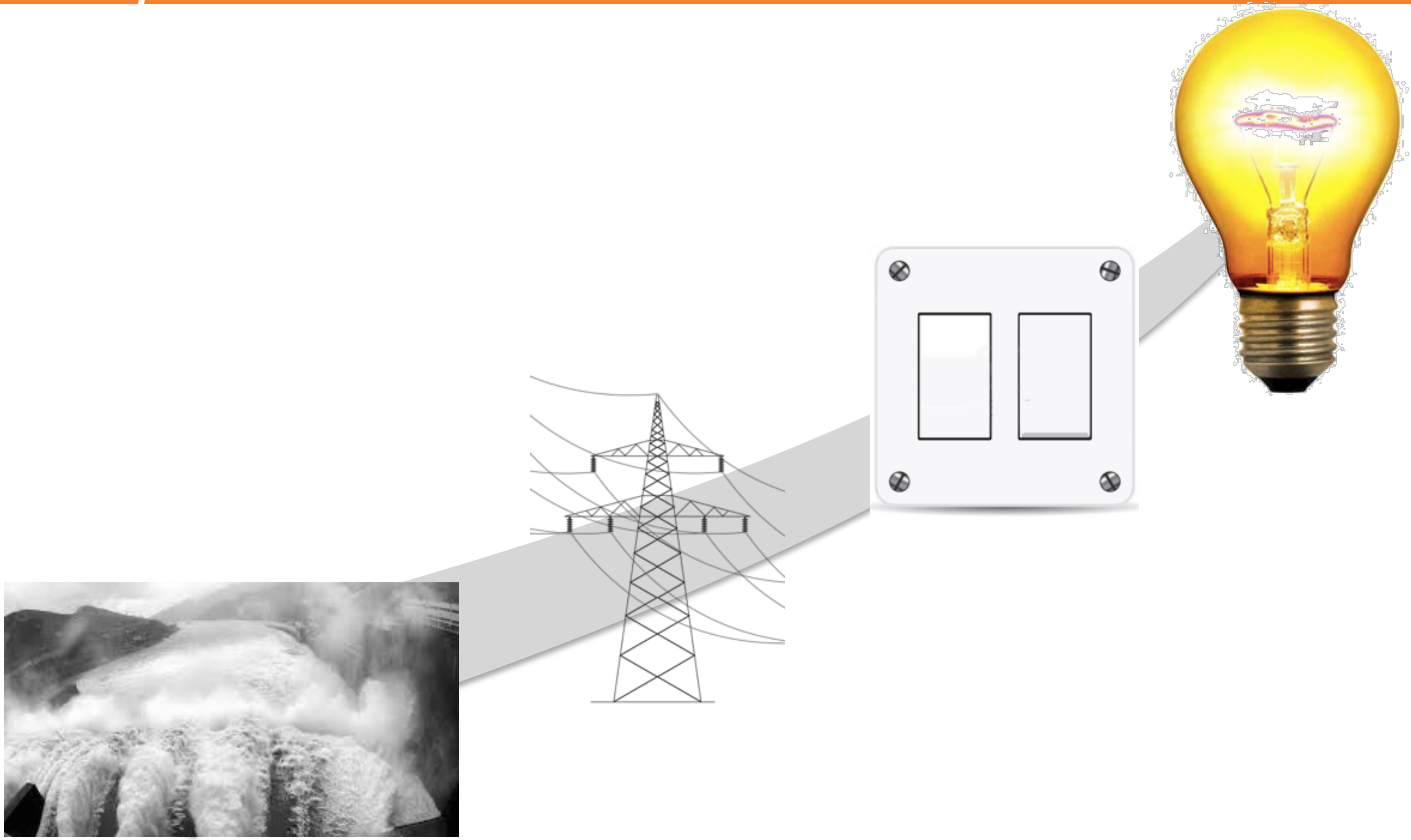
European life-science data infrastructure: Data, Computing and Services to Communities

Niklas Blomberg

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

What appears to be a simple, reliable user experience...



...is made possible by robust, non-trivial infrastructure that often goes unnoticed.

Biomedical research is requiring increasingly sophisticated infrastructure.

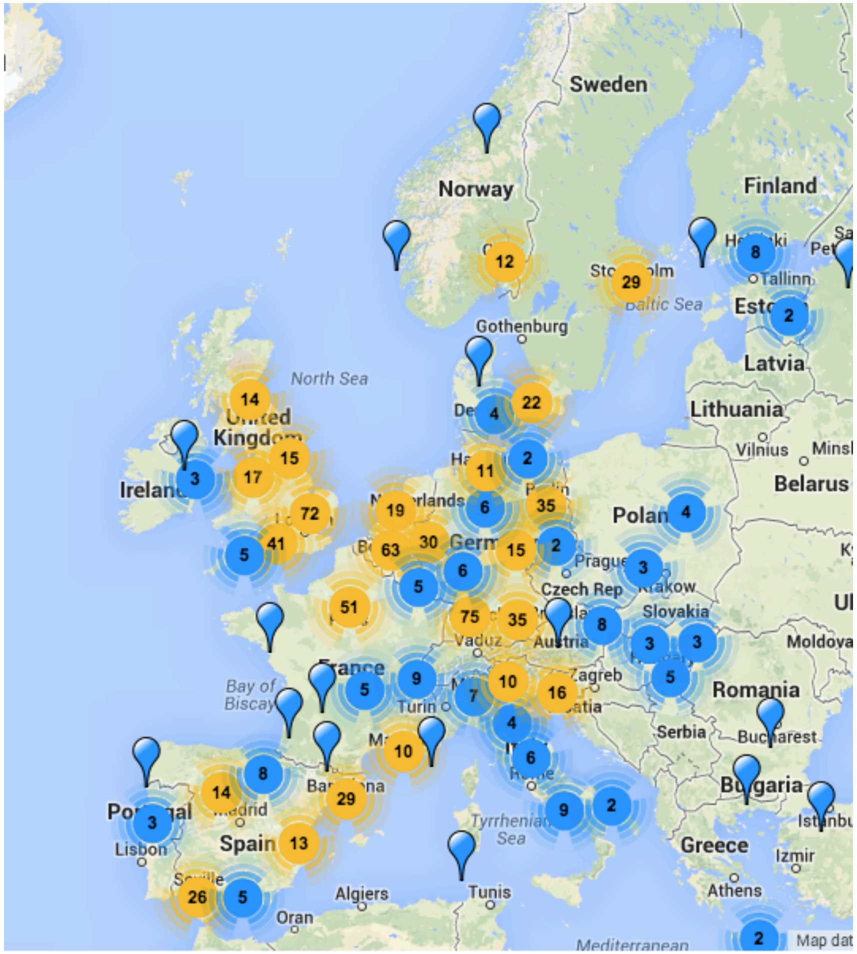
No formal infrastructure
(single use)

Basic infrastructure
(local)

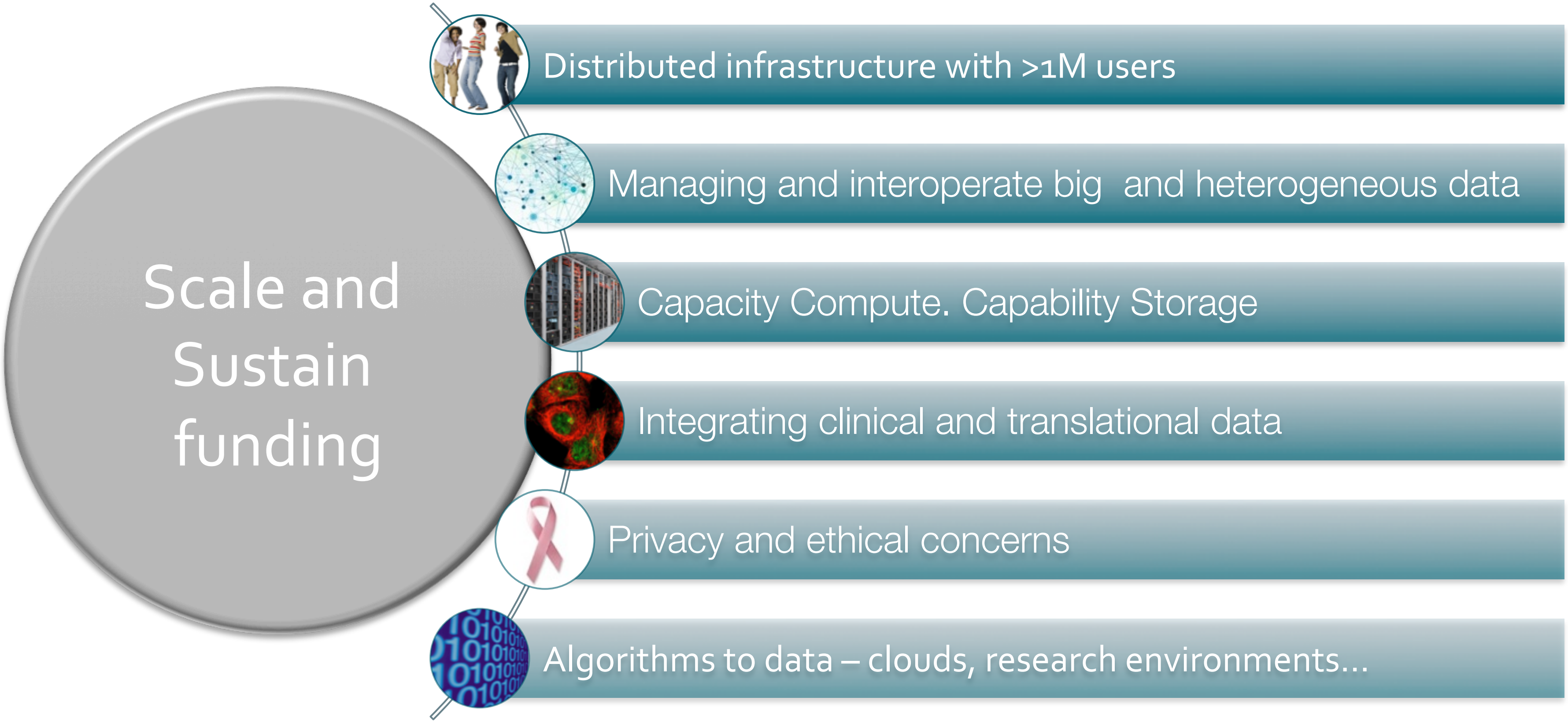
Robust infrastructure
(large-scale, non-trivial,
interconnected)

Life-science and data infrastructure

- Data production and using at a large number of sites across Europe
  - (European Illumina sales up 20% 2013)
- Human genomics projects but also plants, microbiota, environmental marker organisms
- Metabolomics & Proteomics coming of age
  - UK National Phenome facility
- Be scalable to 1000s of sites
- Deal with incomplete, conflicting, and incorrect data



