ELIXIR

Technical activities in ELIXIR Europe

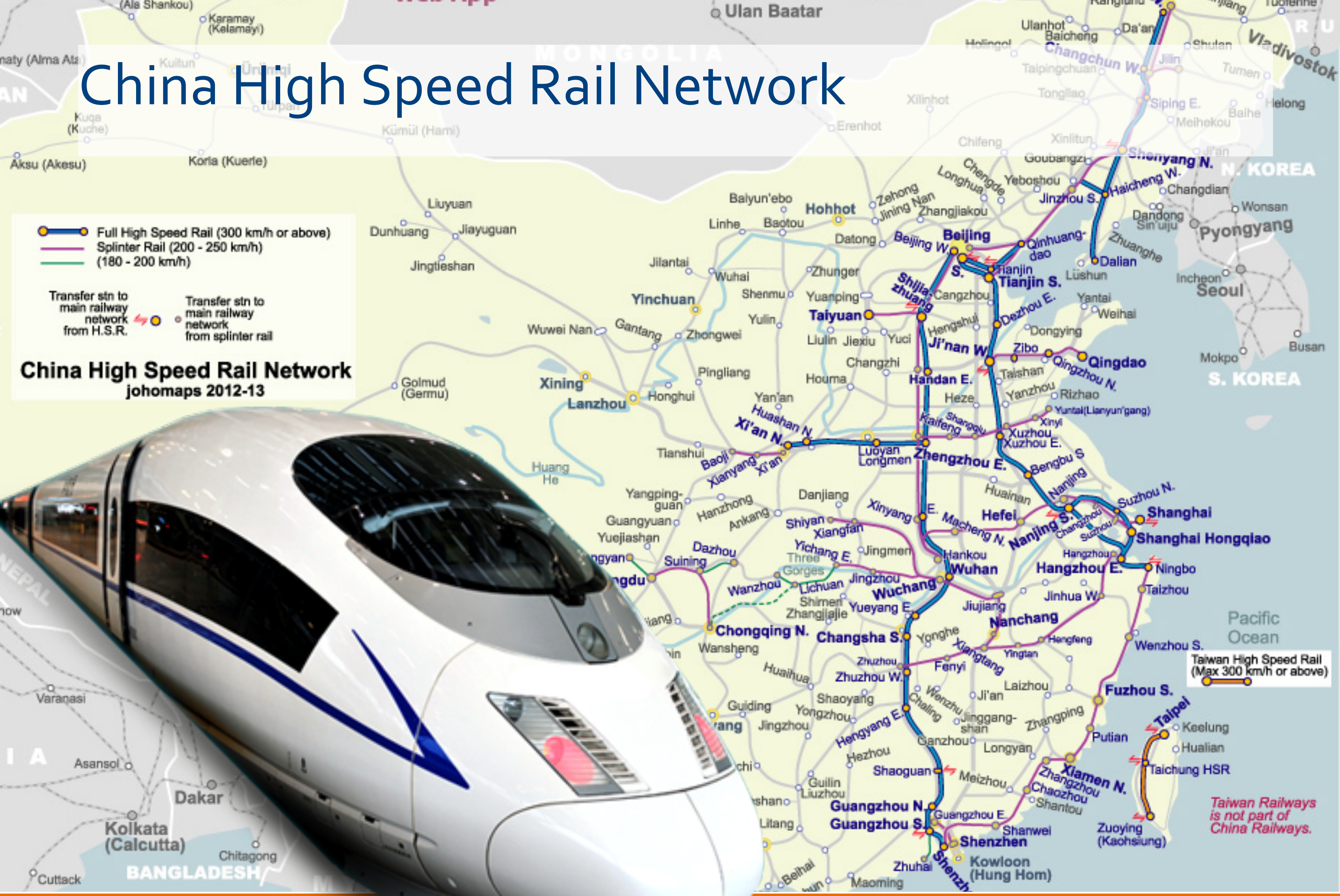
Rafael C Jimenez

ELIXIR CTO

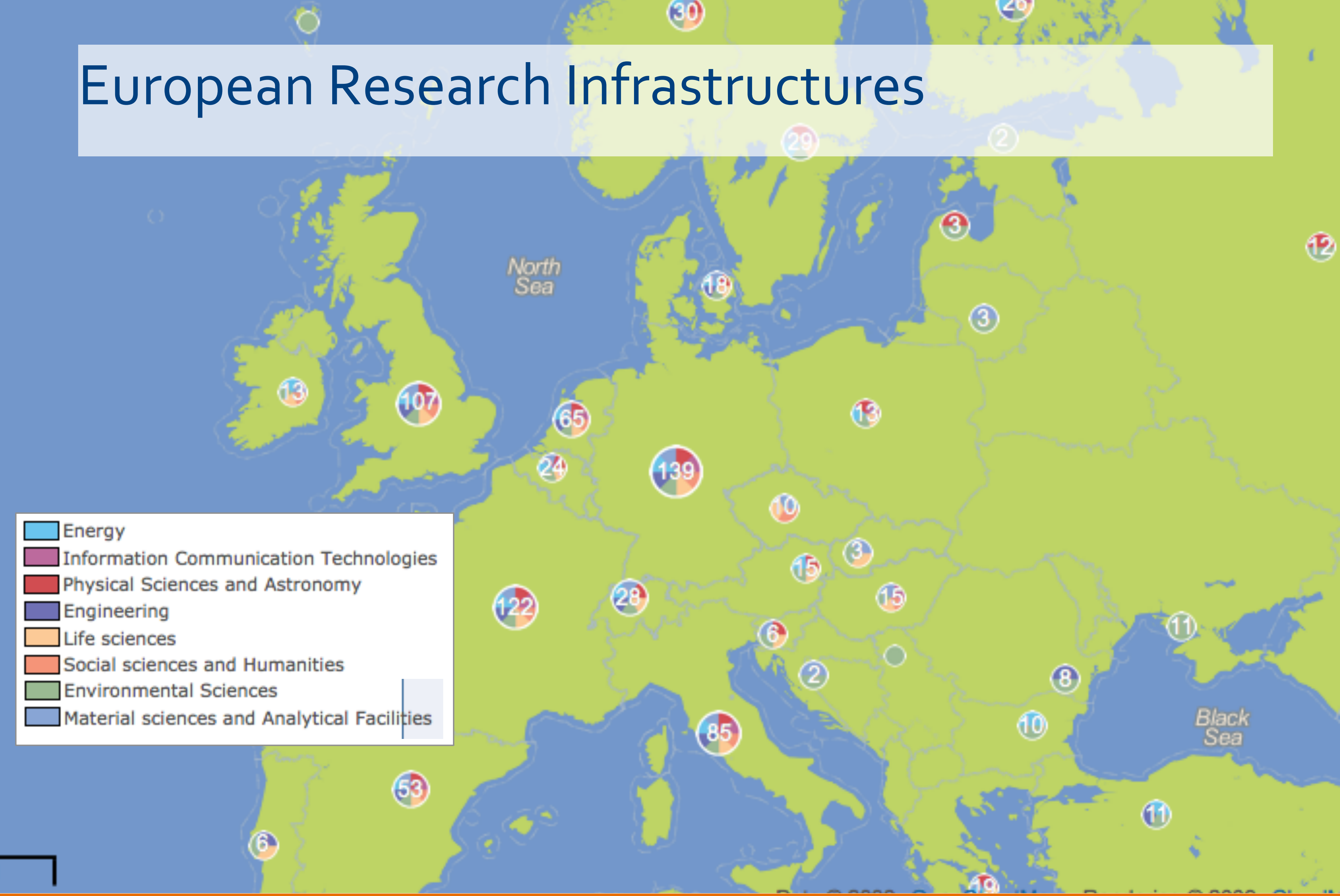
EGI forum 2014, Tuesday 20 May

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

China High Speed Rail Network



European Research Infrastructures


Energy
Information Communication Technologies
Physical Sciences and Astronomy
Engineering
Life sciences
Social sciences and Humanities
Environmental Sciences
Material sciences and Analytical Facilities
North Sea
Black Sea
30
26
29
2
3
12
18
3
13
107
65
13
24
139
10
3
15
15
122
28
6
11
2
8
10
85
53
6
11
19

European Research Infrastructures

life sciences

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURES NETWORK

BBMRI
Biobanking and
Biomolecular
Resources Research
Infrastructure

instruct
Integrating
Biology

elixir

EATRIS
European Infrastructure for
Translational Medicine

EU-OPENSCREEN
ESFRI ROADMAP 2008

EMBRC
EUROPEAN
MARINE
BIOLOGICAL
RESOURCE
CENTRE

EURO-BIOIMAGING

LS

e-infrastructures

EUDAT

GÉANT

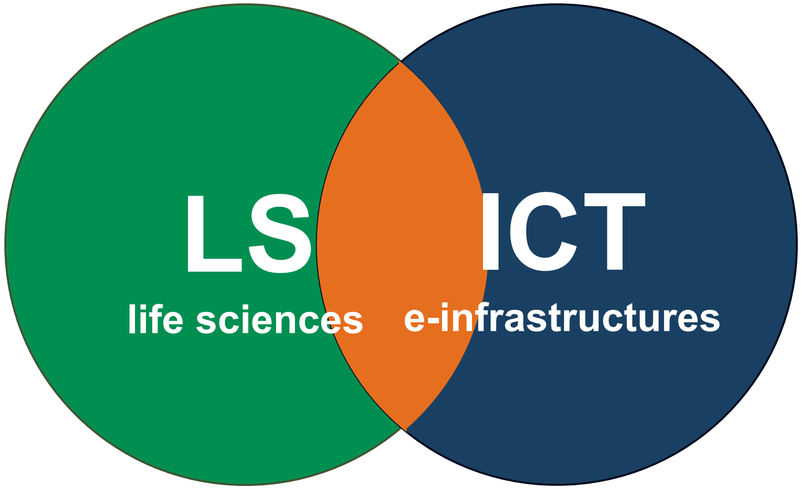
egi

PRACE

ICT

Research infrastructures

Facilitate research



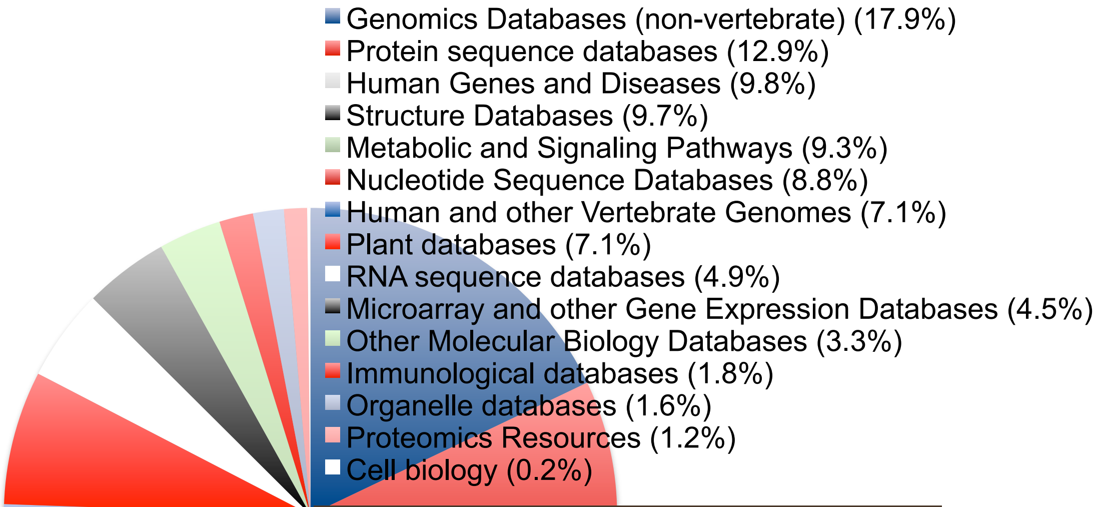
Physical facilities
Scientific information

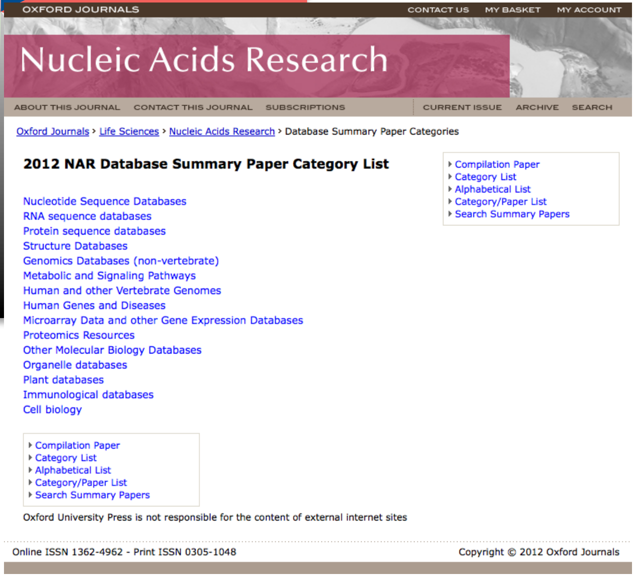
Transfer
Computation
Storage

Data

Data resources in life science

- Many
- Diverse
- Disperse





Oxford Journals › Life Sciences › Nucleic Acids Research › Database Summary Paper Categories

