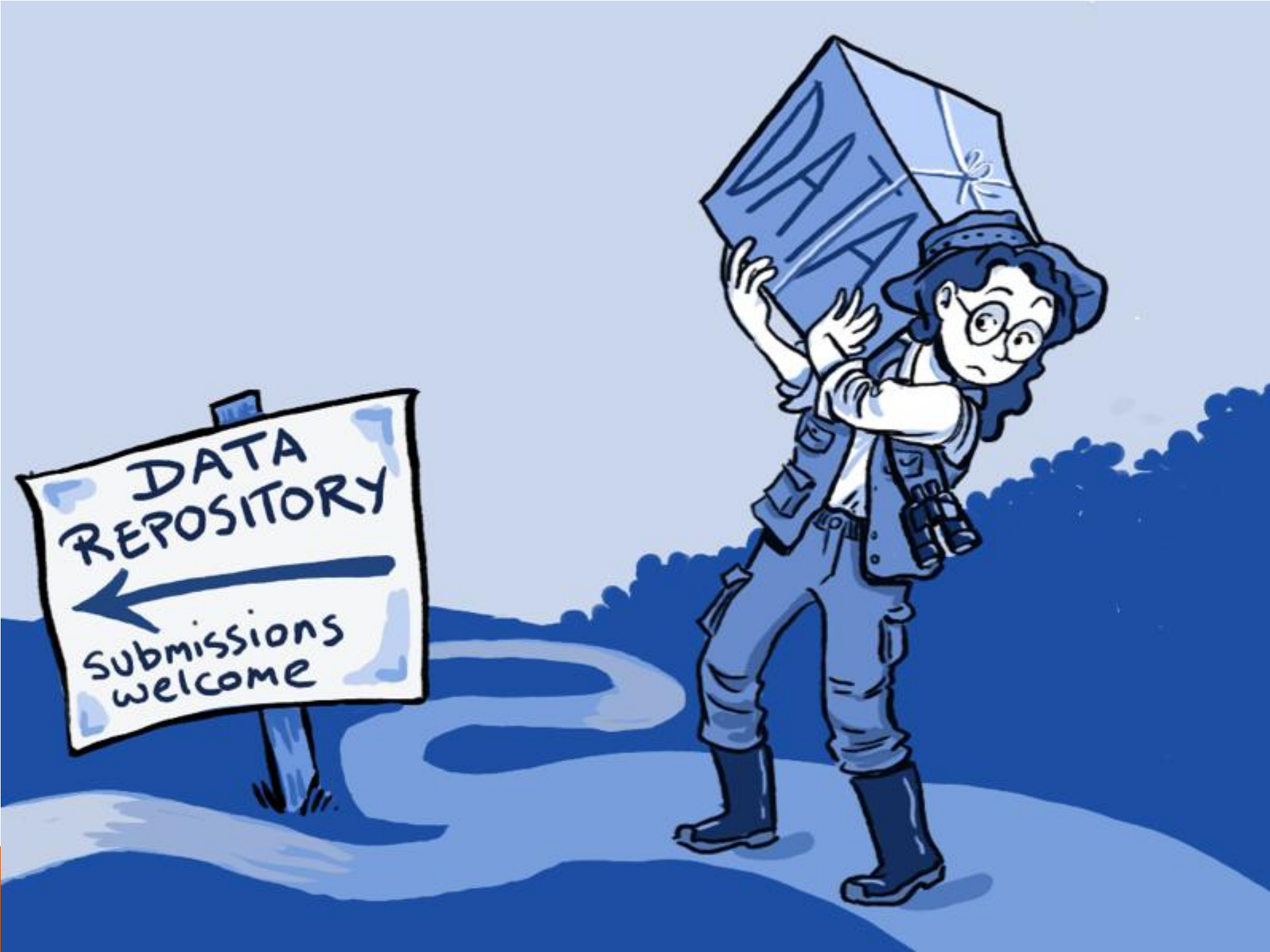
1. Cloud computing
   "Embassy cloud": Access reference data in a virtual environment - work as though you are at EMBL-EBI or SIB, Switzerland
2. Authentication & Authorisation
   Improved methods and processes for access to clinical data
3. High-Performance Computing
   "Lightpath": Connections for on-demand reference data to remote HPC centres at EMBL-EBI and CSC Finland