Challenges for life-science data services



Scale and Sustain funding

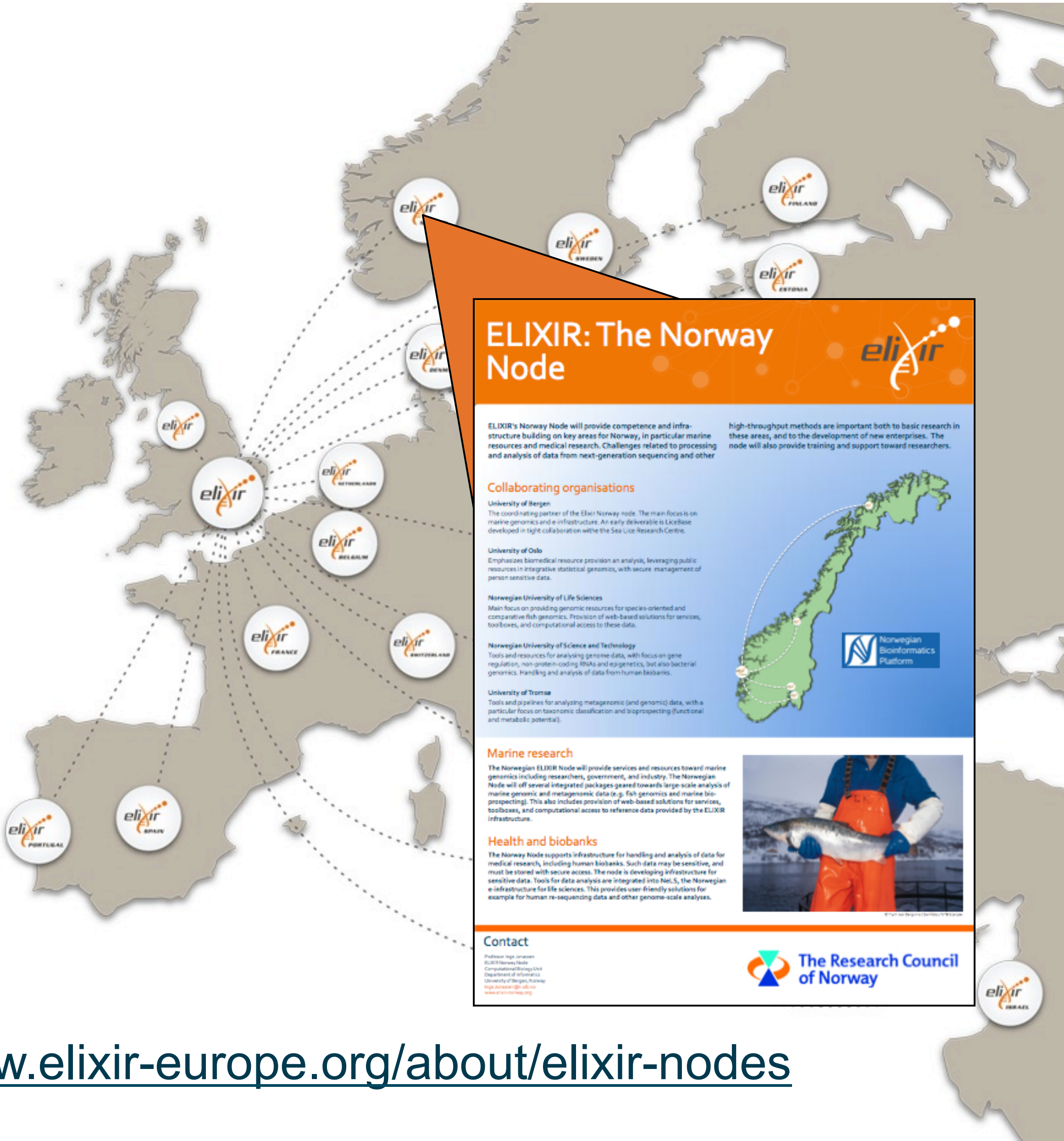
- Distributed infrastructure with >1M users
- Managing and interoperate big and heterogeneous data
- Capacity Compute. Capability Storage
- Integrating clinical and translational data
- Privacy and ethical concerns
- Algorithms to data – clouds, research environments…

A distributed infrastructure to scale with the challenges

- ELIXIR deliver services through national ELIXIR Nodes

- ELIXIR Nodes build local bioinformatics capacity throughout Europe

- ELIXIR Nodes build on national strengths and priorities



ELIXIR: The Norway Node

ELIXIR's Norway Node will provide competence and infrastructure building on key areas for Norway, in particular marine resources and medical research. Challenges related to processing and analysis of data from next-generation sequencing and other high-throughput methods are important both to basic research in these areas, and to the development of new enterprises. The node will also provide training and support toward researchers.

Collaborating organisations

University of Bergen
The coordinating partner of the Elixir Norway node. The main focus is on marine genomics and e-infrastructure. An early deliverable is LiceBase developed in tight collaboration withe the Sea Lice Research Centre.

University of Oslo
Emphasizes biomedical resource provision an analysis, leveraging public resources in integrative statistical genomics, with secure management of person sensitive data.

Norwegian University of Life Sciences
Main focus on providing genomic resources for species-oriented and comparative fish genomics. Provision of web-based solutions for services, toolboxes, and computational access to these data.

Norwegian University of Science and Technology
Tools and resources for analysing genome data, with focus on gene regulation, non-protein-coding RNAs and epigenetics, but also bacterial genomics. Handling and analysis of data from human biobanks.

University of Tromsø
Tools and pipelines for analyzing metagenomic (and genomic) data, with a particular focus on taxonomic classification and bioprospecting (functional and metabolic potential).

Norwegian Bioinformatics Platform

Marine research

The Norwegian ELIXIR Node will provide services and resources toward marine genomics including researchers, government, and industry. The Norwegian Node will off several integrated packages geared towards large-scale analysis of marine genomic and metagenomic data (e.g. fish genomics and marine bioprospecting). This also includes provision of web-based solutions for services, toolboxes, and computational access to reference data provided by the ELIXIR infrastructure.

Health and biobanks

The Norway Node supports infrastructure for handling and analysis of data for medical research, including human biobanks. Such data may be sensitive, and must be stored with secure access. The node is developing infrastructure for sensitive data. Tools for data analysis are integrated into NeLS, the Norwegian e-infrastructure for life sciences. This provides user-friendly solutions for example for human re-sequencing data and other genome-scale analyses.

Contact

Professor Inge Jonassen
ELIXIR Norway Node
Computational Biology Unit
Department of Informatics
University of Bergen, Norway
inge.jonassen@ii.uib.no
www.elixir-norway.org

The Research Council of Norway

http://www.elixir-europe.org/about/elixir-nodes

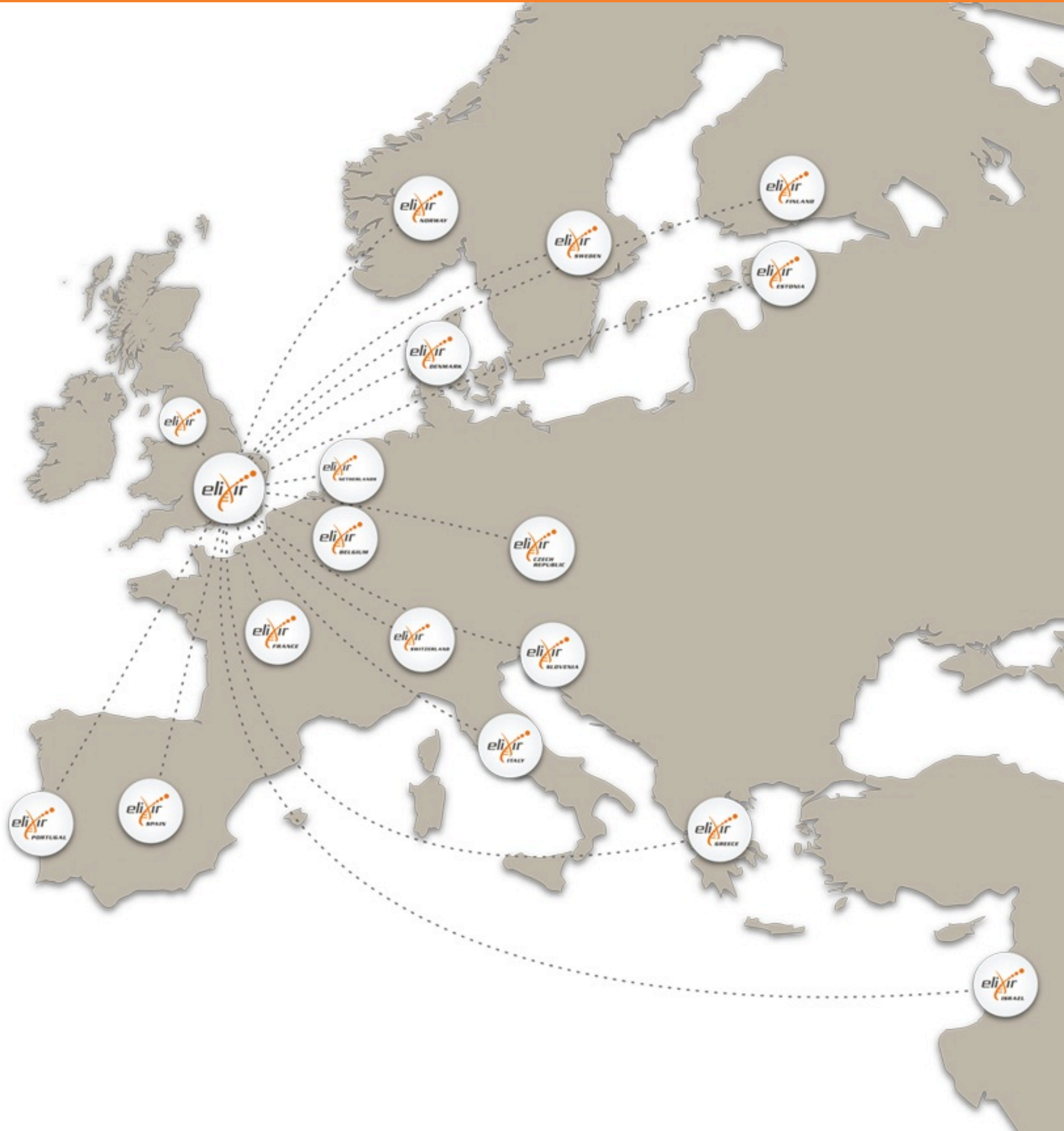
ELIXIR

- Elixir Consortium Agreement (ECA) entered into legal force Jan 2014

- 8 countries signed to date
  - Czech Republic, Denmark, Estonia, Netherlands, Norway, Sweden, Switzerland and the UK

- Further 9 countries have signed MoU and are working towards national signatures

- Discussions on-going with additional prospective member states



ELIXIR Infrastructure

- **Data**

  Sustain core data resources

- **Tools**

  Services & connectors to drive access and exploitation

- **Compute**

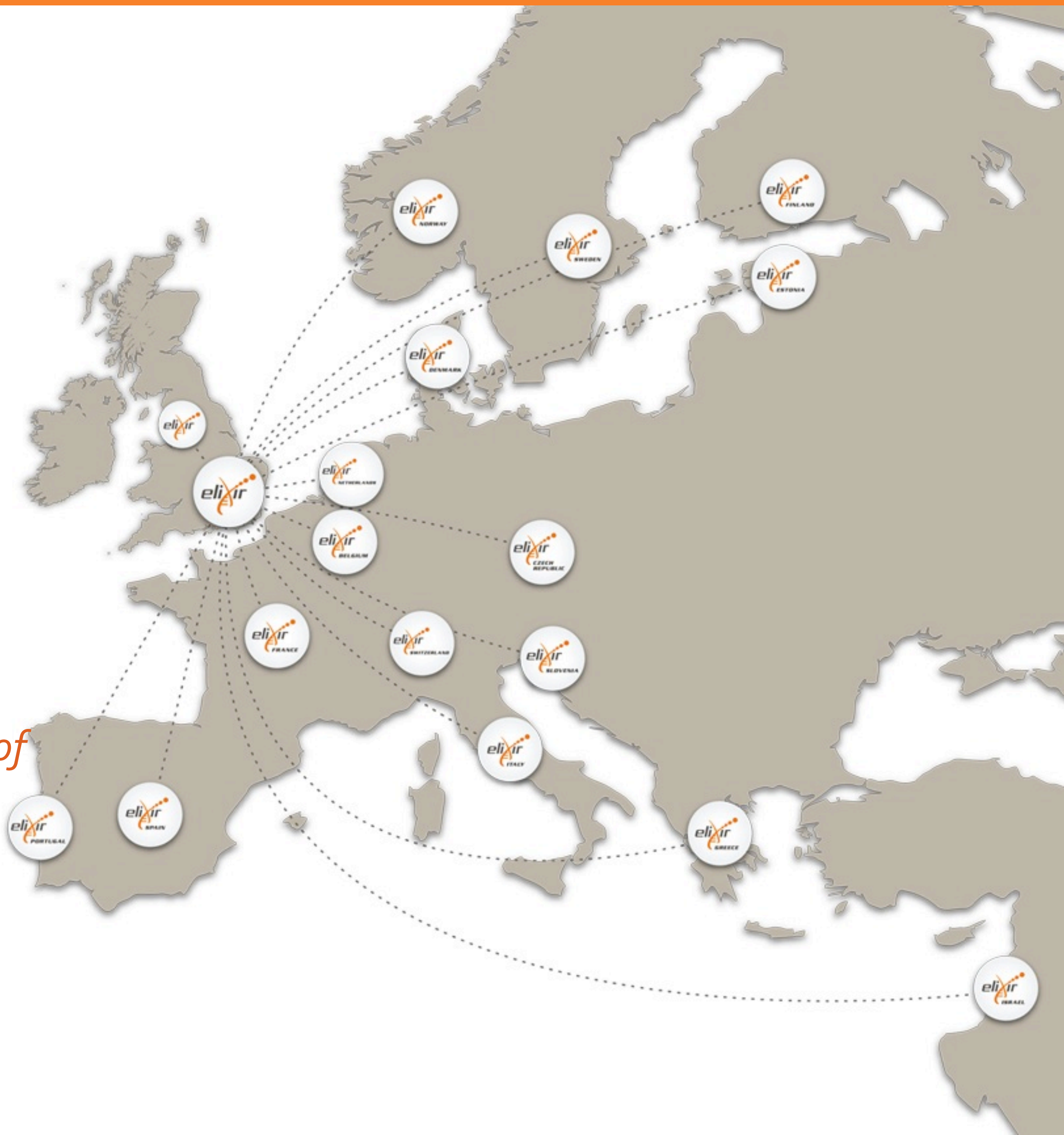
  Access, Exchange & Compute on sensitive data