2012 NAR Database Summary Paper Category List

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Oxford University Press is not responsible for the content of external internet sites

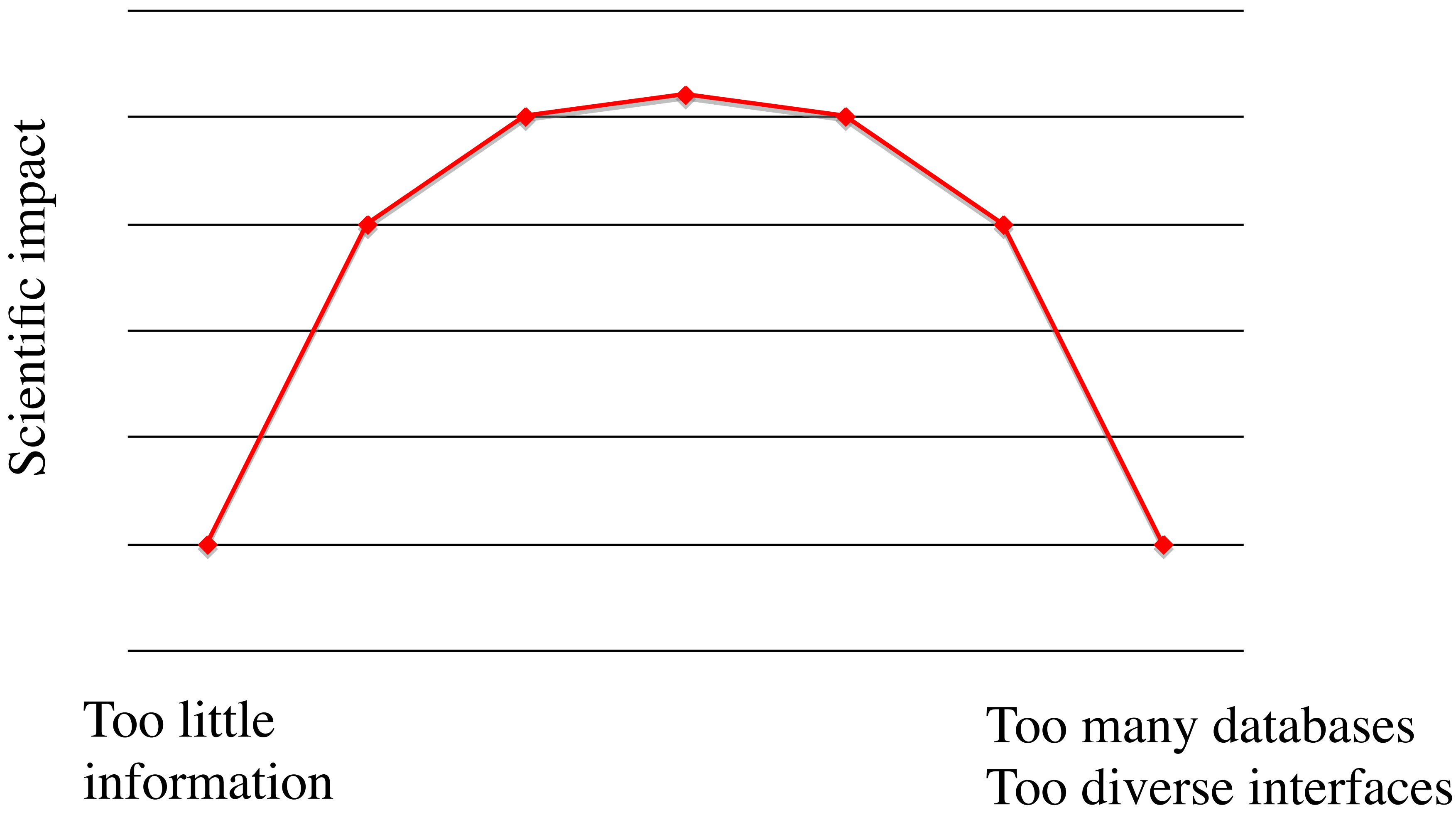
Online ISSN 1362-4962 - Print ISSN 0305-1048

Copyright © 2012 Oxford Journals

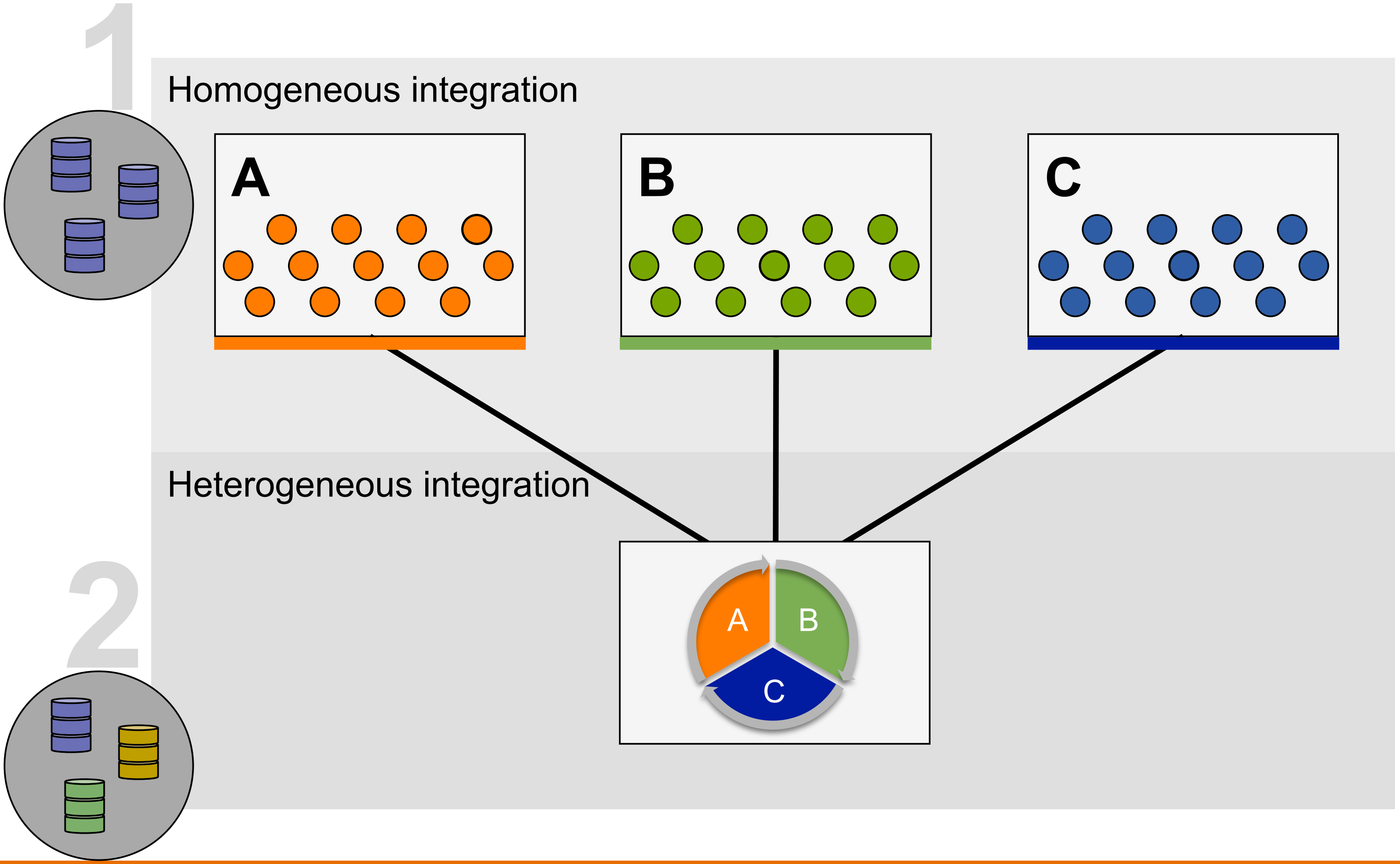
~1800
molecular biology data resources

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



1

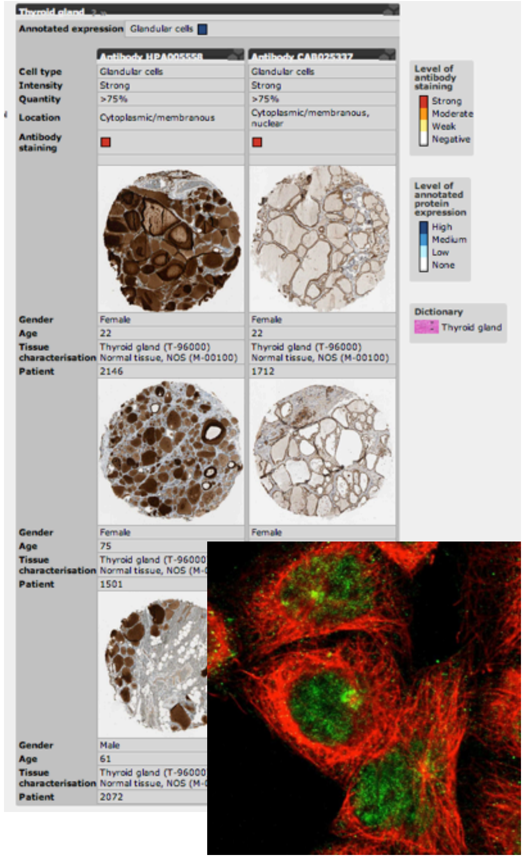
Homogeneous integration

2

Heterogeneous integration

Improving Links Between distributed European resources

ELIXIR pilot: Interoperability of protein expressions resources



Thyroid gland

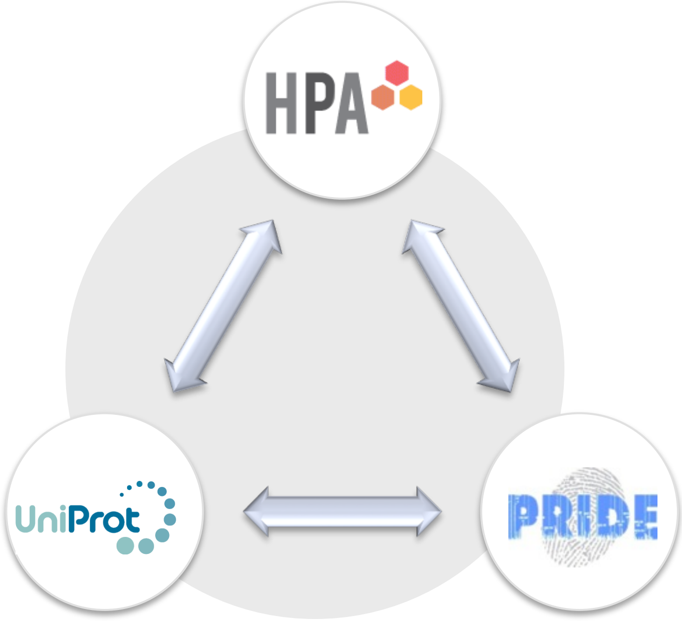
Annotated expression: Glandular cells 🟦

	Antibody HPA005558	Antibody CAB025337
Cell type	Glandular cells	Glandular cells
Intensity	Strong	Strong
Quantity	>75%	>75%
Location	Cytoplasmic/membranous	Cytoplasmic/membranous, nuclear
Antibody staining	🟥	🟥
Gender	Female	Female
Age	22	22
Tissue characterisation	Thyroid gland (T-96000) Normal tissue, NOS (M-00100)	Thyroid gland (T-96000) Normal tissue, NOS (M-00100)
Patient	2146	1712
Gender	Female	Female
Age	75	
Tissue characterisation	Thyroid gland (T-96000 Normal tissue, NOS (M-	
Patient	1501	
Gender	Male	
Age	61	
Tissue characterisation	Thyroid gland (T-96000 Normal tissue, NOS (M-	
Patient	2072	