Minimum metadata

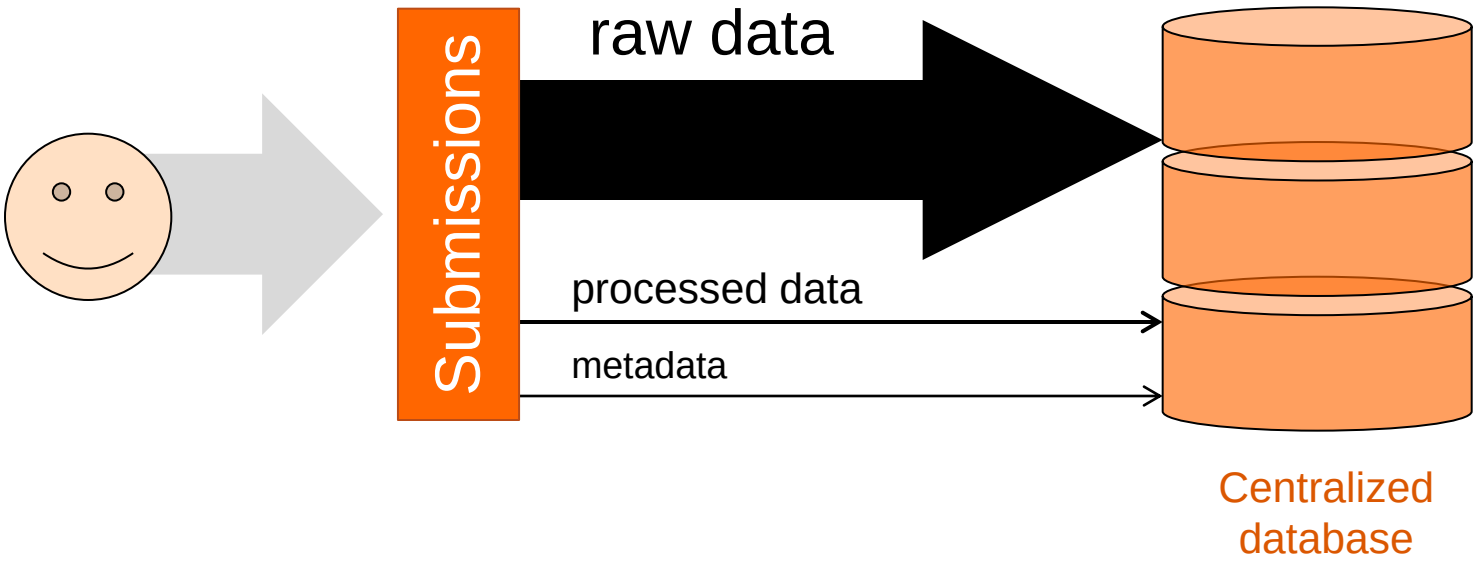
life sciences

- Title
- Description
- Creator
- Publication Date
- **Topics**
- **Audience**
- …