- **Standards**

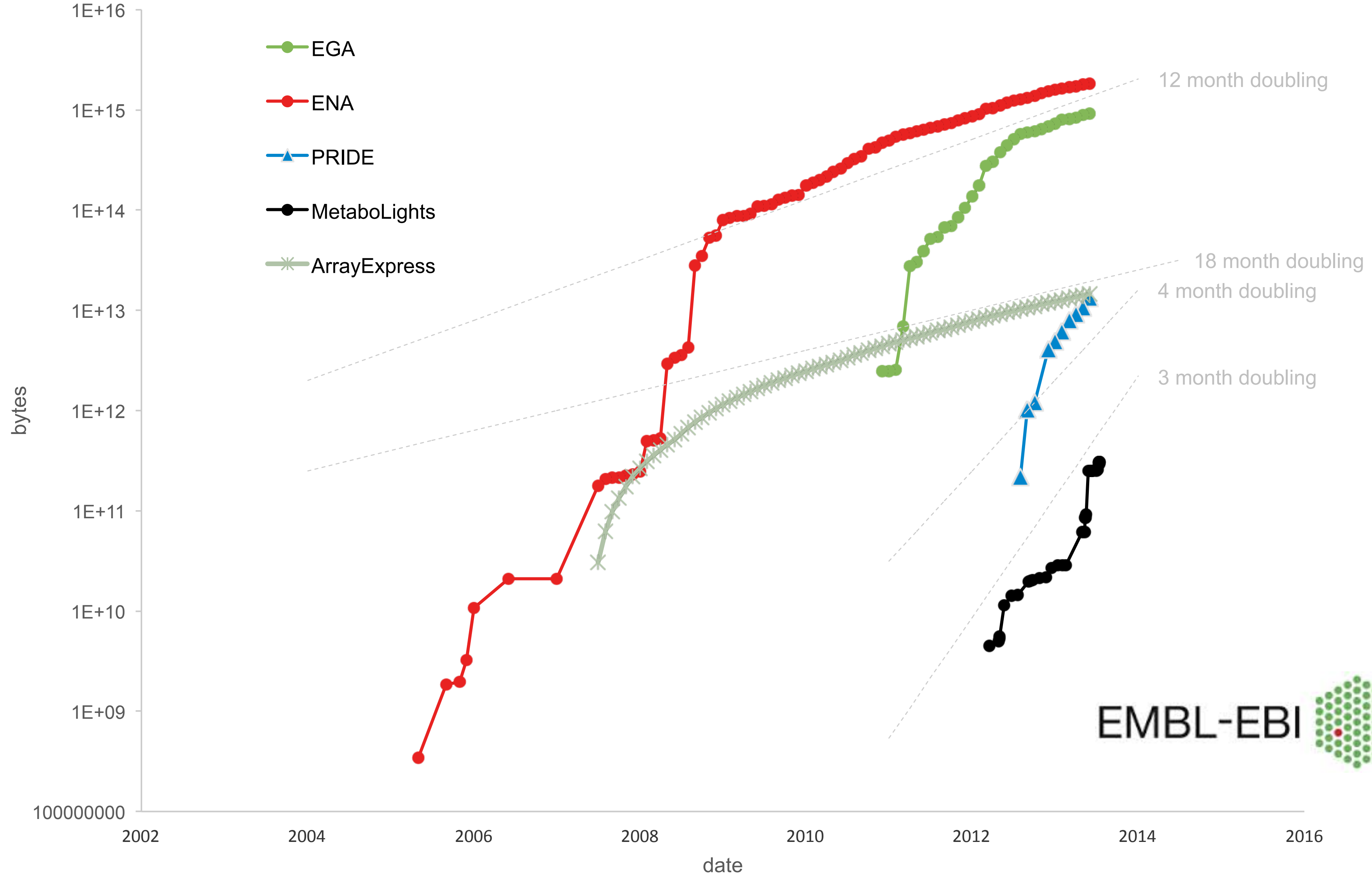
  Integration and interoperability of data and services.

- **Training**

  Professional skills for managing and exploiting data

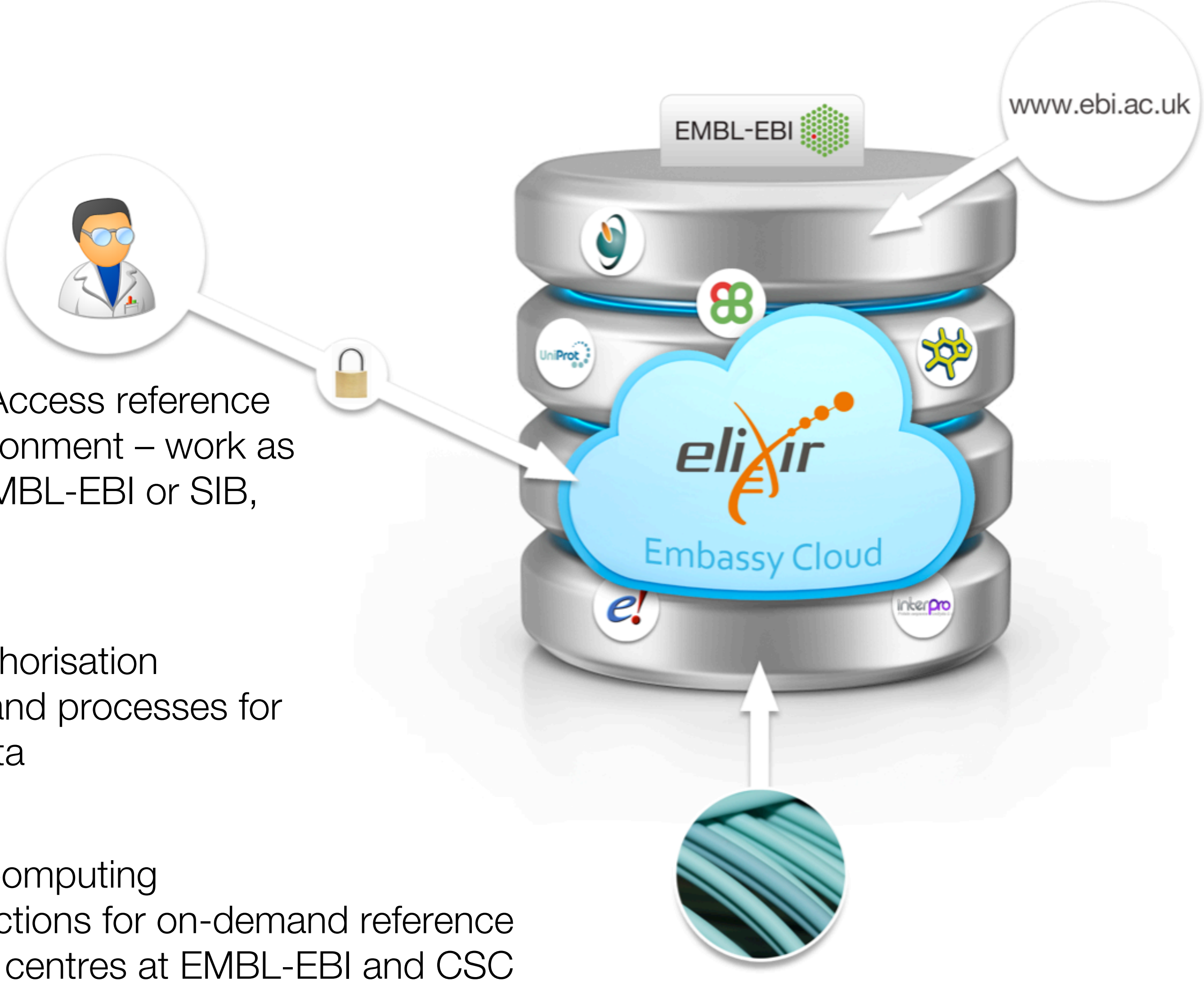


Growing data



EMBL-EBI

ELIXIR pilots to address key challenges in biomedical research:



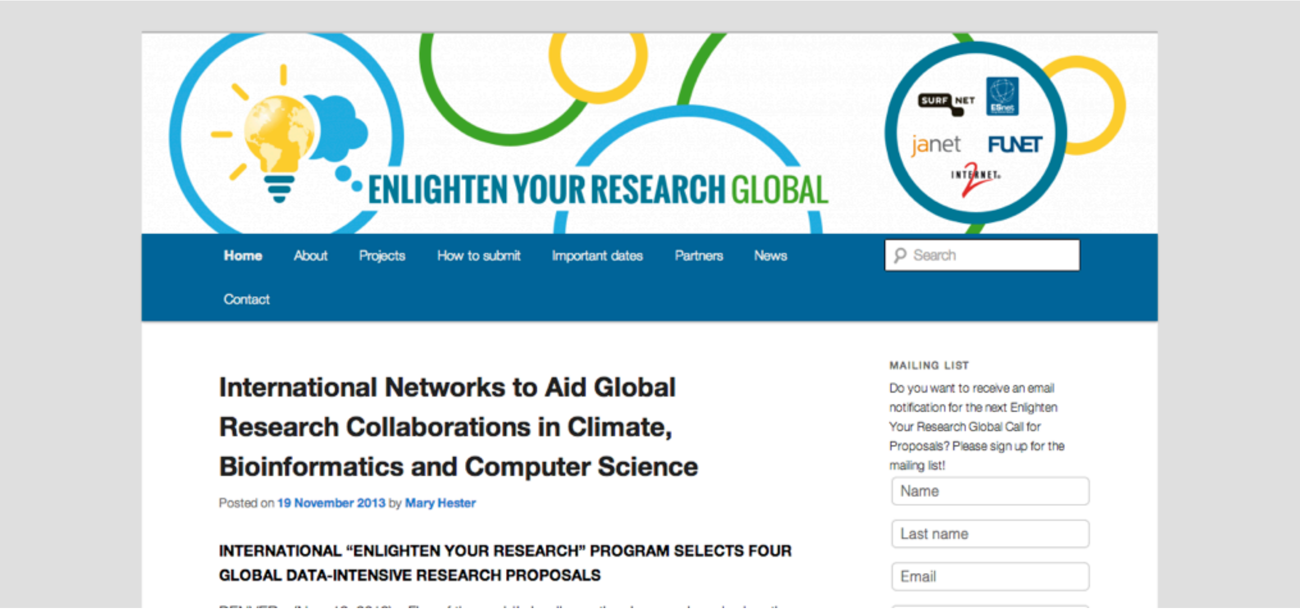
1. Cloud computing
   “Embassy cloud”: Access reference data in a virtual environment – work as though you are at EMBL-EBI or SIB, Switzerland