Level of antibody staining
- Strong
- Moderate
- Weak
- Negative

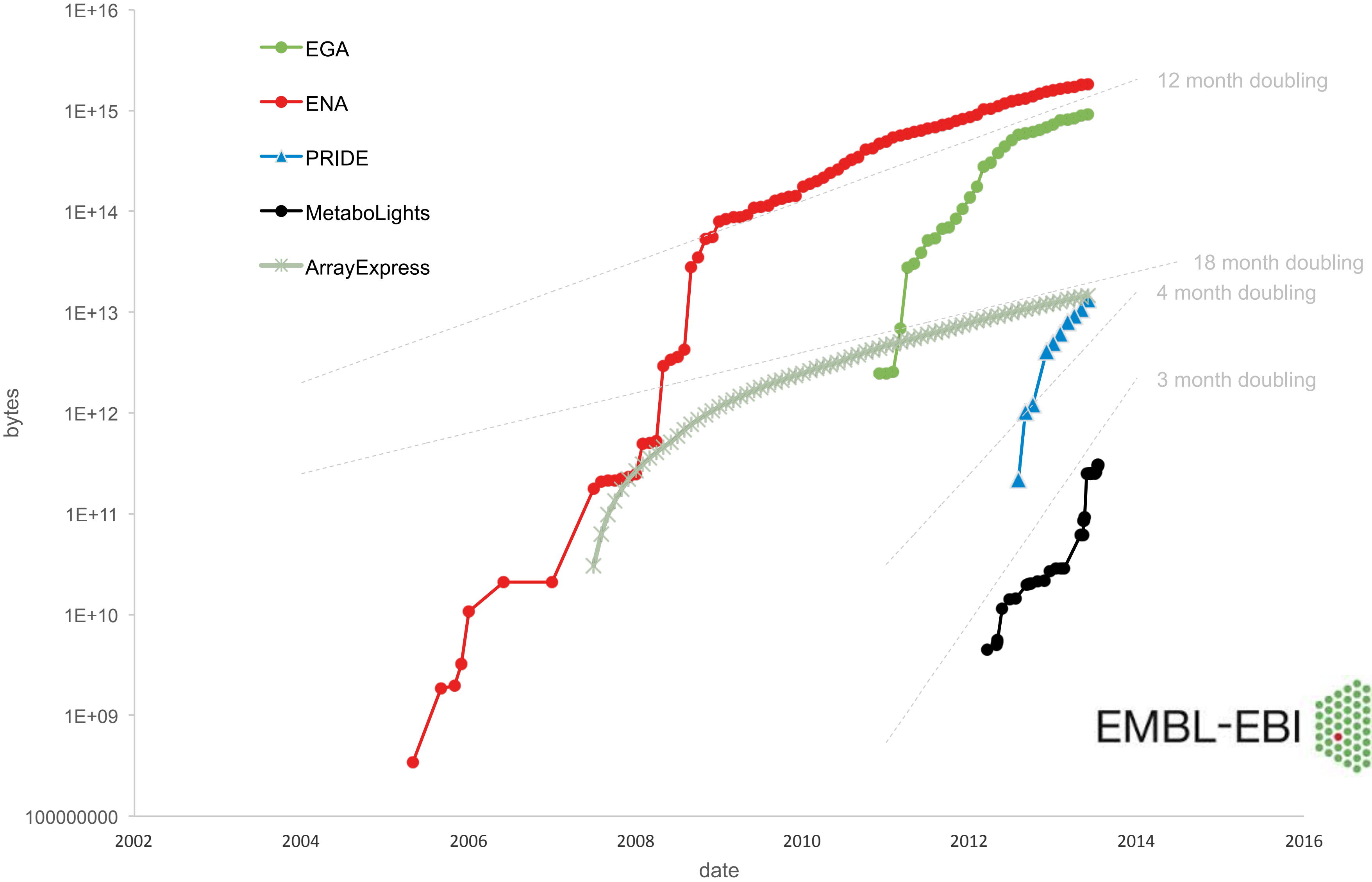
Level of annotated protein expression
- High
- Medium
- Low
- None

Dictionary
- Thyroid gland



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

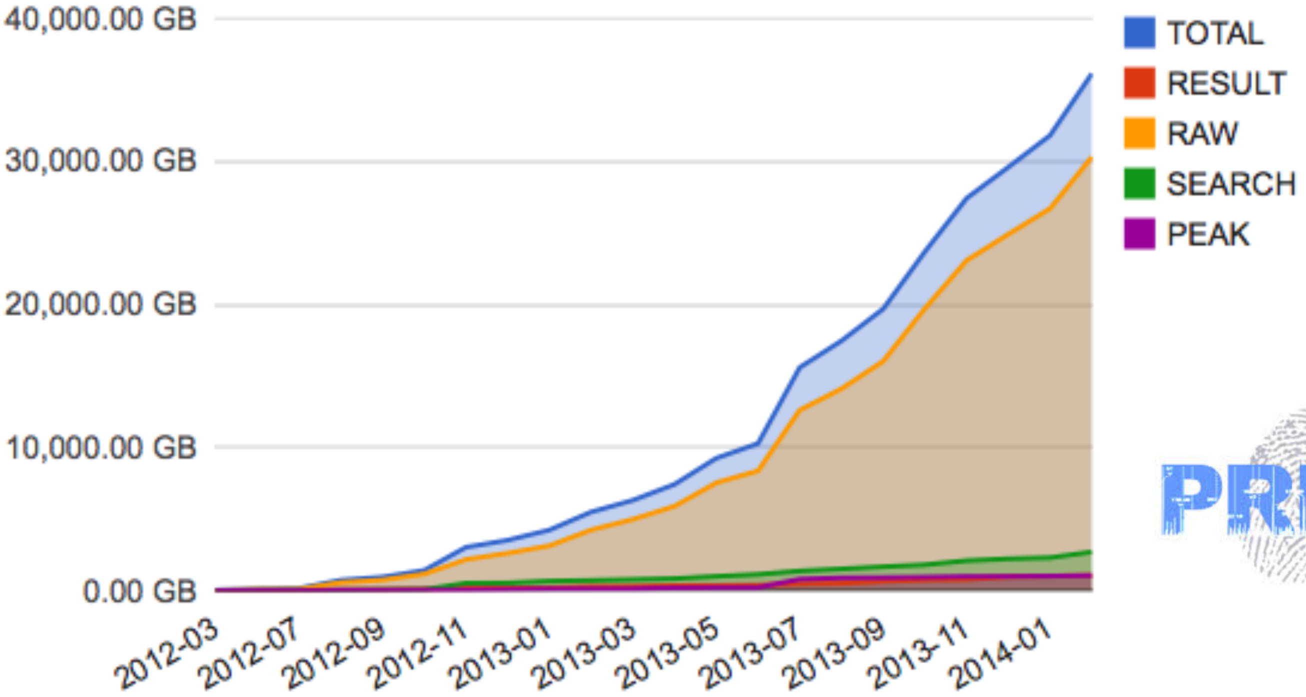


EMBL-EBI

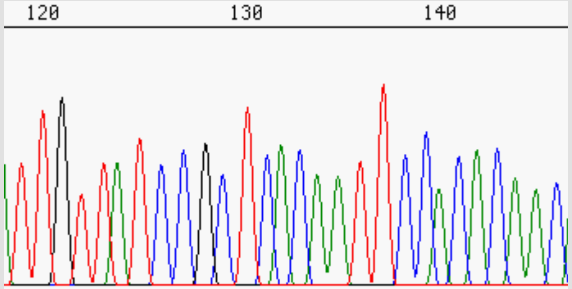
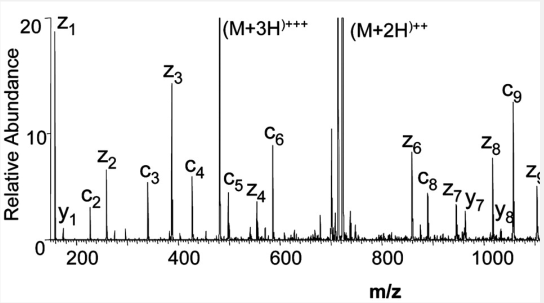
elixir

Proteomics data in PRIDE

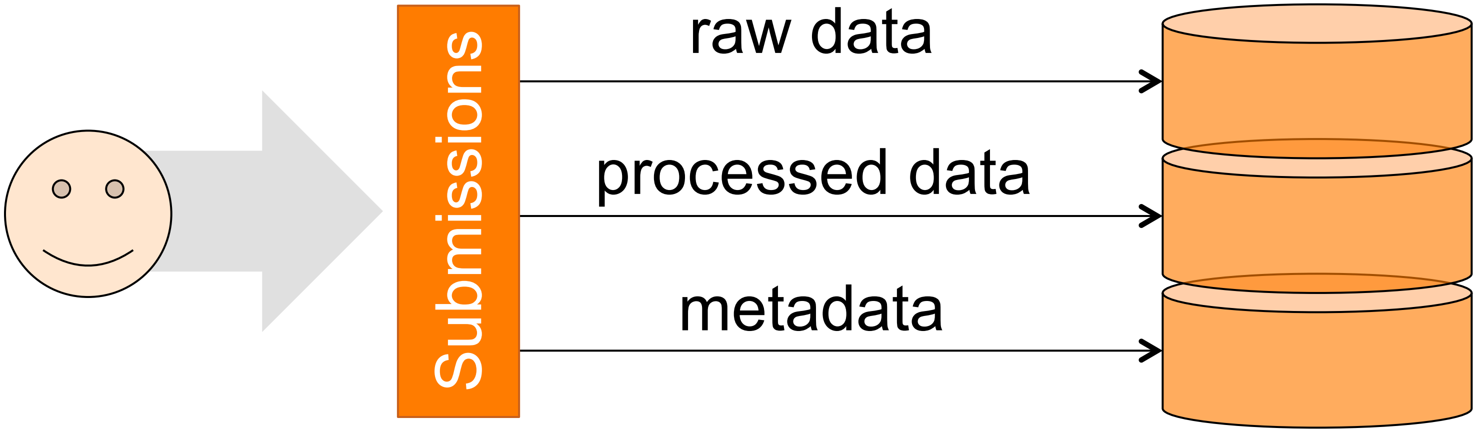
~85% raw data



Data types examples

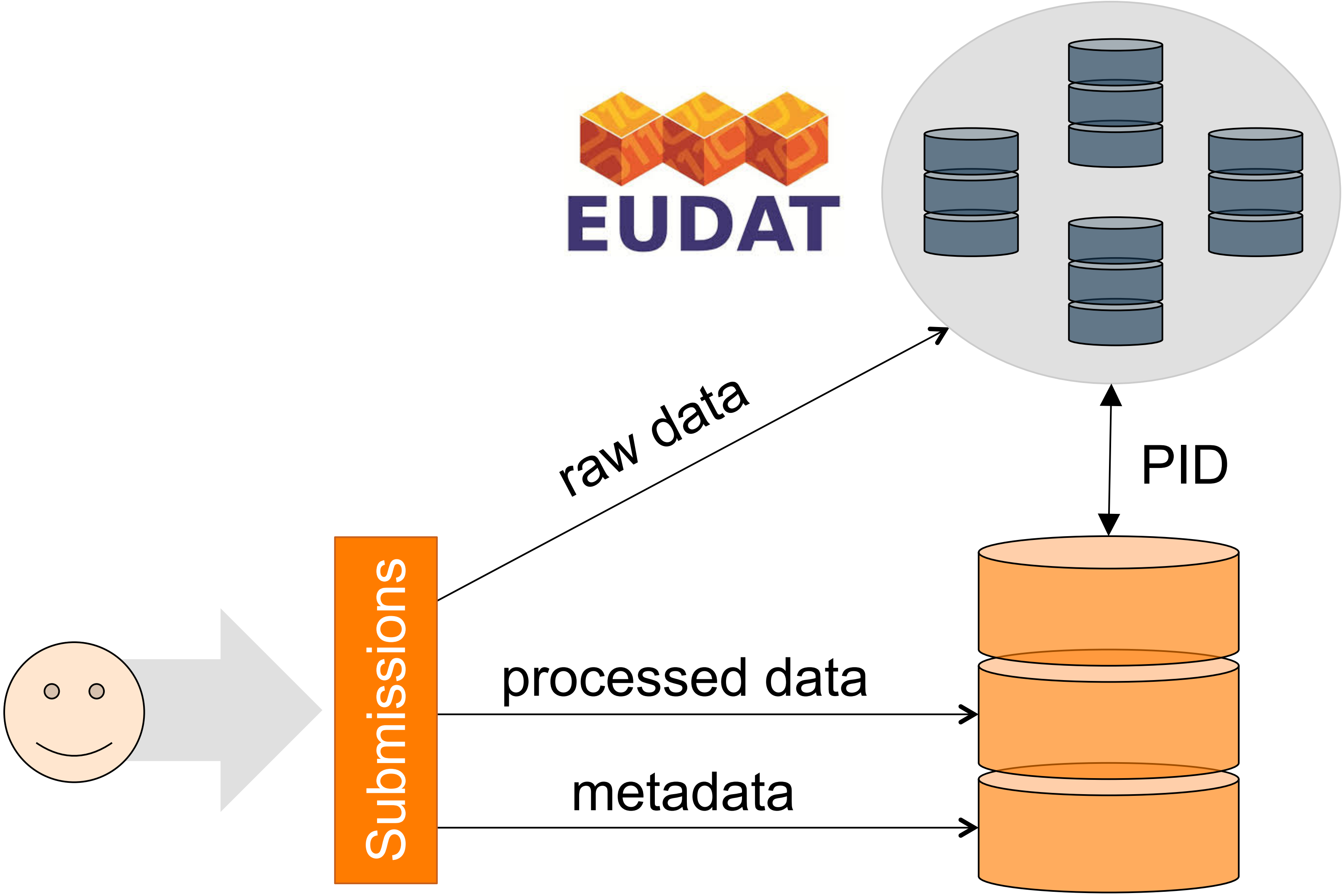
Raw data	Process data	Metadata
	TTGTTATCCG…	***DNA*** *Human* *Liver* *Mitochondria* *W. Smith* …
	LPISASHSSK…	***Peptide*** *Mouse* *Heart* *Nucleus* *J. Heinz* …
…	…	…