DATA
DATA REPOSITORY
submissions welcome

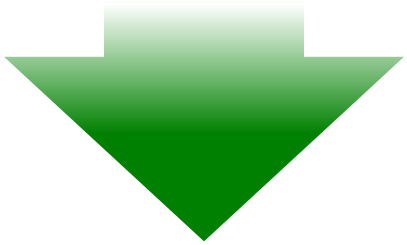
Data submission



Data sharing

The casual approach

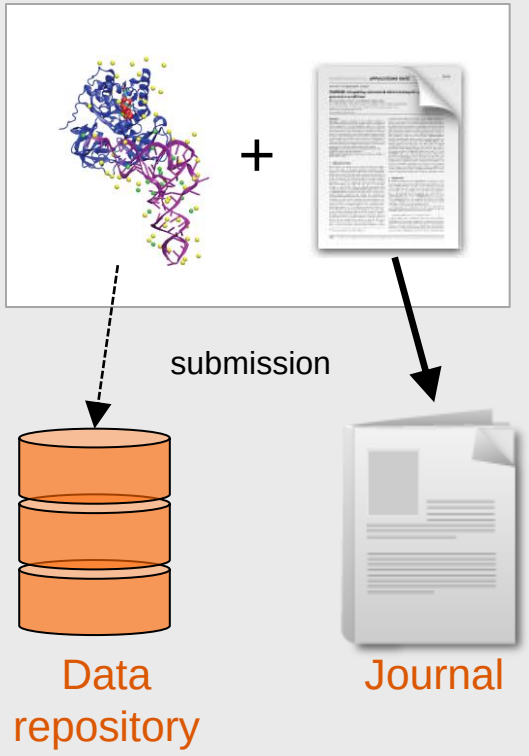
'data on my disk and available to anyone who requests it'



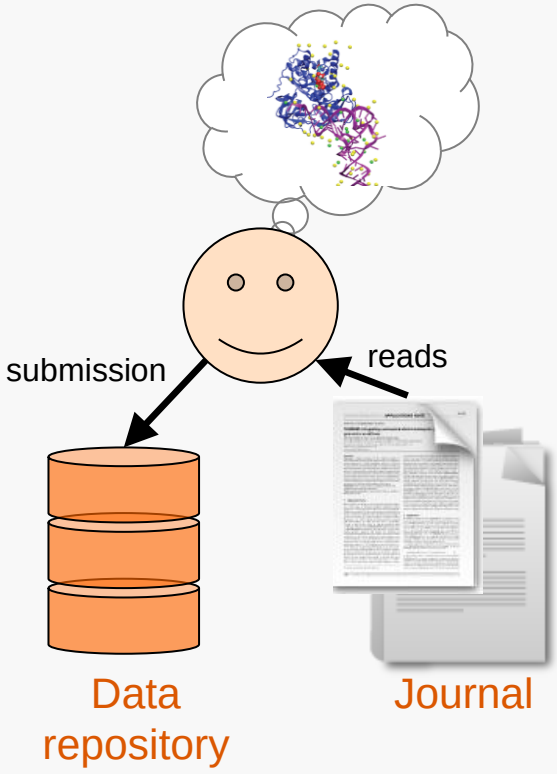
Submission to data repositories

Data submissions

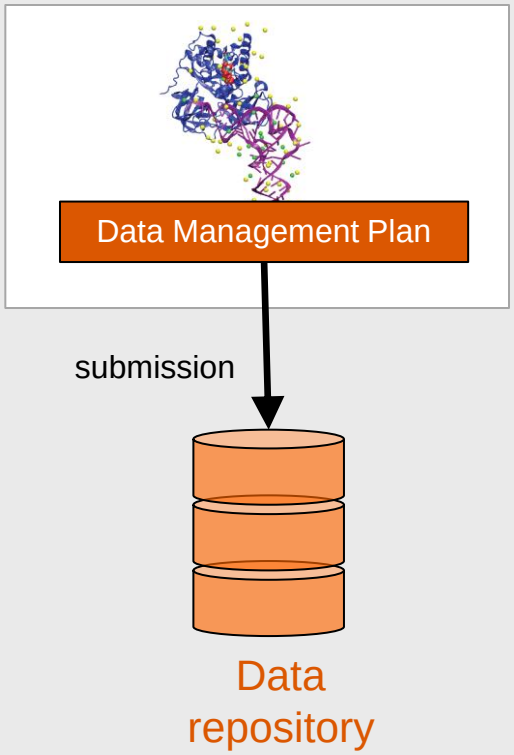
Journal request



Curator



Data management

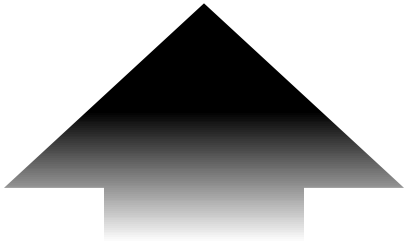


Data sharing

Will big data affect data deposition?

The casual approach

'data on my disk and available to anyone who requests it'



Submission to data repositories

Data submissions

How much data?

How much available data?