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

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

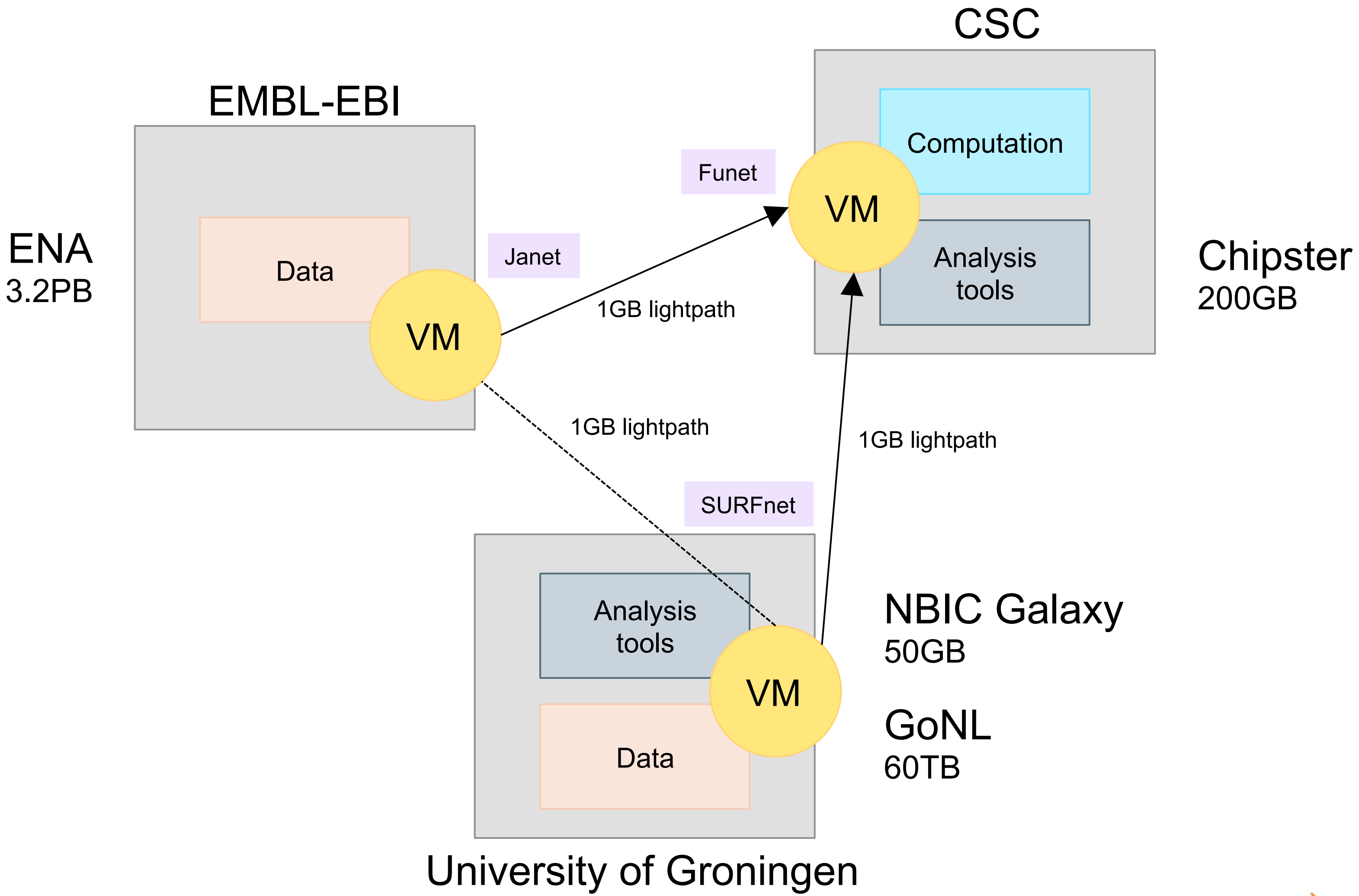
Do you want to receive an email notification for the next Enlighten Your Research Global Call for Proposals? Please sign up for the mailing list!

Name

Last name

Email

Cross-site VM Operation



CSC

EMBL-EBI

Computation

Funet

VM

Analysis tools

ENA
3.2PB

Data

Janet

VM

1GB lightpath

Chipster
200GB

1GB lightpath

1GB lightpath

SURFnet

NBIC Galaxy
50GB

Analysis tools

VM

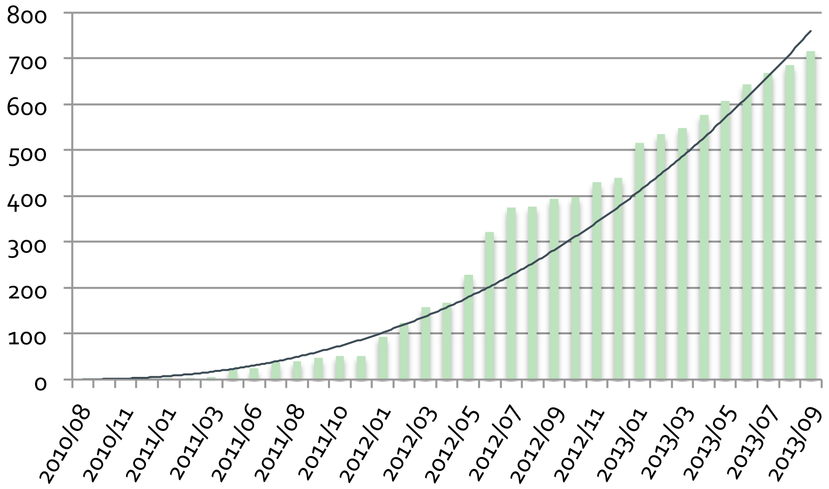
GoNL
60TB

Data

University of Groningen

ELIXIR Pilot: EGA as a joint venture

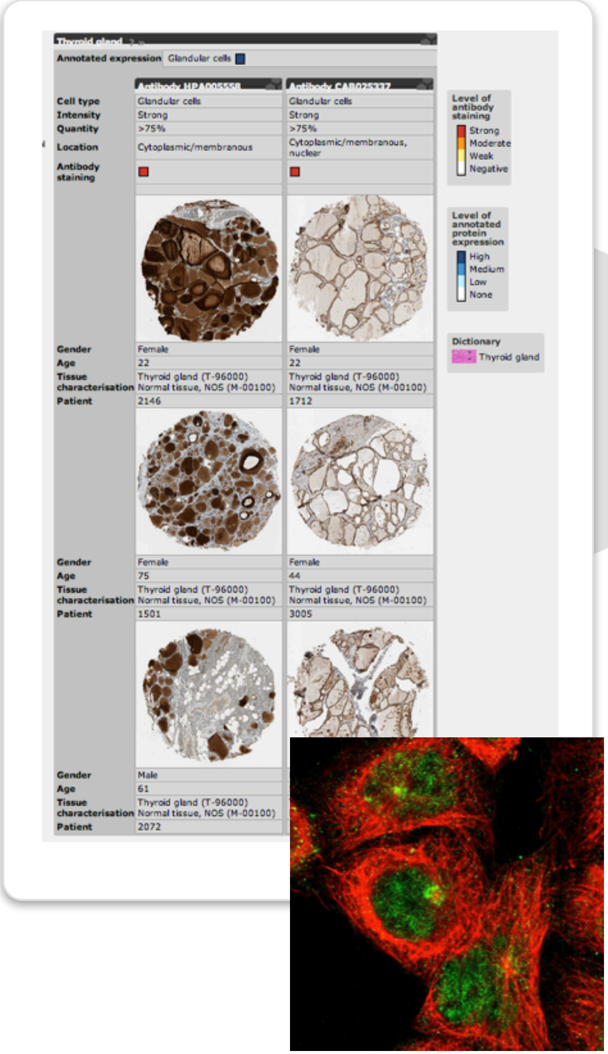
- Primary archive for any data consented for sharing in the context of research but not for fully public distribution
  - Secure storage, management and dissemination of data – raw or processed - from biomedical research projects.
  - Phenotypic data collected from the subjects.
  - Submissions must be de-identified and in accordance with the informed consent.
  - Data are packed into datasets that are governed by a Data Access Committee (DAC).
    - Authentication - each DAC approved individual will have a personal EGA account.
    - Authorization – DACs attach access permission(s) to the EGA account(s).
- EGA hosts more than 450 studies and discoverability to the 732 that are in both EGA and dbGaP
- EGA supports more than 400 user requests per month



Under the ELIXIR pilot project, the CRG and the EBI have agreed to “Explore ways in which the CRG’s emerging Node could share responsibility for production of the EGA in future”…

... which translates into managing peer database representations of the EGA Project hosted jointly by the Hub and the Spanish node of ELIXIR

Data interoperability – Human Protein Atlas



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	44
Tissue characterisation	Thyroid gland (T-96000) Normal tissue, NOS (M-00100)	Thyroid gland (T-96000) Normal tissue, NOS (M-00100)
Patient	1501	3005
Gender	Male	
Age	61	
Tissue characterisation	Thyroid gland (T-96000) Normal tissue, NOS (M-00100)	
Patient	2072	

Level of antibody staining

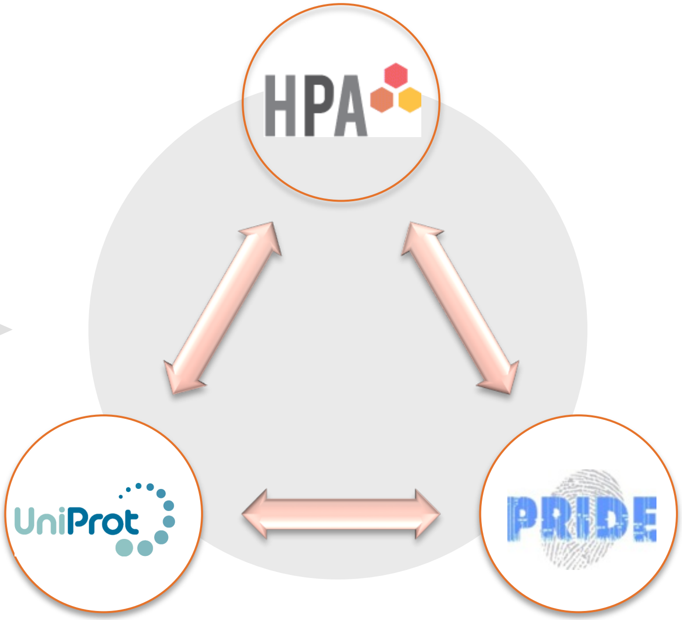
- Strong
- Moderate
- Weak
- Negative

Level of annotated protein expression

- High
- Medium
- Low
- None

Dictionary

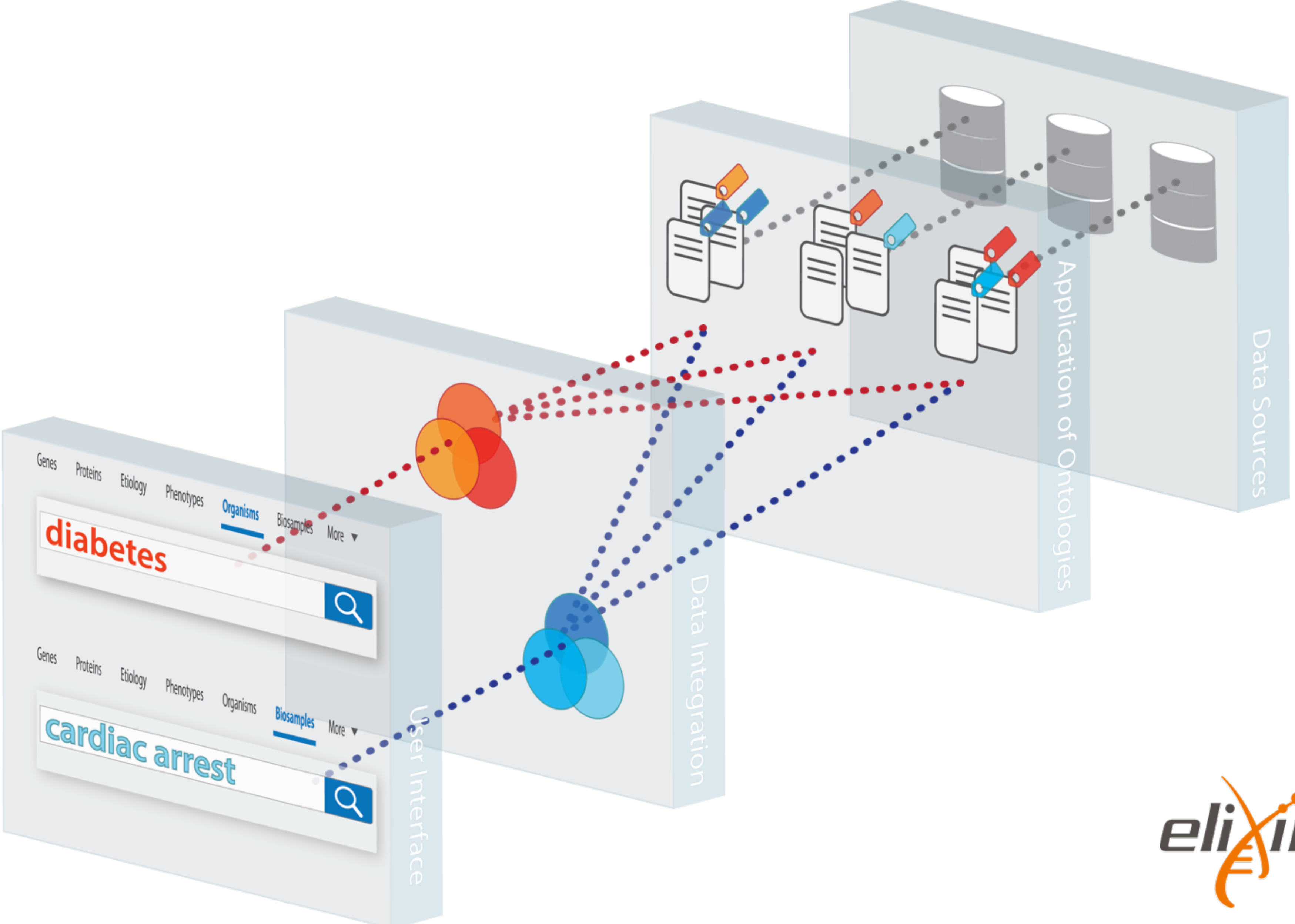
Thyroid gland



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.