Data submission

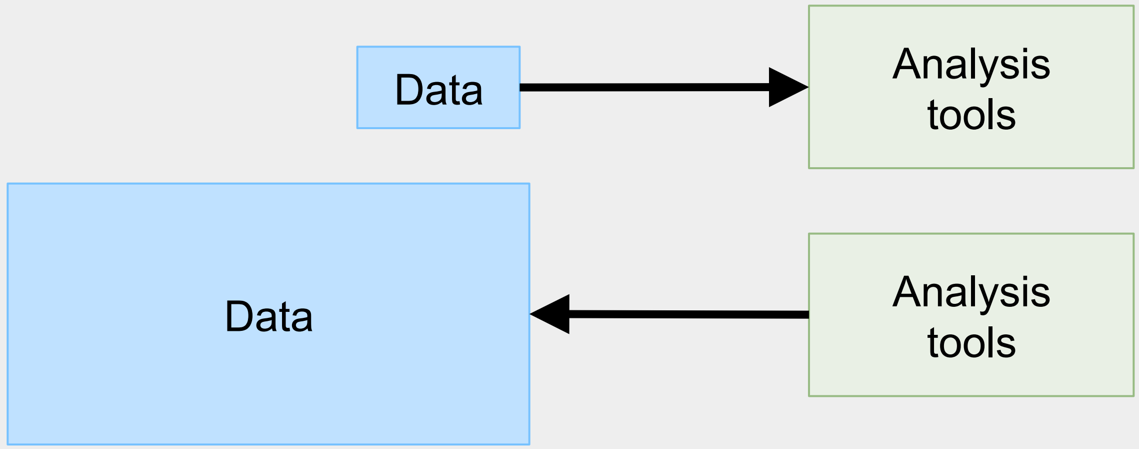


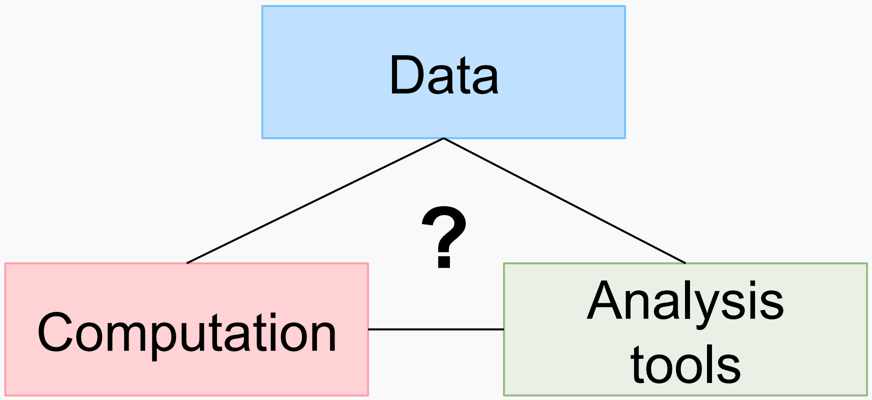
Centralized database

Data submission - pilot



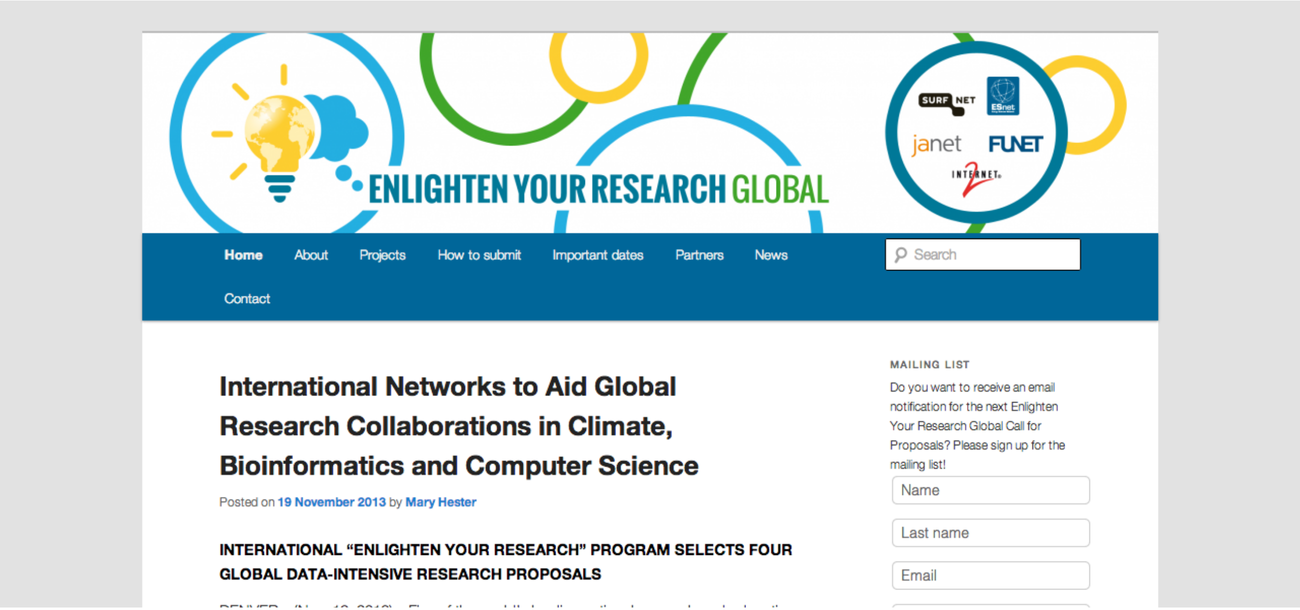
Data analysis





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 egi



International Networks to Aid Global Research Collaborations in Climate, Bioinformatics and Computer Science

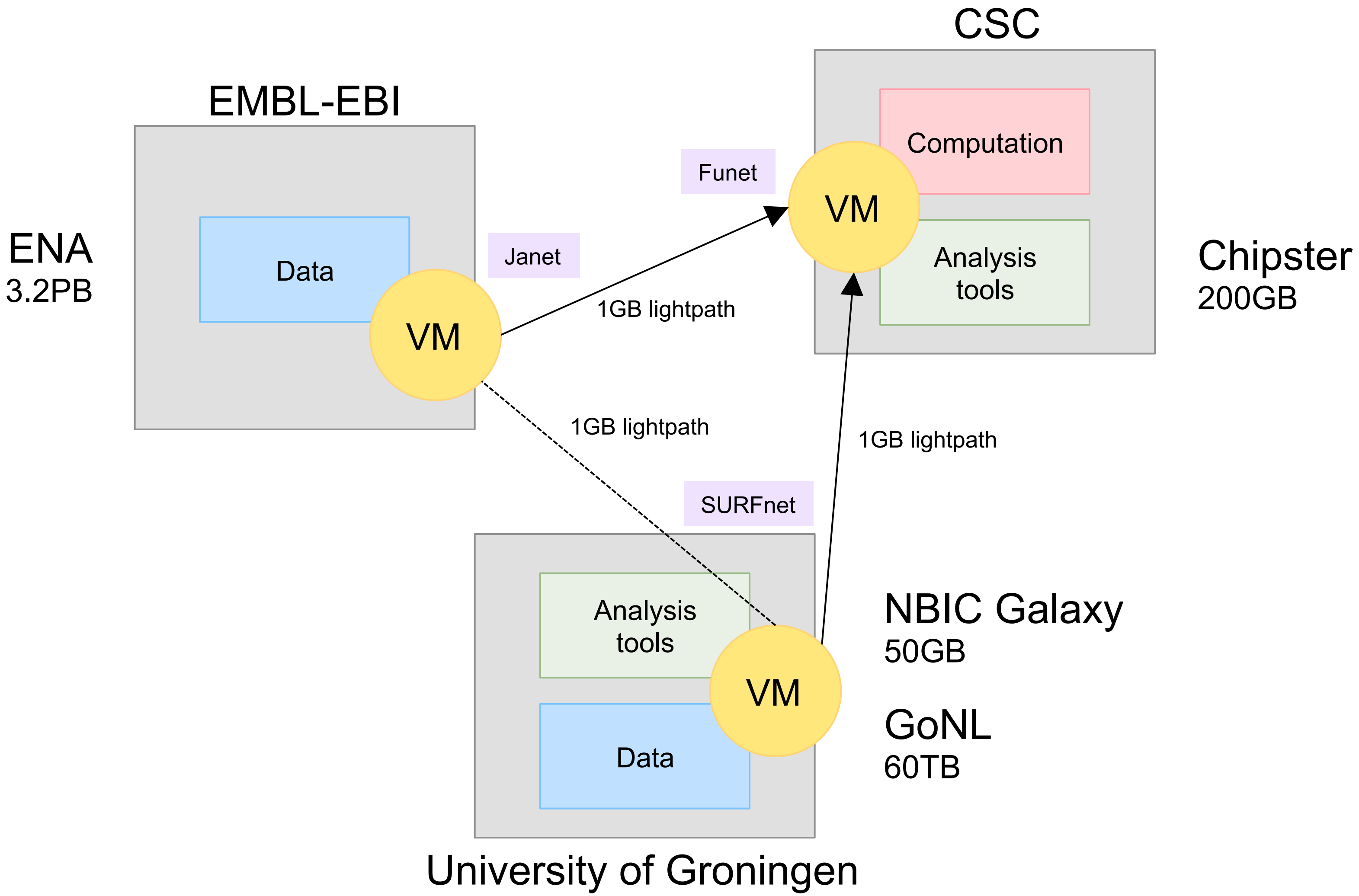
Posted on 19 November 2013 by Mary Hester

INTERNATIONAL "ENLIGHTEN YOUR RESEARCH" PROGRAM SELECTS FOUR GLOBAL DATA-INTENSIVE RESEARCH PROPOSALS

MAILING LIST

Do you want to receive an email notification for the next Enlighten Your Research Global Call for Proposals? Please sign up for the mailing list!