Improving Links Between distributed European resources

Provide standardss, toolkits, and interoprable components for all layers



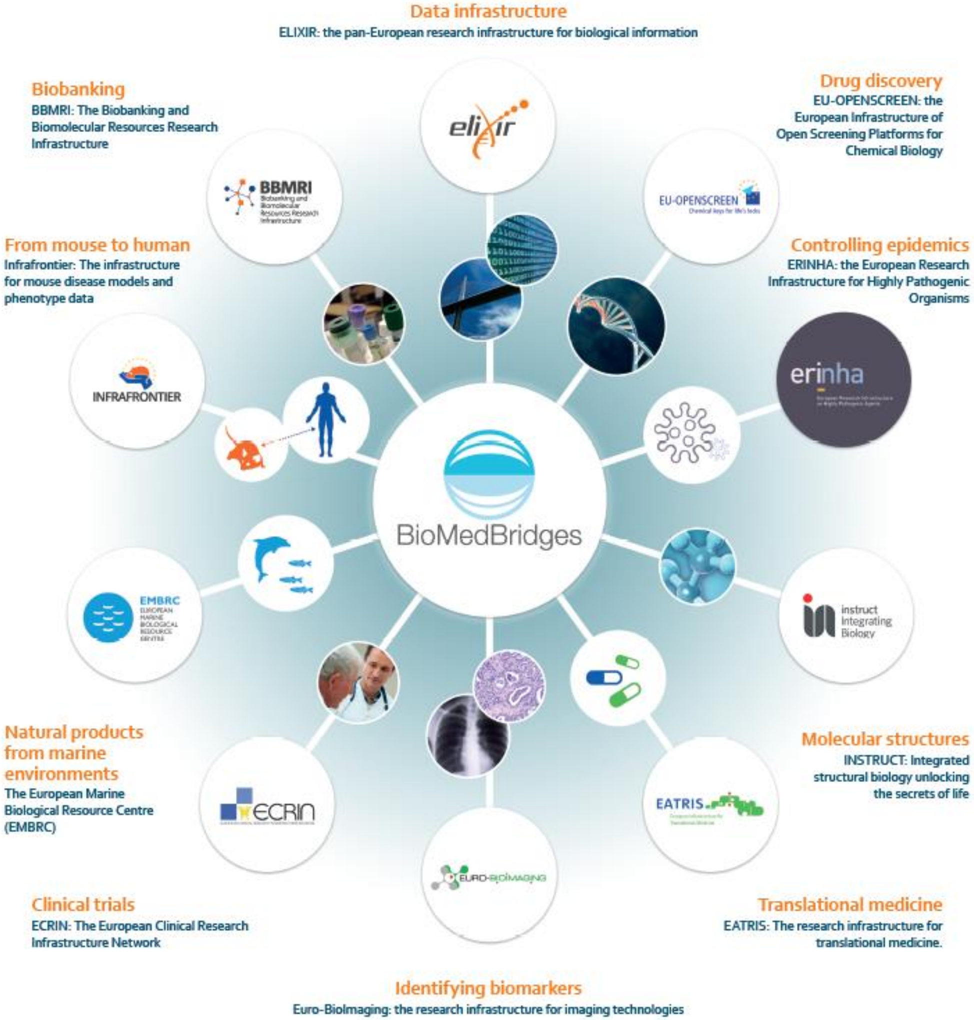
BioMedBridges

Ten new biomedical sciences research infrastructures stronger through common links

- Computational 'data and service' bridges between the BMS RIs

- Interoperability between data and services in the biological, medical, translational and clinical domains

- Link basic biological research data with clinical research and associated data



Data infrastructure
ELIXIR: the pan-European research infrastructure for biological information

Biobanking
BBMRI: The Biobanking and Biomolecular Resources Research Infrastructure

Drug discovery
EU-OPENSCREEN: the European Infrastructure of Open Screening Platforms for Chemical Biology

From mouse to human
Infrafrontier: The infrastructure for mouse disease models and phenotype data

Controlling epidemics
ERINHA: the European Research Infrastructure for Highly Pathogenic Organisms

Natural products from marine environments
The European Marine Biological Resource Centre (EMBRC)

Molecular structures
INSTRUCT: Integrated structural biology unlocking the secrets of life

Clinical trials
ECRIN: The European Clinical Research Infrastructure Network

Translational medicine
EATRIS: The research infrastructure for translational medicine.

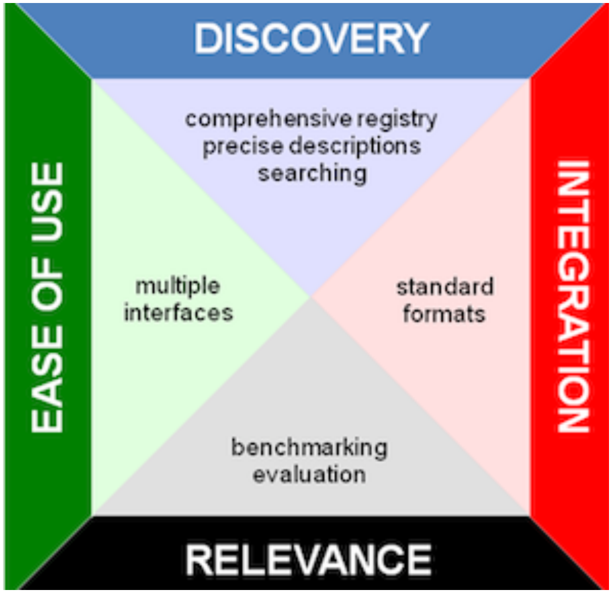
Identifying biomarkers
Euro-BioImaging: the research infrastructure for imaging technologies

SEVENTH FRAMEWORK PROGRAMME

elixir

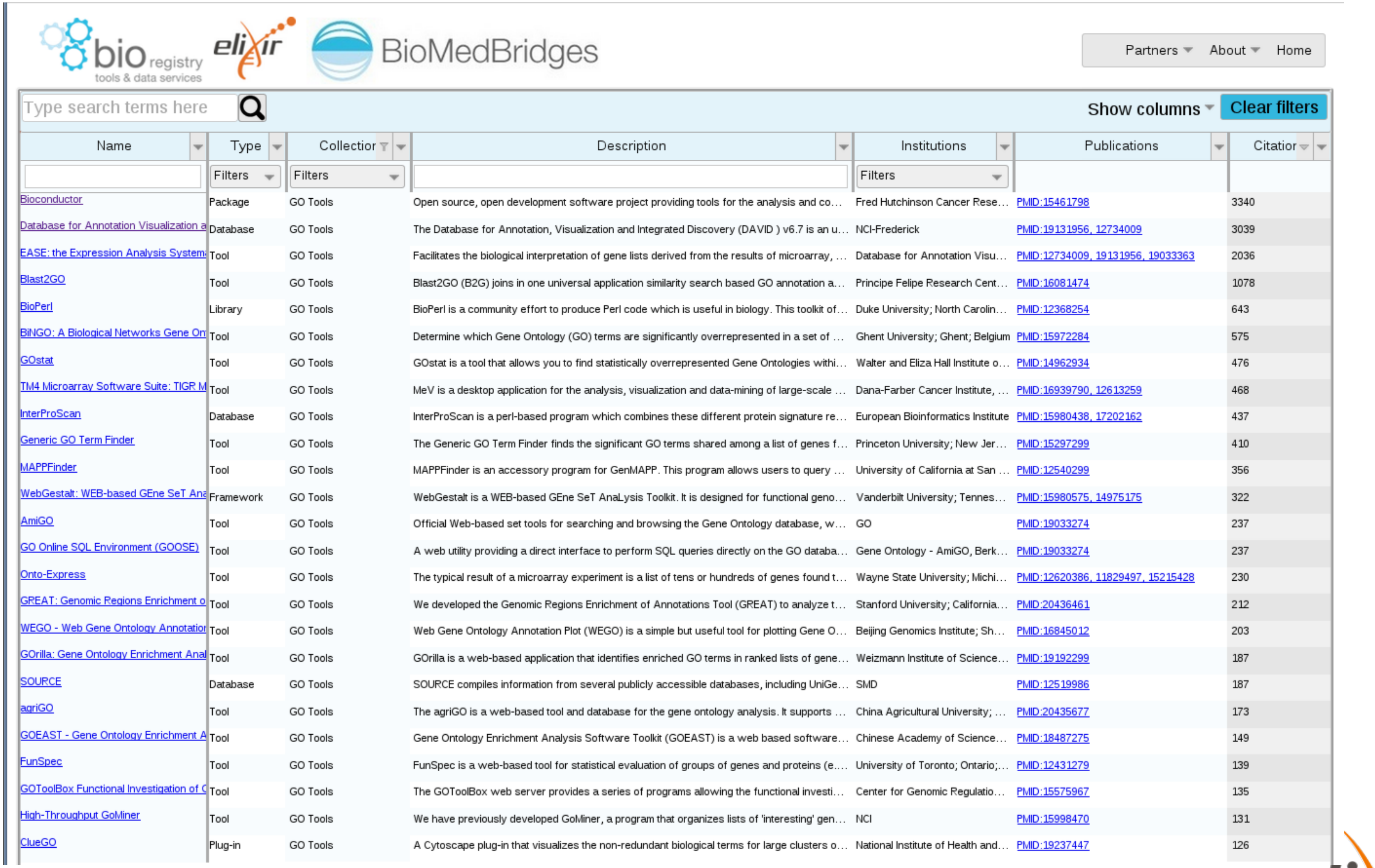
Service and Resource Registry

- Provides a simple search interface
- Content: 1943 tools etc., 22,232 annotation
  - E.g. URL, text, ontology term: type, formats ..
- Classifies tools using an ontology
  - E.g. Sequence analysis tool
- Download complete content
- Supports a wide scope of tools
- Provides an interface to the literature
- Simple spreadsheet population
- Domain neutral



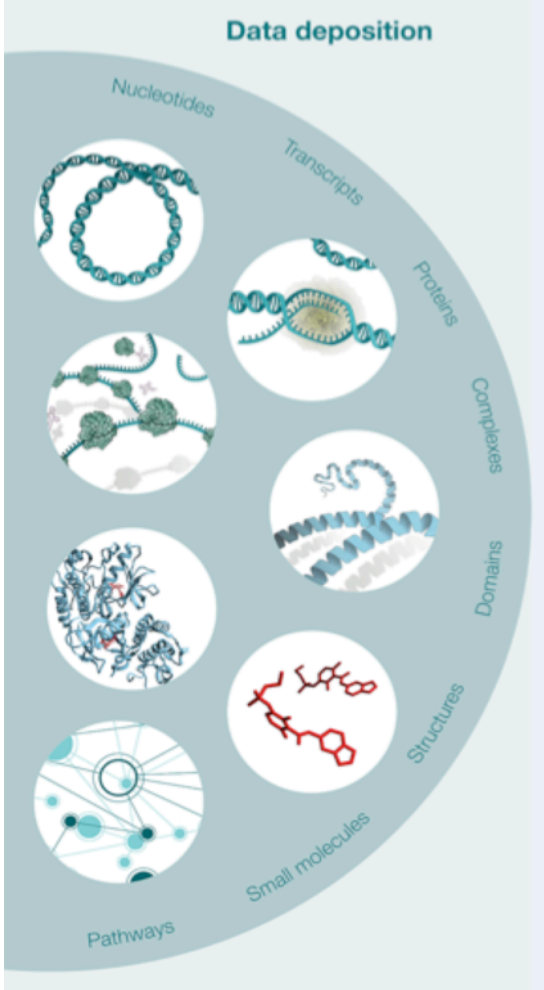
http://bioregistry.cbs.dtu.dk/

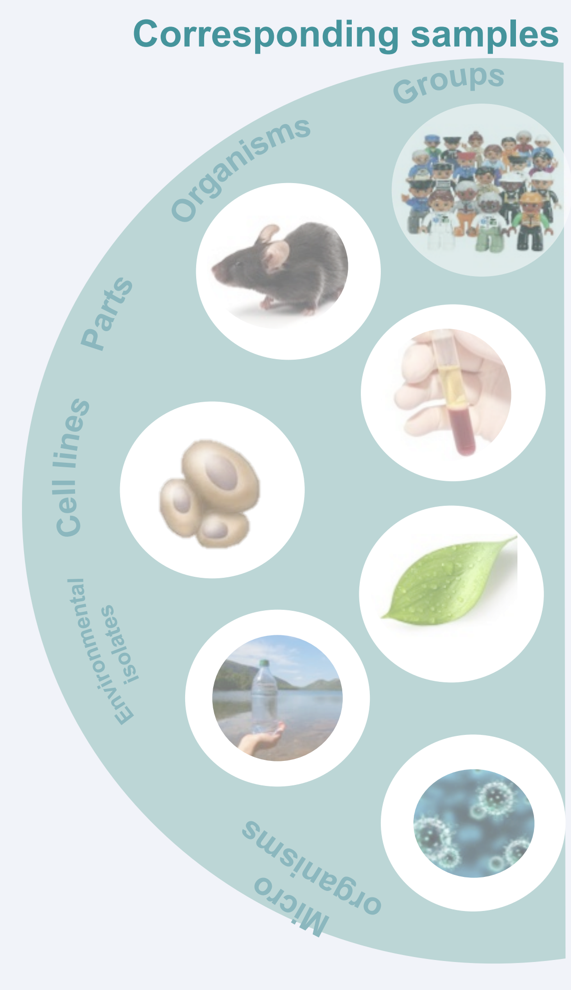
Service and Resource Registry



Name	Type	Collectior	Description	Institutions	Publications	Citatior
Bioconductor	Package	GO Tools	Open source, open development software project providing tools for the analysis and co…	Fred Hutchinson Cancer Rese…	PMID:15461798	3340
Database for Annotation Visualization a	Database	GO Tools	The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an u…	NCI-Frederick	PMID:19131956, 12734009	3039
EASE: the Expression Analysis Systema	Tool	GO Tools	Facilitates the biological interpretation of gene lists derived from the results of microarray, …	Database for Annotation Visu…	PMID:12734009, 19131956, 19033363	2036
Blast2GO	Tool	GO Tools	Blast2GO (B2G) joins in one universal application similarity search based GO annotation a…	Principe Felipe Research Cent…	PMID:16081474	1078
BioPerl	Library	GO Tools	BioPerl is a community effort to produce Perl code which is useful in biology. This toolkit of…	Duke University; North Carolin…	PMID:12368254	643
BiNGO: A Biological Networks Gene On	Tool	GO Tools	Determine which Gene Ontology (GO) terms are significantly overrepresented in a set of …	Ghent University; Ghent; Belgium	PMID:15972284	575
GOstat	Tool	GO Tools	GOstat is a tool that allows you to find statistically overrepresented Gene Ontologies withi…	Walter and Eliza Hall Institute o…	PMID:14962934	476
TM4 Microarray Software Suite: TIGR M	Tool	GO Tools	MeV is a desktop application for the analysis, visualization and data-mining of large-scale …	Dana-Farber Cancer Institute, …	PMID:16939790, 12613259	468
InterProScan	Database	GO Tools	InterProScan is a perl-based program which combines these different protein signature re…	European Bioinformatics Institute	PMID:15980438, 17202162	437
Generic GO Term Finder	Tool	GO Tools	The Generic GO Term Finder finds the significant GO terms shared among a list of genes f…	Princeton University; New Jer…	PMID:15297299	410
MAPPFinder	Tool	GO Tools	MAPPFinder is an accessory program for GenMAPP. This program allows users to query …	University of California at San …	PMID:12540299	356
WebGestalt: WEB-based GEne SeT Ana	Framework	GO Tools	WebGestalt is a WEB-based GEne SeT AnaLysis Toolkit. It is designed for functional geno…	Vanderbilt University; Tennes…	PMID:15980575, 14975175	322
AmiGO	Tool	GO Tools	Official Web-based set tools for searching and browsing the Gene Ontology database, w…	GO	PMID:19033274	237
GO Online SQL Environment (GOOSE)	Tool	GO Tools	A web utility providing a direct interface to perform SQL queries directly on the GO databa…	Gene Ontology - AmiGO, Berk…	PMID:19033274	237
Onto-Express	Tool	GO Tools	The typical result of a microarray experiment is a list of tens or hundreds of genes found t…	Wayne State University; Michi…	PMID:12620386, 11829497, 15215428	230
GREAT: Genomic Regions Enrichment o	Tool	GO Tools	We developed the Genomic Regions Enrichment of Annotations Tool (GREAT) to analyze t…	Stanford University; California…	PMID:20436461	212
WEGO - Web Gene Ontology Annotation	Tool	GO Tools	Web Gene Ontology Annotation Plot (WEGO) is a simple but useful tool for plotting Gene O…	Beijing Genomics Institute; Sh…	PMID:16845012	203
GOrilla: Gene Ontology Enrichment Anal	Tool	GO Tools	GOrilla is a web-based application that identifies enriched GO terms in ranked lists of gene…	Weizmann Institute of Science…	PMID:19192299	187
SOURCE	Database	GO Tools	SOURCE compiles information from several publicly accessible databases, including UniGe…	SMD	PMID:12519986	187
agriGO	Tool	GO Tools	The agriGO is a web-based tool and database for the gene ontology analysis. It supports …	China Agricultural University; …	PMID:20435677	173
GOEAST - Gene Ontology Enrichment A	Tool	GO Tools	Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) is a web based software…	Chinese Academy of Science…	PMID:18487275	149
FunSpec	Tool	GO Tools	FunSpec is a web-based tool for statistical evaluation of groups of genes and proteins (e.…	University of Toronto; Ontario;…	PMID:12431279	139
GOToolBox Functional Investigation of C	Tool	GO Tools	The GOToolBox web server provides a series of programs allowing the functional investi…	Center for Genomic Regulatio…	PMID:15575967	135
High-Throughput GoMiner	Tool	GO Tools	We have previously developed GoMiner, a program that organizes lists of 'interesting' gen…	NCI	PMID:15998470	131
ClueGO	Plug-in	GO Tools	A Cytoscape plug-in that visualizes the non-redundant biological terms for large clusters o…	National Institute of Health and…	PMID:19237447	126

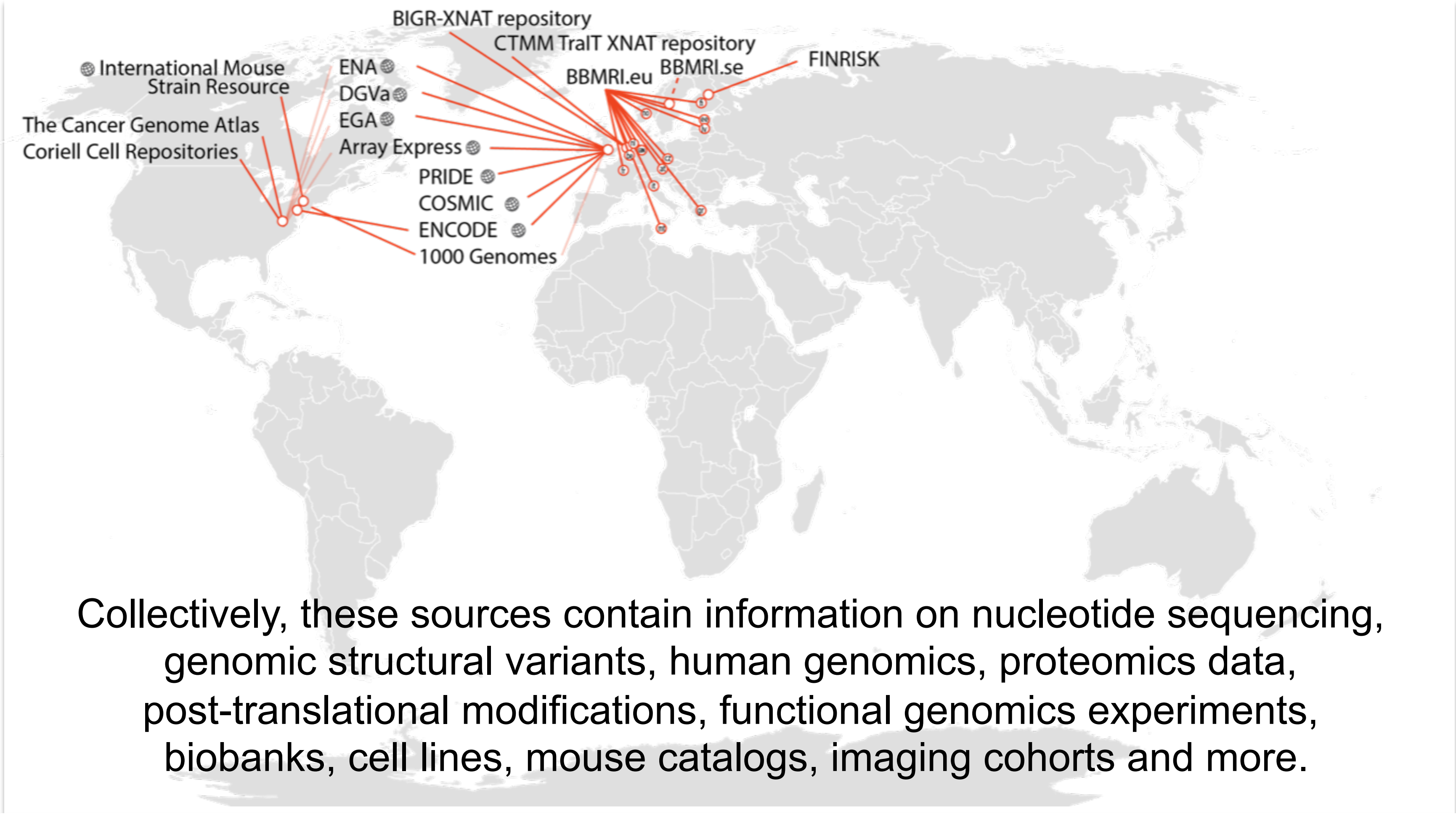
BMB: Biosample information integration and discovery