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

GÉANT

CONNECT

THE MAGAZINE FROM THE GÉANT COMMUNITY | ISSUE 13 OCTOBER 2013

DNA DATA DELUGE

GÉANT AND THE GLOBAL SHARING OF GENOMIC DATA

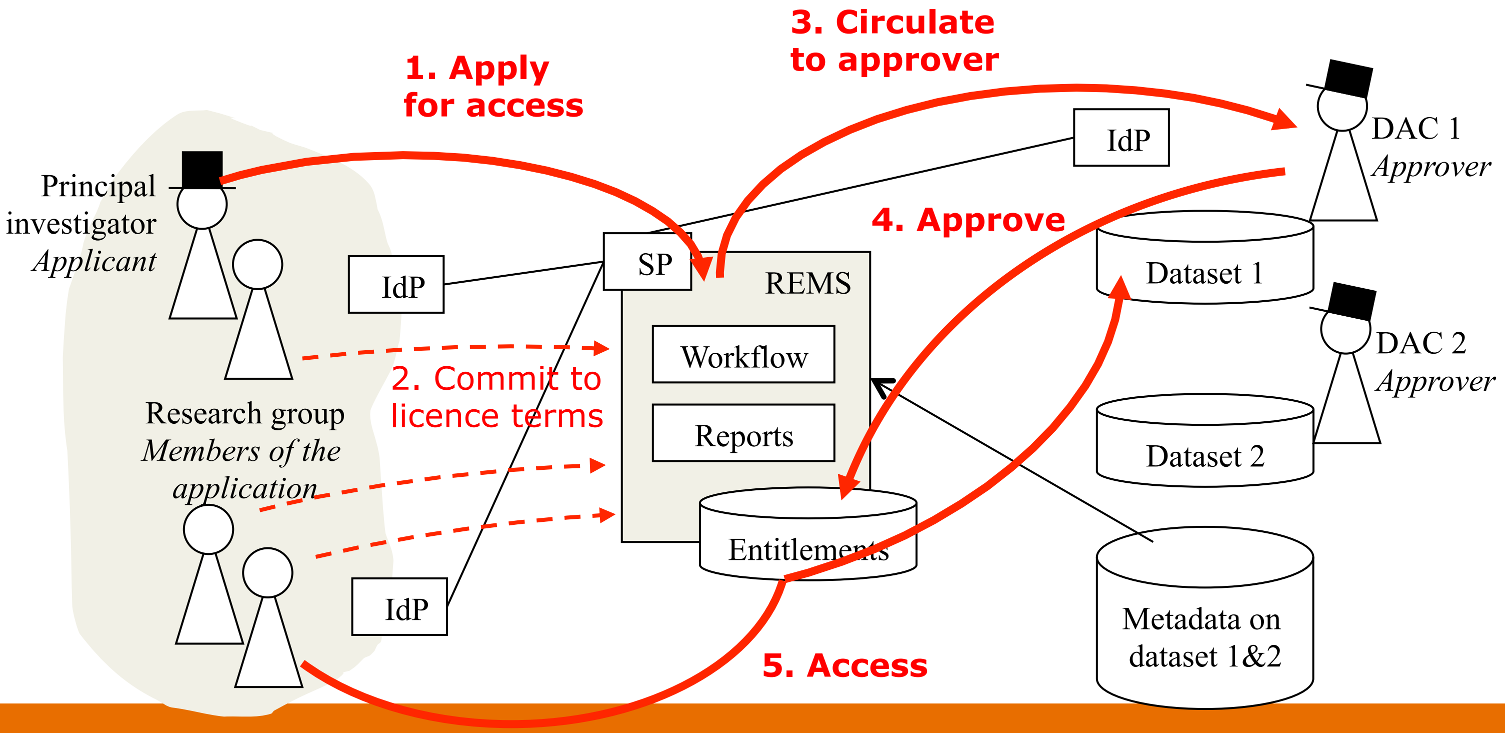
CREATING NETWORKS: BEHIND THE SCENES AT GÉANT'S NETWORK MIGRATION

BUILDING COMMUNITIES: VIA NETWORK PERFORMANCE AND MONITORING

SUPPORTING NRENS: CELEBRATING 20 YEARS OF EENET, THE INTERNET AND INDEPENDENCE

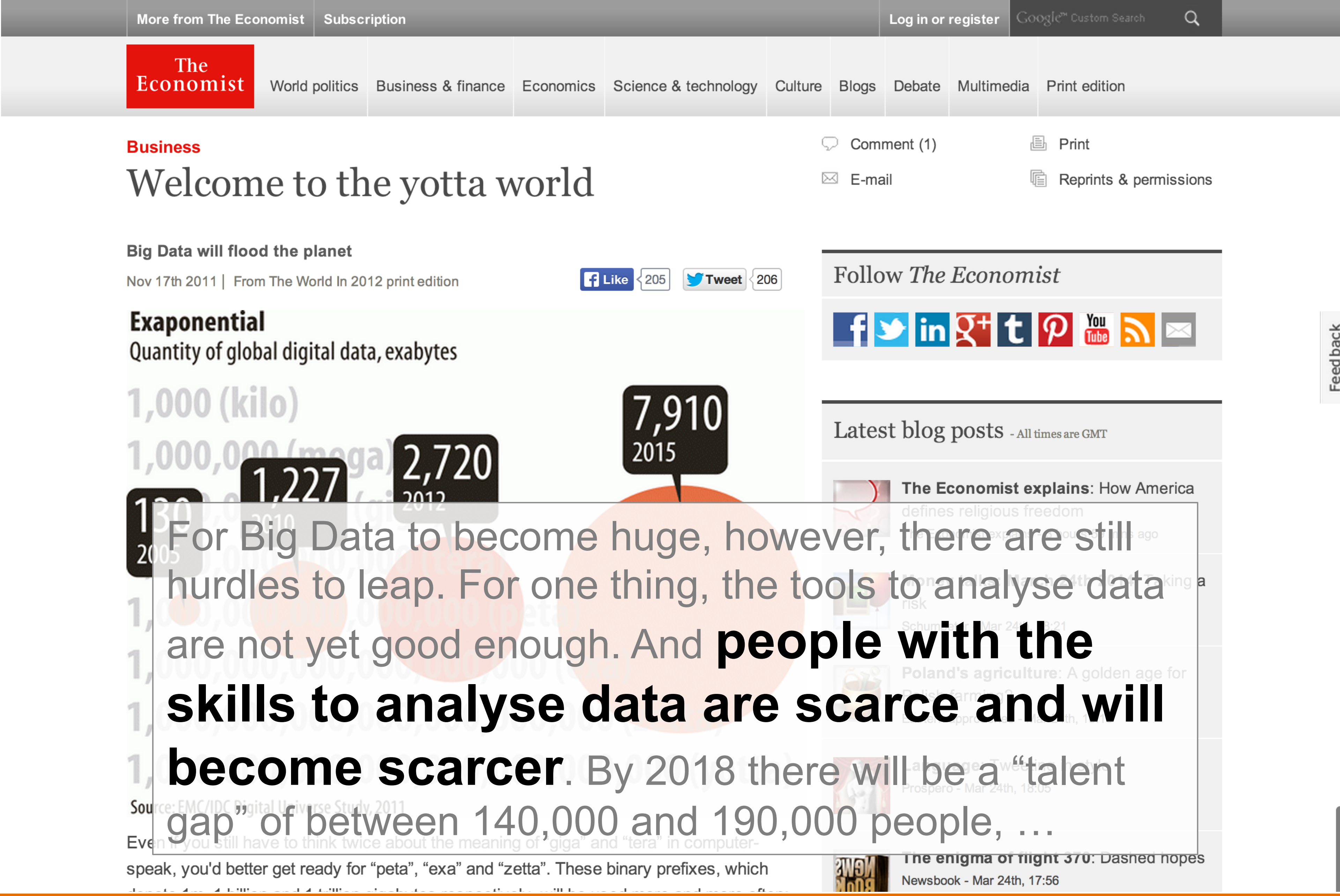
REMS - Resource Entitlement Management System

- 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/



More from The Economist | Subscription | Log in or register | Google™ Custom Search

The Economist | World politics | Business & finance | Economics | Science & technology | Culture | Blogs | Debate | Multimedia | Print edition

Business

Welcome to the yotta world

Comment (1) | Print | E-mail | Reprints & permissions

Big Data will flood the planet

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

Like 205 | Tweet 206

Exaponential

Quantity of global digital data, exabytes

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

Follow *The Economist*

Latest blog posts - All times are GMT

The Economist explains: How America defines religious freedom

The enigma of flight 370: Dashed hopes
Newsbook - Mar 24th, 17:56

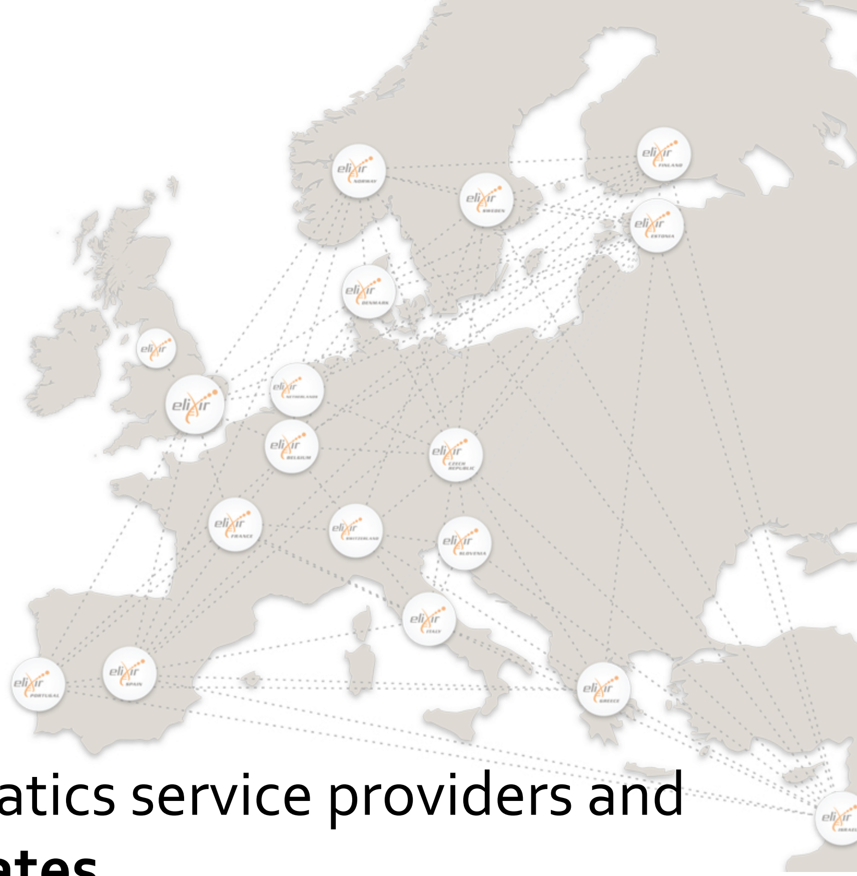
For Big Data to become huge, however, there are still hurdles to leap. For one thing, the tools to analyse data are not yet good enough. And **people with the skills to analyse data are scarce and will become scarcer**. By 2018 there will be a "talent gap" of between 140,000 and 190,000 people, …

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

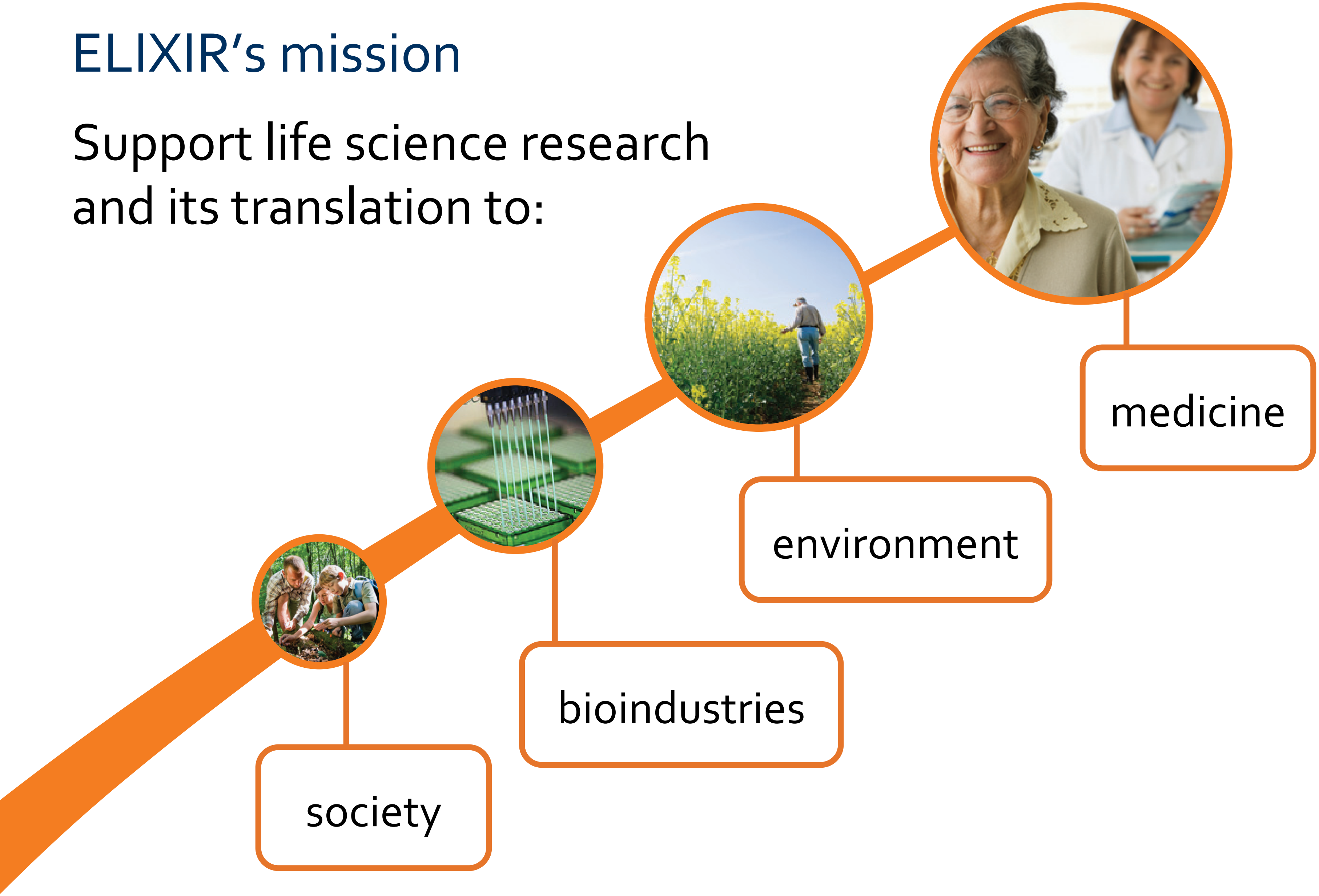
- European **life sciences** research infrastructure for **biological information** to **facilitate research**
- **Safeguard data** and build **sustainable data services**



- Participated by major bioinformatics service providers and supported by **17 EU member states**
- Creating a robust infrastructure for biological information is a **bigger task than any individual organisation** or nation can take on alone

ELIXIR’s mission

Support life science research and its translation to:



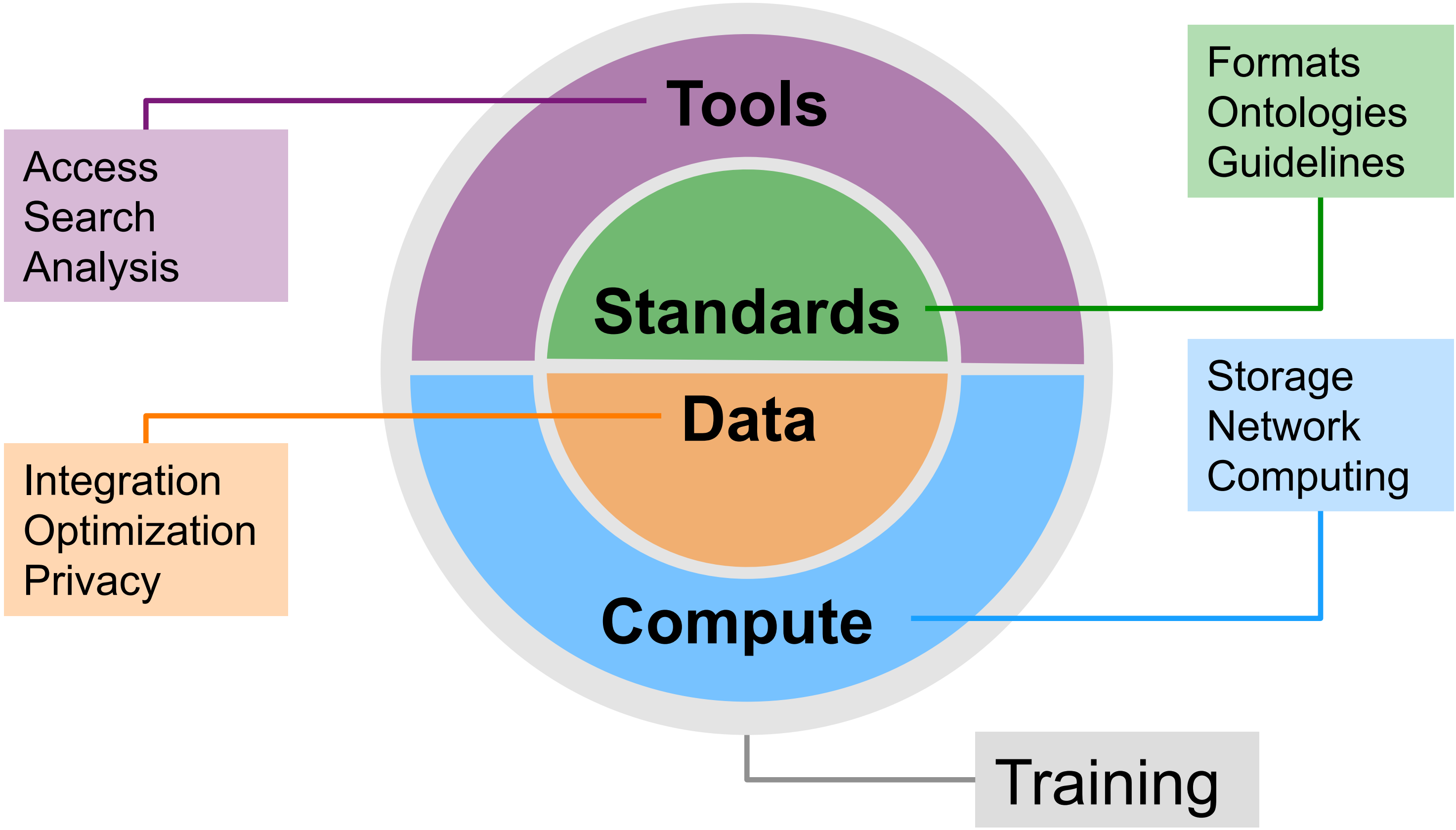
society

bioindustries

environment

medicine

Infrastructure for Life Sciences



Access
Search
Analysis

Tools

Standards

Formats
Ontologies
Guidelines

Data

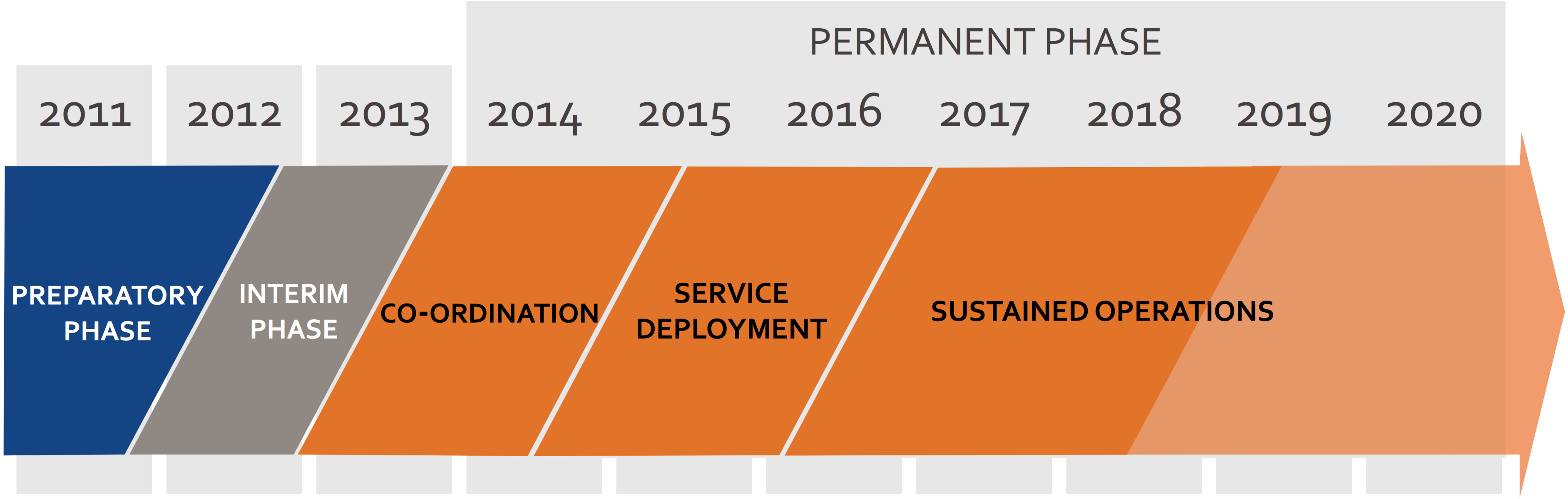
Integration
Optimization
Privacy

Compute

Storage
Network
Computing

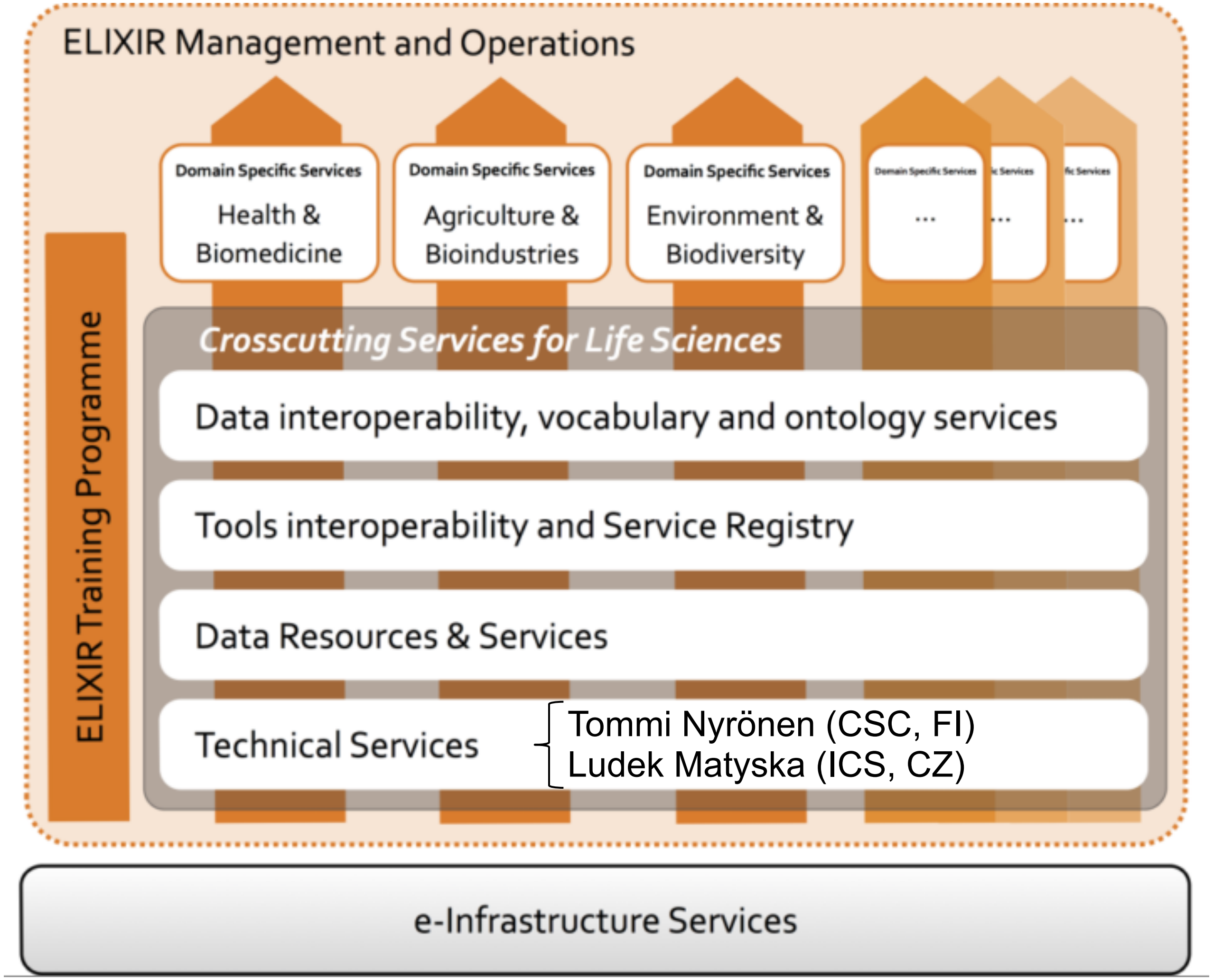
Training

Acceleration towards sustained operations



- Key objectives:
  - In 2014: Build on success of Interim Phase, formal establishment of Nodes
  - In 2015: Start delivering ELIXIR services
  - From 2016 onwards: sustained operations across ELIXIR

Work streams



ELIXIR Management and Operations

ELIXIR Training Programme

Domain Specific Services
Health & Biomedicine

Domain Specific Services
Agriculture & Bioindustries

Domain Specific Services
Environment & Biodiversity

Domain Specific Services
…

…

…

Crosscutting Services for Life Sciences

Data interoperability, vocabulary and ontology services

Tools interoperability and Service Registry

Data Resources & Services

Technical Services

Tommi Nyrönen (CSC, FI)
Ludek Matyska (ICS, CZ)

e-Infrastructure Services

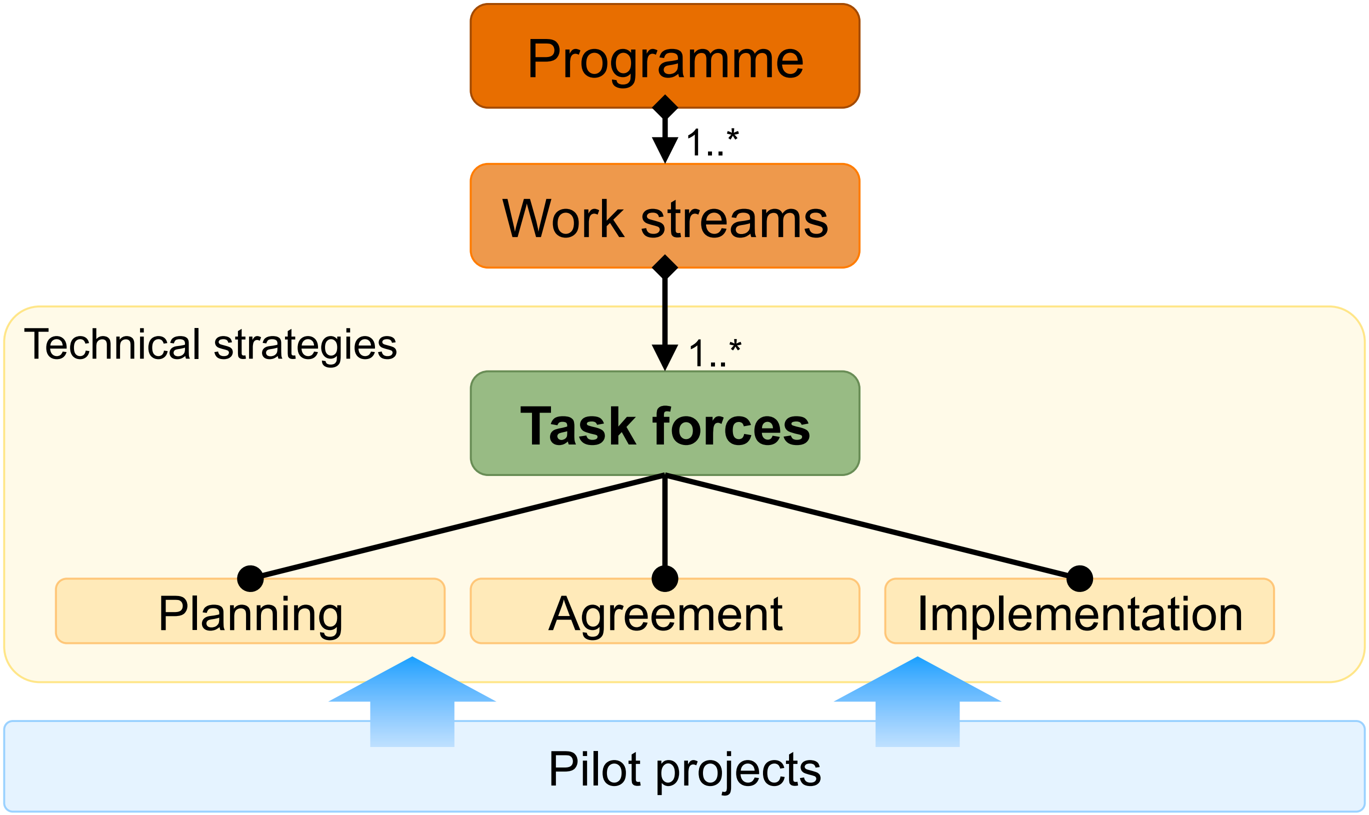
TCG - Technical Coordinators Group

- Coordinate the ELIXIR technical strategy
- Composed of technical representatives from each node