http://www.ebi.ac.uk/biosamples/

Solution: BioSD finds & links 2.8M samples from 32 sources worldwide



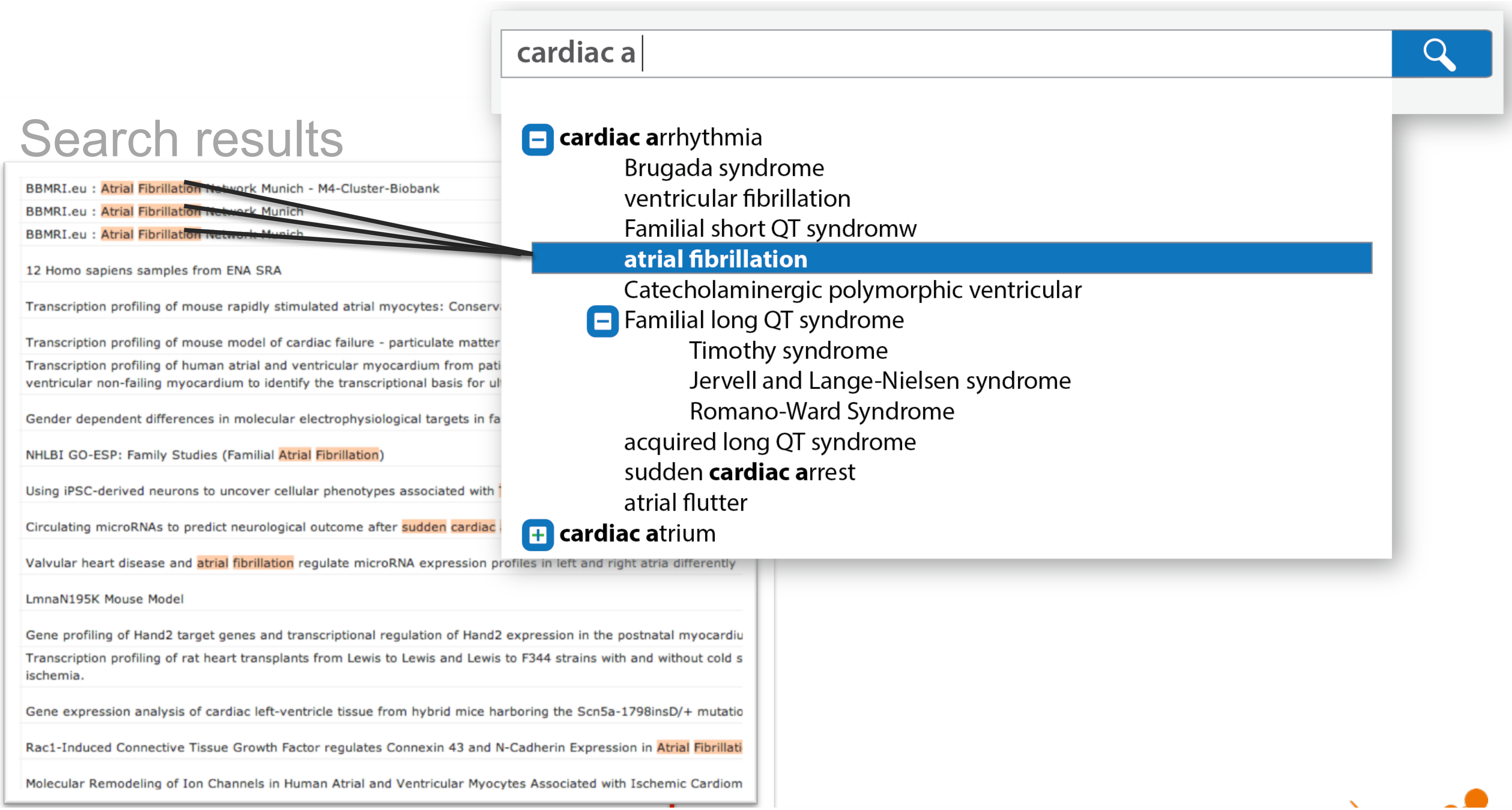
Collectively, these sources contain information on nucleotide sequencing, genomic structural variants, human genomics, proteomics data, post-translational modifications, functional genomics experiments, biobanks, cell lines, mouse catalogs, imaging cohorts and more.

BioMedBridges

elixir

SEVENTH FRAMEWORK PROGRAMME

Example query: "Cardiac Arrhythmia" yields over 450 samples from 10 groups (via BBMRI, Array Express, ENA)



cardiac a

⊟ **cardiac a**rrhythmia
- Brugada syndrome
- ventricular fibrillation
- Familial short QT syndromw
- **atrial fibrillation**
- Catecholaminergic polymorphic ventricular
- ⊟ Familial long QT syndrome
  - Timothy syndrome
  - Jervell and Lange-Nielsen syndrome
  - Romano-Ward Syndrome
- acquired long QT syndrome
- sudden **cardiac a**rrest
- atrial flutter

⊞ **cardiac a**trium

Search results

BBMRI.eu : Atrial Fibrillation Network Munich - M4-Cluster-Biobank

BBMRI.eu : Atrial Fibrillation Network Munich

BBMRI.eu : Atrial Fibrillation Network Munich

12 Homo sapiens samples from ENA SRA

Transcription profiling of mouse rapidly stimulated atrial myocytes: Conserv

Transcription profiling of mouse model of cardiac failure - particulate matter

Transcription profiling of human atrial and ventricular myocardium from pati ventricular non-failing myocardium to identify the transcriptional basis for ul

Gender dependent differences in molecular electrophysiological targets in fa

NHLBI GO-ESP: Family Studies (Familial Atrial Fibrillation)

Using iPSC-derived neurons to uncover cellular phenotypes associated with

Circulating microRNAs to predict neurological outcome after sudden cardiac

Valvular heart disease and atrial fibrillation regulate microRNA expression profiles in left and right atria differently

LmnaN195K Mouse Model

Gene profiling of Hand2 target genes and transcriptional regulation of Hand2 expression in the postnatal myocardiu

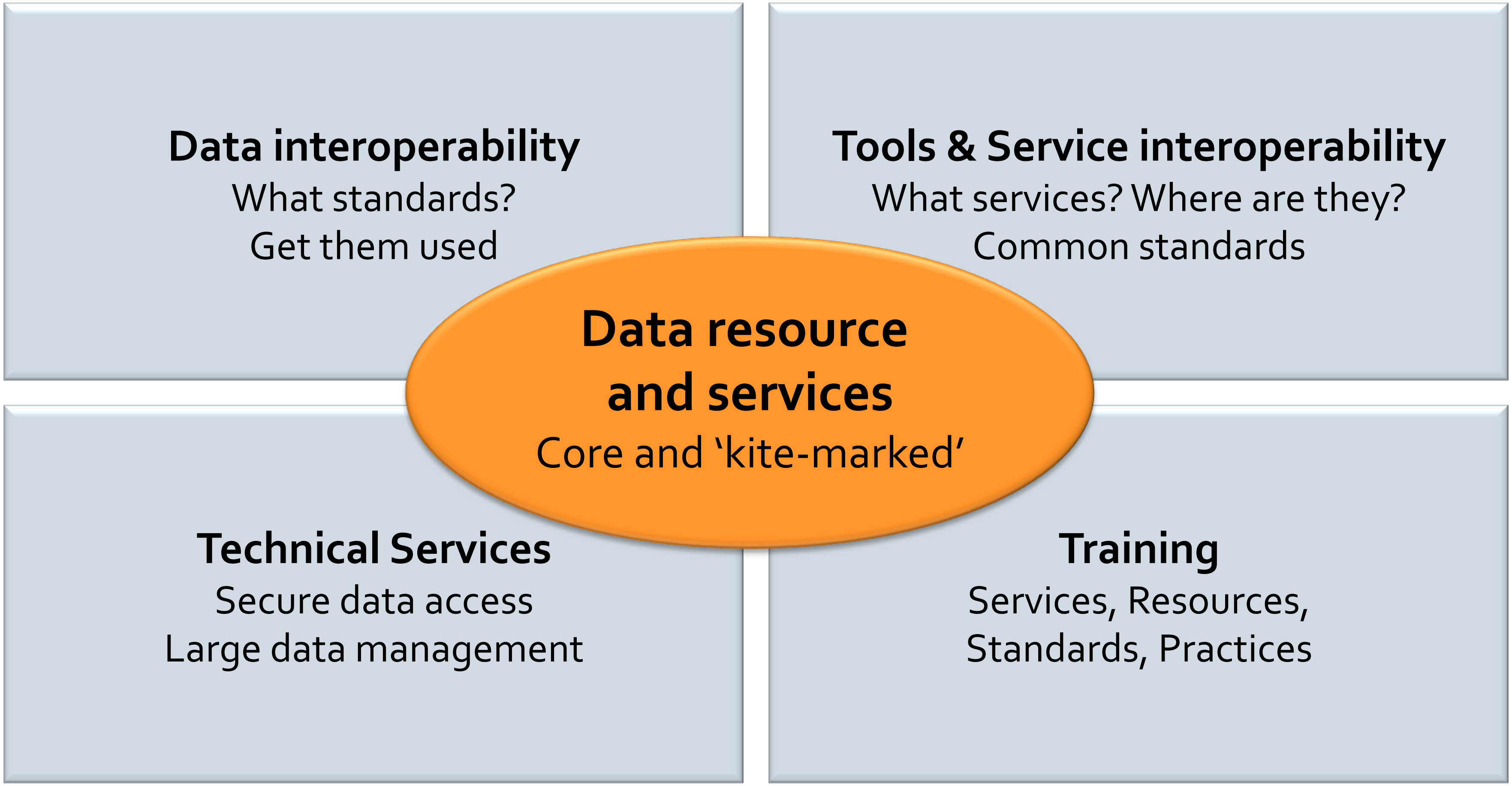
Transcription profiling of rat heart transplants from Lewis to Lewis and Lewis to F344 strains with and without cold s ischemia.

Gene expression analysis of cardiac left-ventricle tissue from hybrid mice harboring the Scn5a-1798insD/+ mutatio

Rac1-Induced Connective Tissue Growth Factor regulates Connexin 43 and N-Cadherin Expression in Atrial Fibrillati

Molecular Remodeling of Ion Channels in Human Atrial and Ventricular Myocytes Associated with Ischemic Cardiom

ELIXIR Programme



Data interoperability
What standards?
Get them used

Tools & Service interoperability
What services? Where are they?
Common standards

Data resource and services
Core and 'kite-marked'