Belgium	Lieven Sterck	**Norway**	Kjell Petersen
Czech Republic	David Antos and Jan Paces	**Portugal**	TBC
Denmark	Kristoffer Rapacki	**Slovenia**	Brane Leskosek
Estonia	Hedi Peterson	**Spain**	Victor de la Torre
Finland	Jarno Laitinen and Olli Tourunen	**Sweden**	Mikael Borg
France	Christophe Blanchet	**Switzerland**	Heinz Stockinger
Greece	TBC	**UK**	Mario Caccamo and Manuel Corpas
Israel	Jaime Prilusky	**EMBL**	Steven Newhouse
Italy	Federico Zambelli	**ELIXIR Hub**	Rafael C Jimenez
Netherlands	Rob Hooft		

Task forces

- Plan, agree and implement technical strategies
- Represent ELIXIR on one specific technical topic

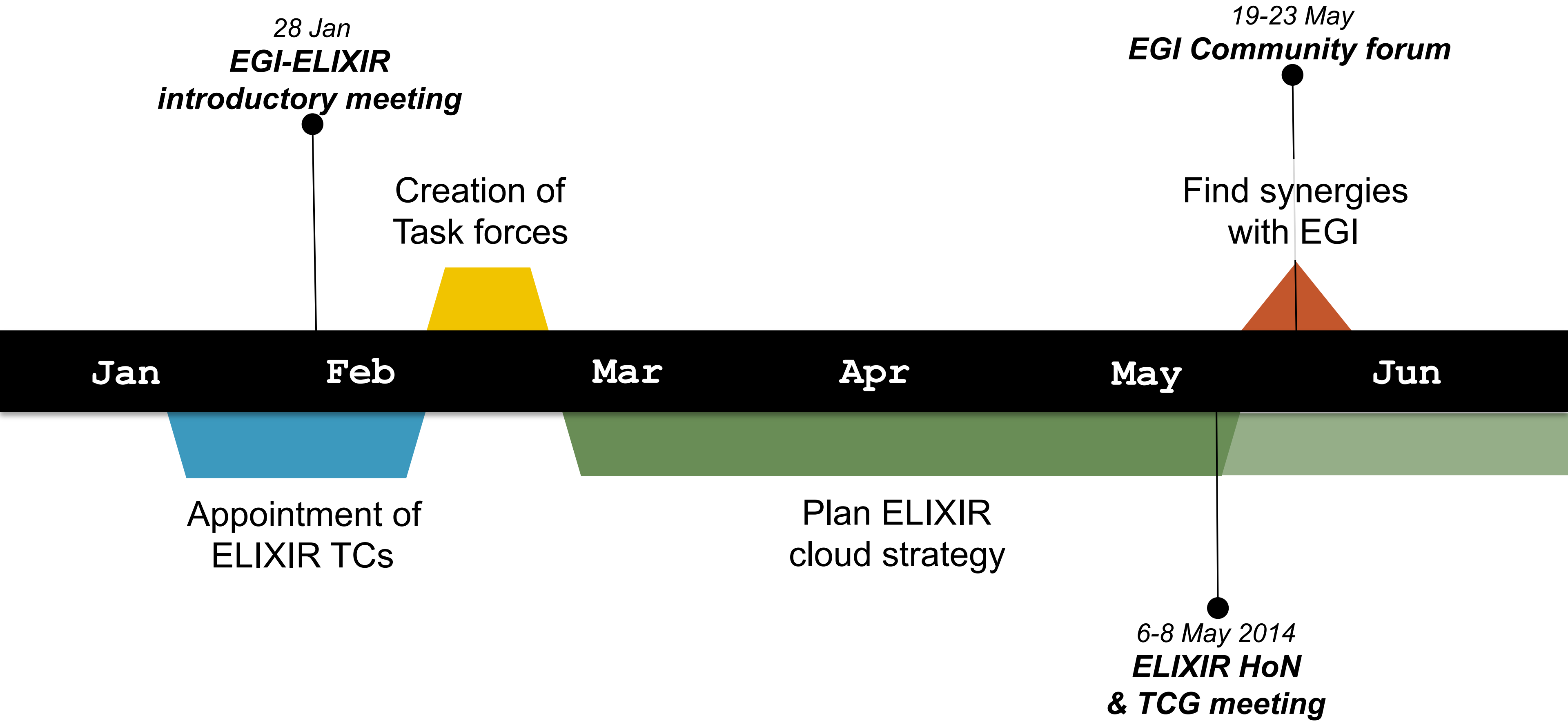


Prioritized task forces

interface with

- **Cloud** → egi
- Service registry
- Communication
- Metrics, monitoring & quality control
- Website
- **Storage** → EUDAT
- **Authentication and authorisation** → GÉANT
- Training portal
- e-Learning and Training

ELIXIR Cloud task force timeline



28 Jan
EGI-ELIXIR introductory meeting

Creation of Task forces

19-23 May
EGI Community forum

Find synergies with EGI

Jan Feb Mar Apr May Jun

Appointment of ELIXIR TCs

Plan ELIXIR cloud strategy

6-8 May 2014
ELIXIR HoN & TCG meeting

Thank you for your attention

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

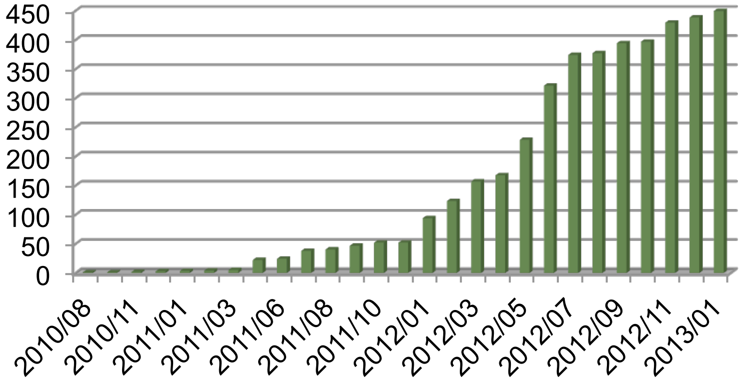
ELIXIR Pilot Projects

- Five short-term Pilot Actions are underway to act as **test beds for integration of ELIXIR services**:

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

European Genome-phenome archive

EGA data growth

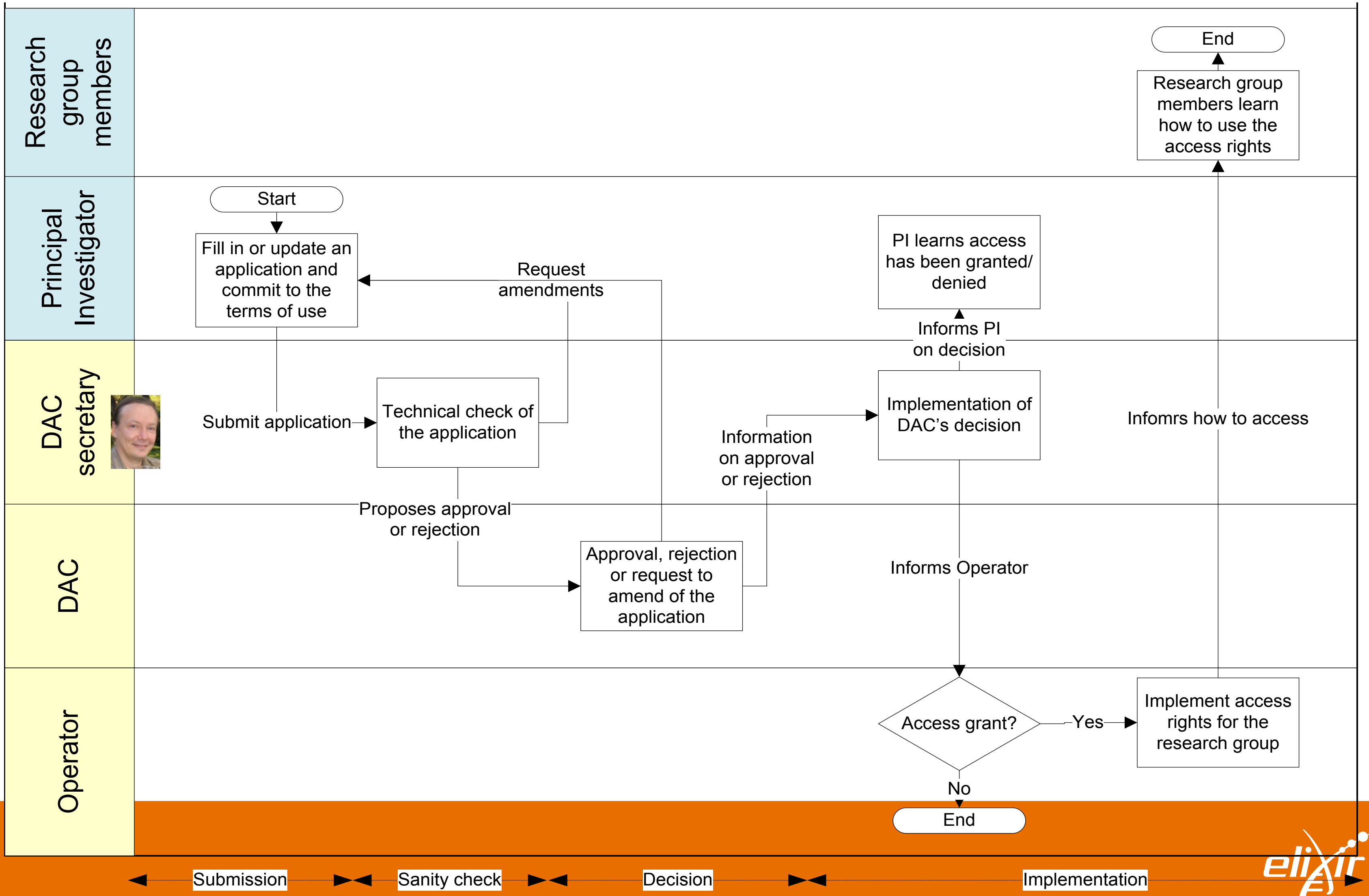


- EGA is to be **distributed effort** with **archive**, **submission**, and **data distribution** capacity at both the EBI and CRG

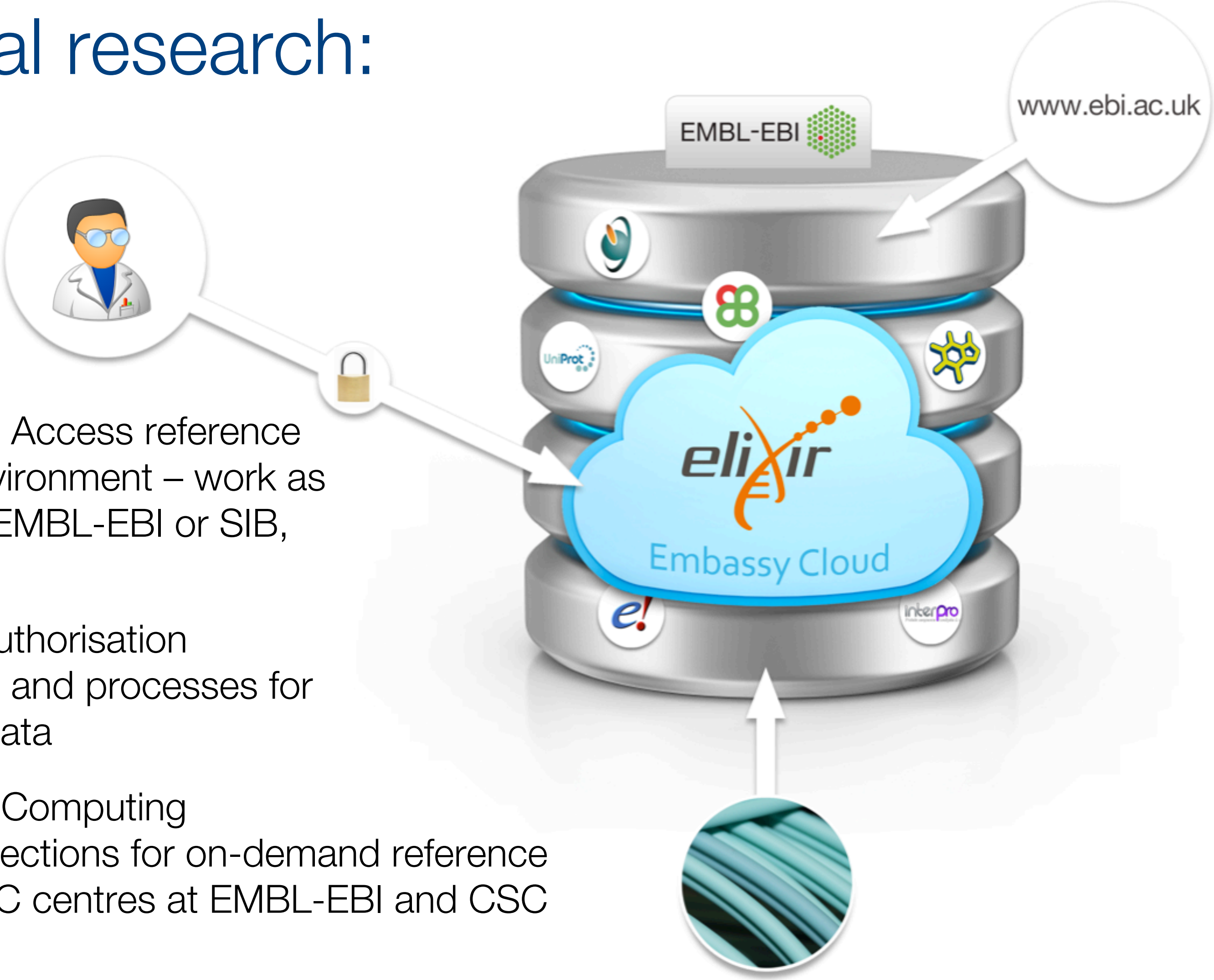
- From the users point of view, EGA remains one **integrated** Archive of **secure human biomedical research data**.