Technical Services
Secure data access
Large data management

Training
Services, Resources,
Standards, Practices

Thank you

Belgium

Czech Republic

Denmark

EMBL

Estonia

Finland

France

Greece

Israel

Italy

Netherlands

Norway

Portugal

Slovenia

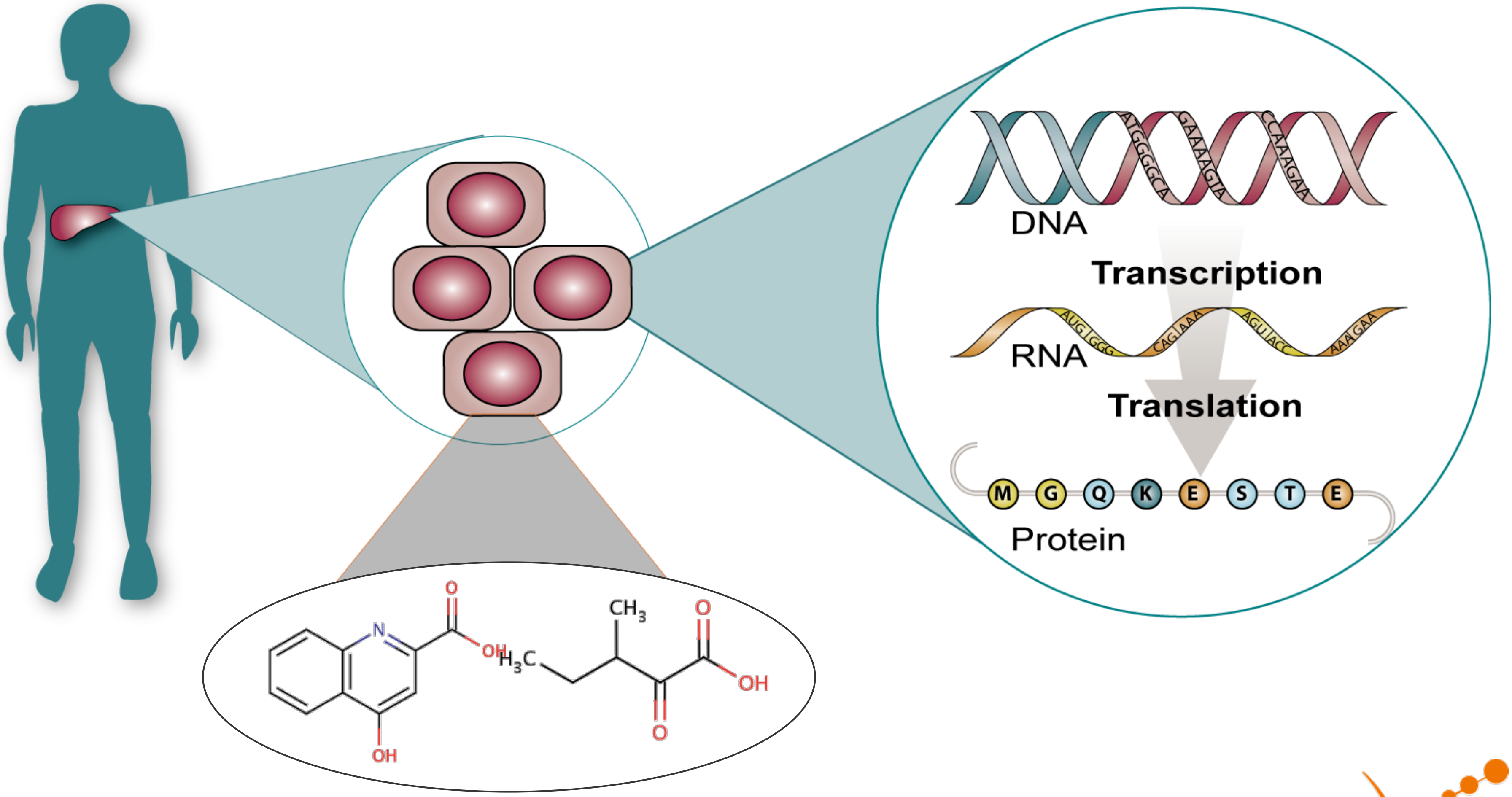
Spain

Sweden

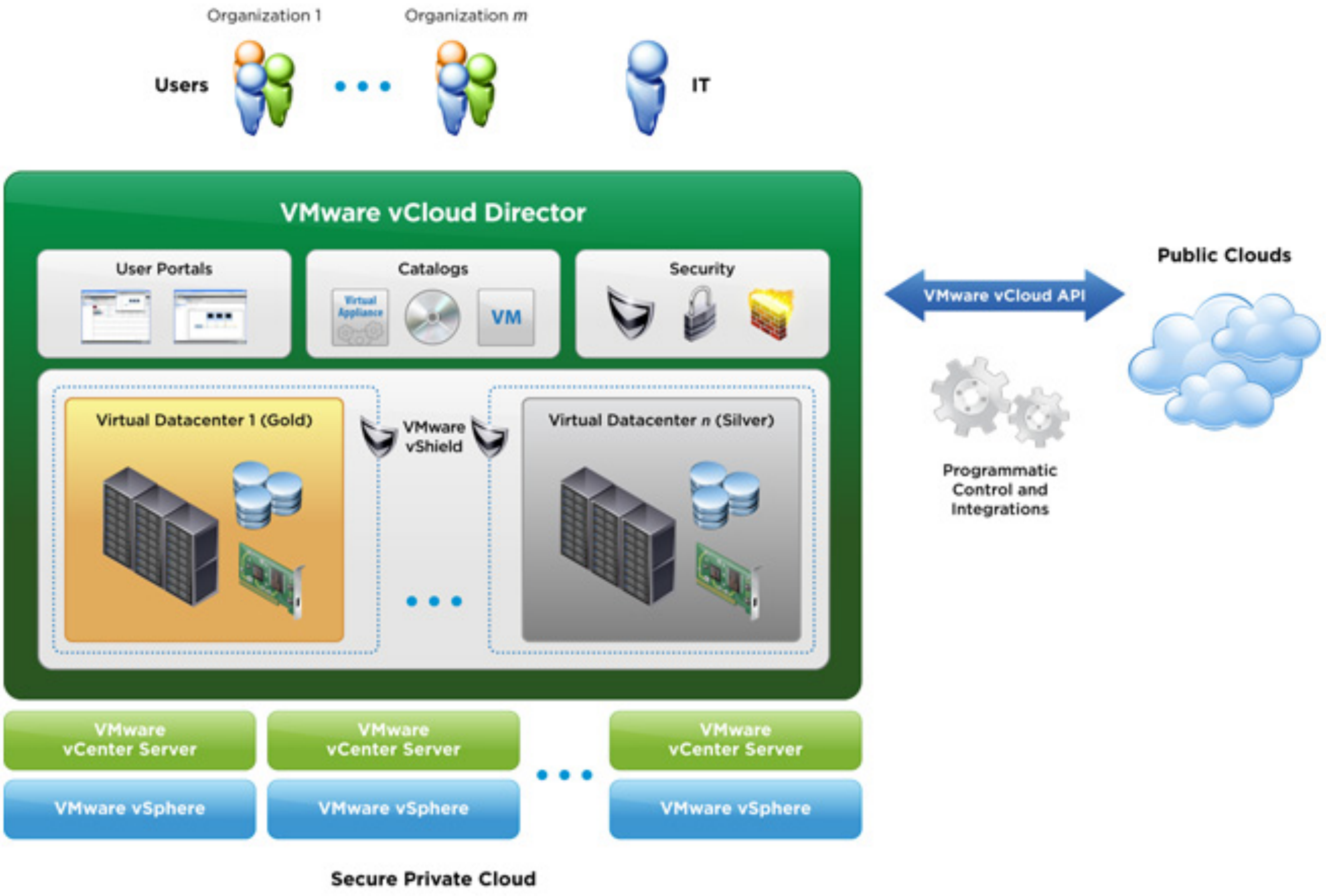
Switzerland

United Kingdom

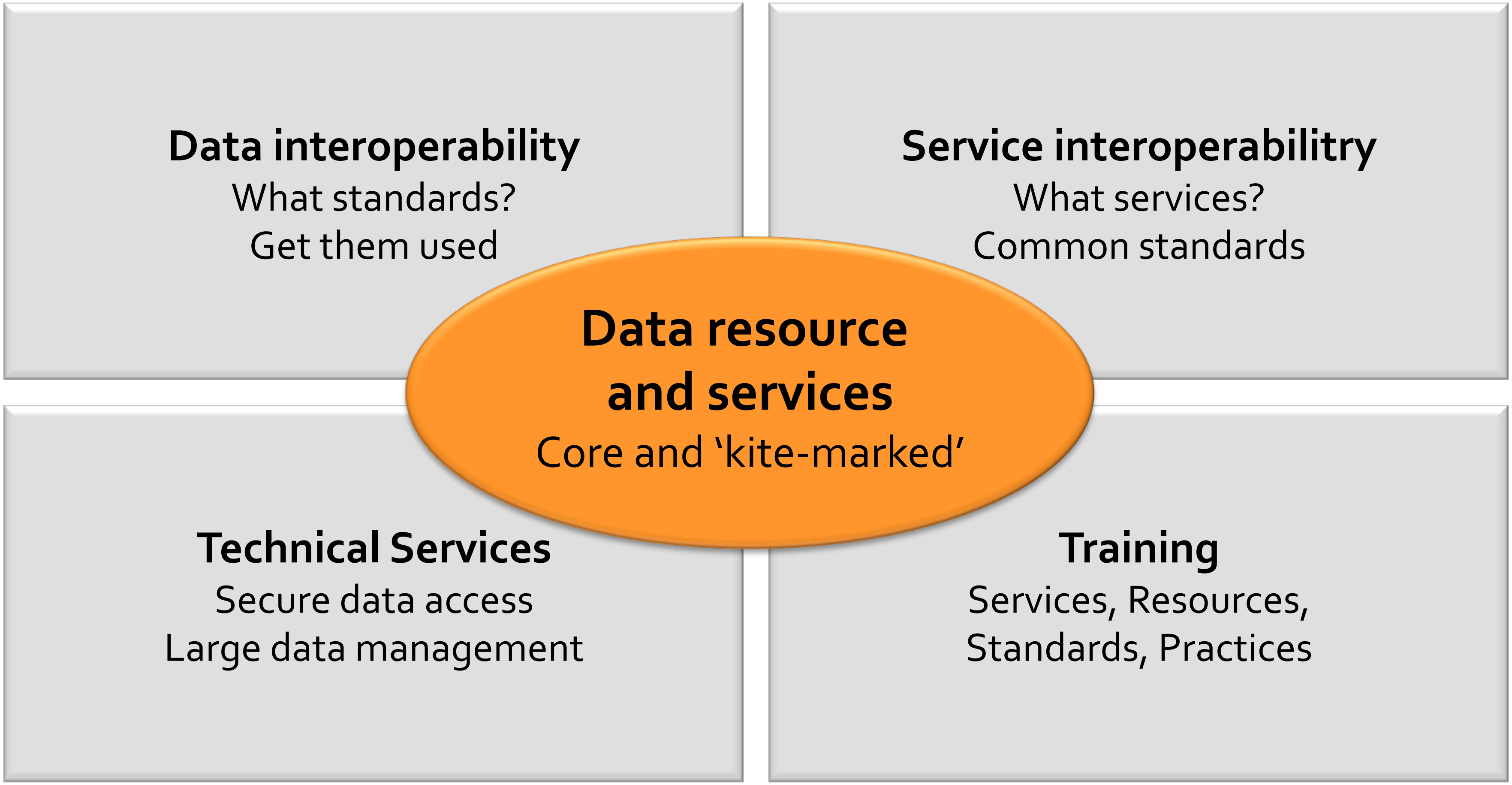
The molecules of life



Embassy Cloud logical view



Data-focused Work Stream Headlines



Data interoperability
What standards?
Get them used

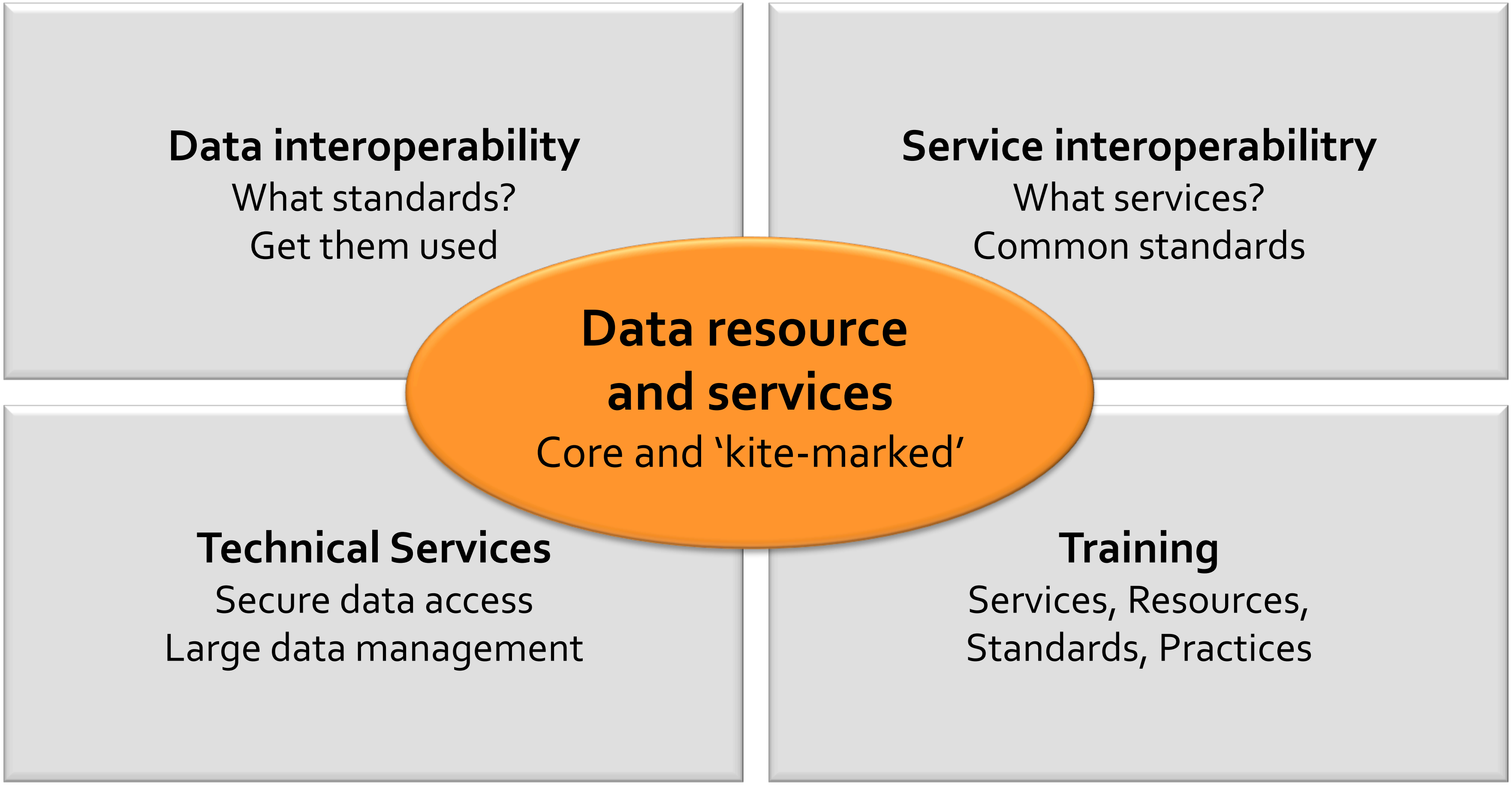
Service interoperability
What services?
Common standards

Data resource and services
Core and 'kite-marked'

Technical Services
Secure data access
Large data management

Training
Services, Resources,
Standards, Practices

Data-focused Work Stream Headlines



Data interoperability
What standards?
Get them used

Service interoperability
What services?
Common standards

Data resource and services
Core and ‘kite-marked’

Technical Services
Secure data access
Large data management

Training
Services, Resources,
Standards, Practices