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

CASE: process for applying access to the Nordic Control Database


Research group members
Principal Investigator
DAC secretary
DAC
Operator
End
Research group members learn how to use the access rights
Start
Fill in or update an application and commit to the terms of use
Request amendments
PI learns access has been granted/ denied
Informs PI on decision
Submit application
Technical check of the application
Information on approval or rejection
Implementation of DAC's decision
Infomrs how to access
Proposes approval or rejection
Approval, rejection or request to amend of the application
Informs Operator
Access grant?
Yes
Implement access rights for the research group
No
End
Submission
Sanity check
Decision
Implementation

ELIXIR pilots to address key challenges in biomedical research:

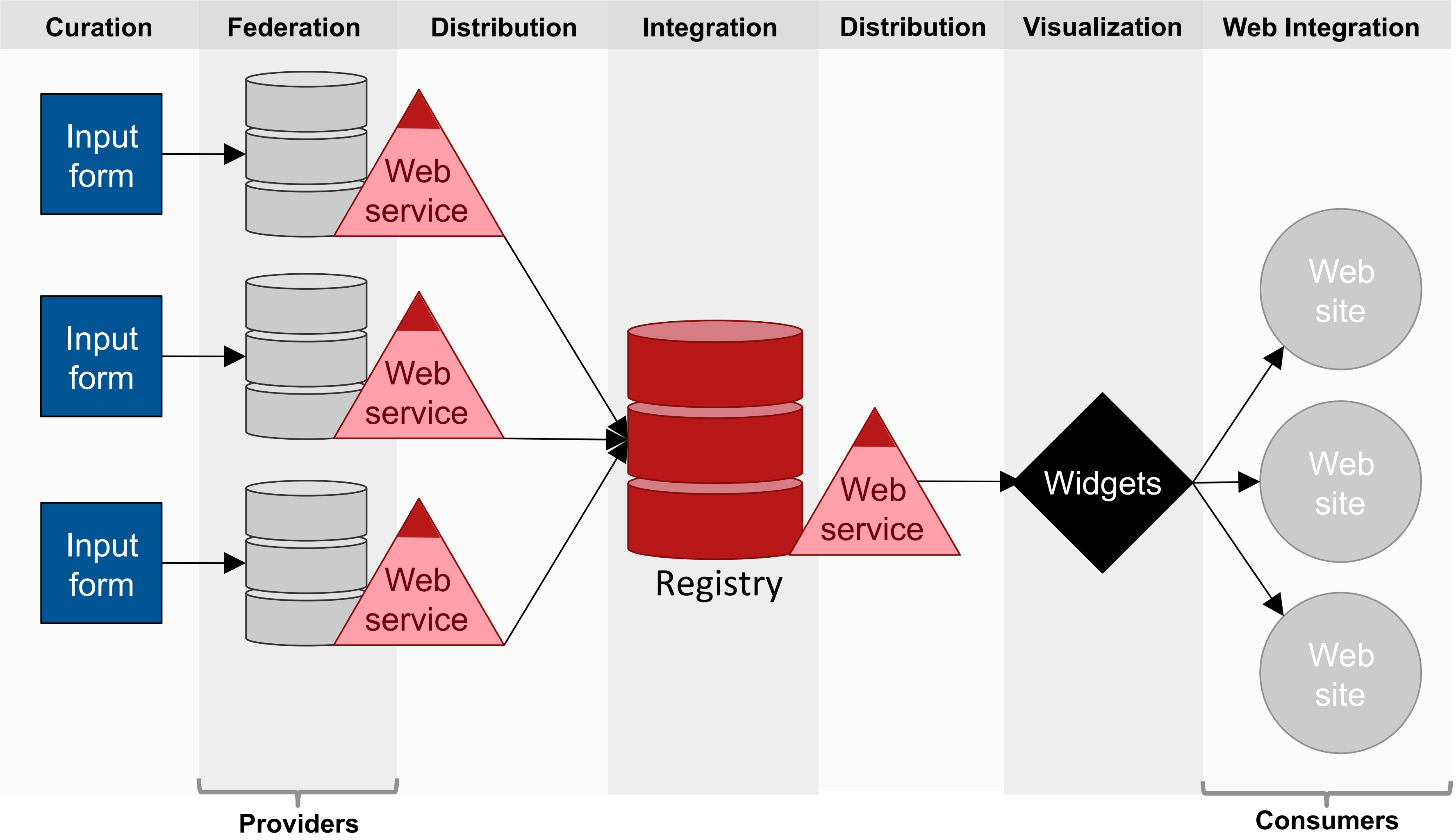


1. Cloud computing
   "Embassy cloud": Access reference data in a virtual environment – work as though you are at EMBL-EBI or SIB, Switzerland

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

Registry of Services and databases



Curation
Federation
Distribution
Integration
Distribution
Visualization
Web Integration

Input form
Input form
Input form

Web service
Web service
Web service

Registry

Web service

Widgets

Web site
Web site
Web site

Providers

Consumers

🟦 Form used by providers to collect data

▤ Data provider

▤ Centralized repository integrating data

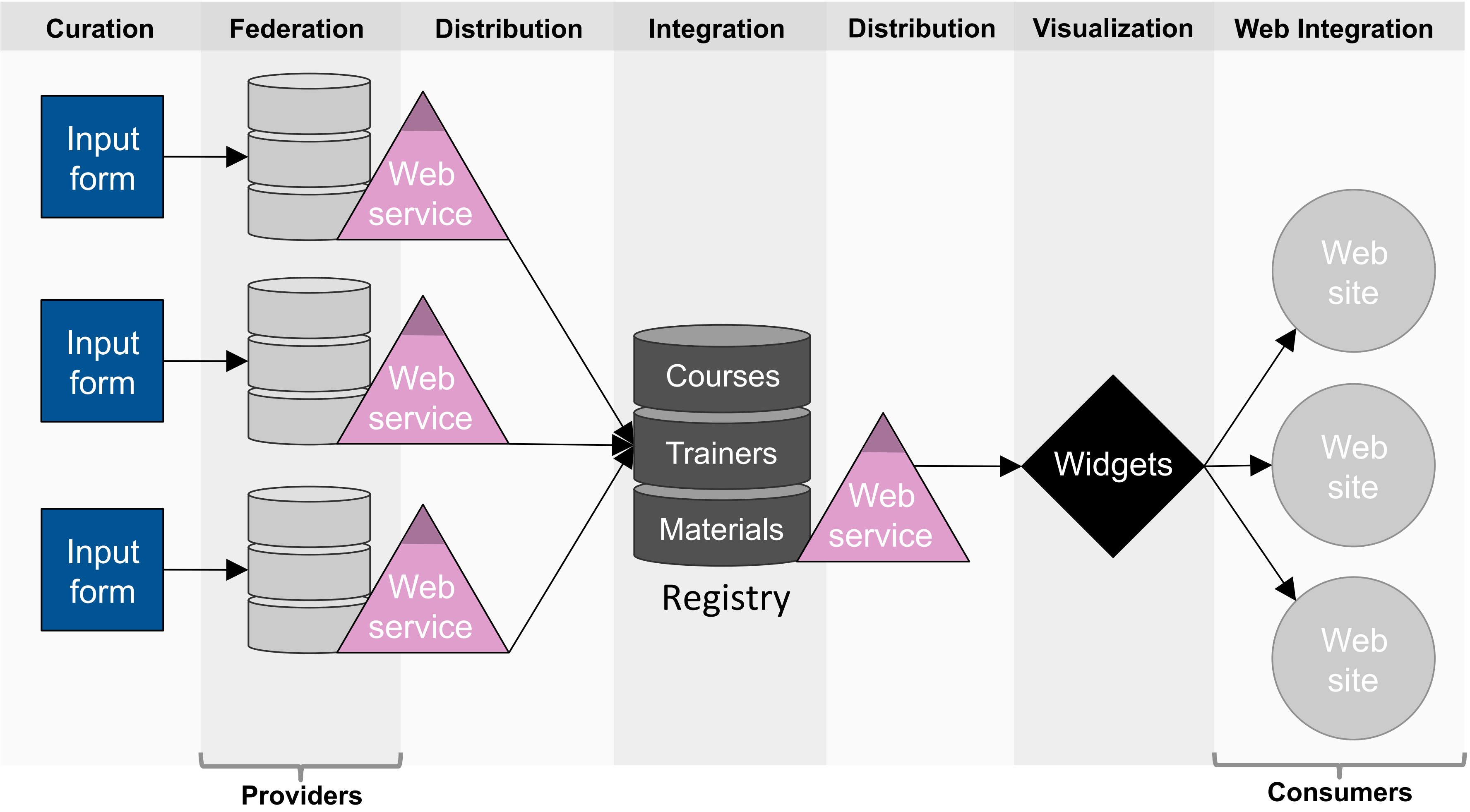
△ Common query interface (web service) based on an agreed standard

◆ Widgets to present information

● Websites interested to integrate information from the centralized

TeSS

Registry of courses, trainers and materials



■ Form used by providers to collect data

⛁ Data provider

⛁ Centralized repository integrating data

▲ Common query interface (web service) based on an agreed standard

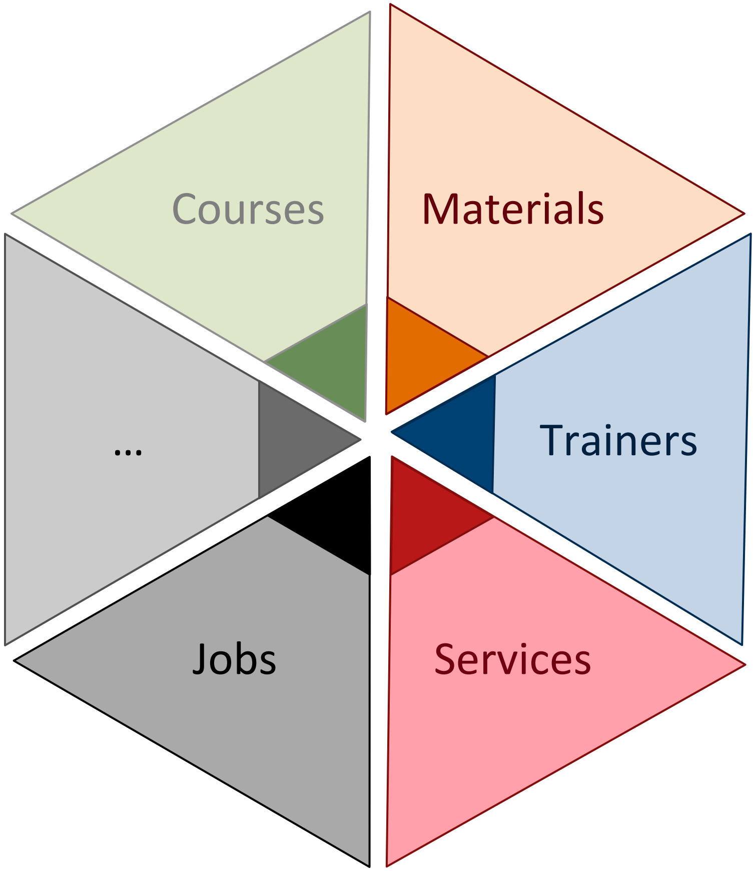
◆ Widgets to present information

● Websites interested to integrate information from the centralized

Minimum metadata

life sciences

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


Courses
Materials
…
Trainers
Jobs
Services