ELIXIR: Data Challenges in the Life Sciences

e-IRG workshop, Athens, 9-10 June 2014

Andrew Smith
ELIXIR Hub

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

ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:

medicine

environment

bioindustries

society

The potential

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

Genome-wide analysis of crop plants

- Population growth and climate change are major challenges to food security.
- Traditional routes to crop improvement are too slow to keep up with this increase in demand.
- Understanding plant genomes helps us identify which species will be most tolerant to drought, salt and pests while still providing optimum nutrition.

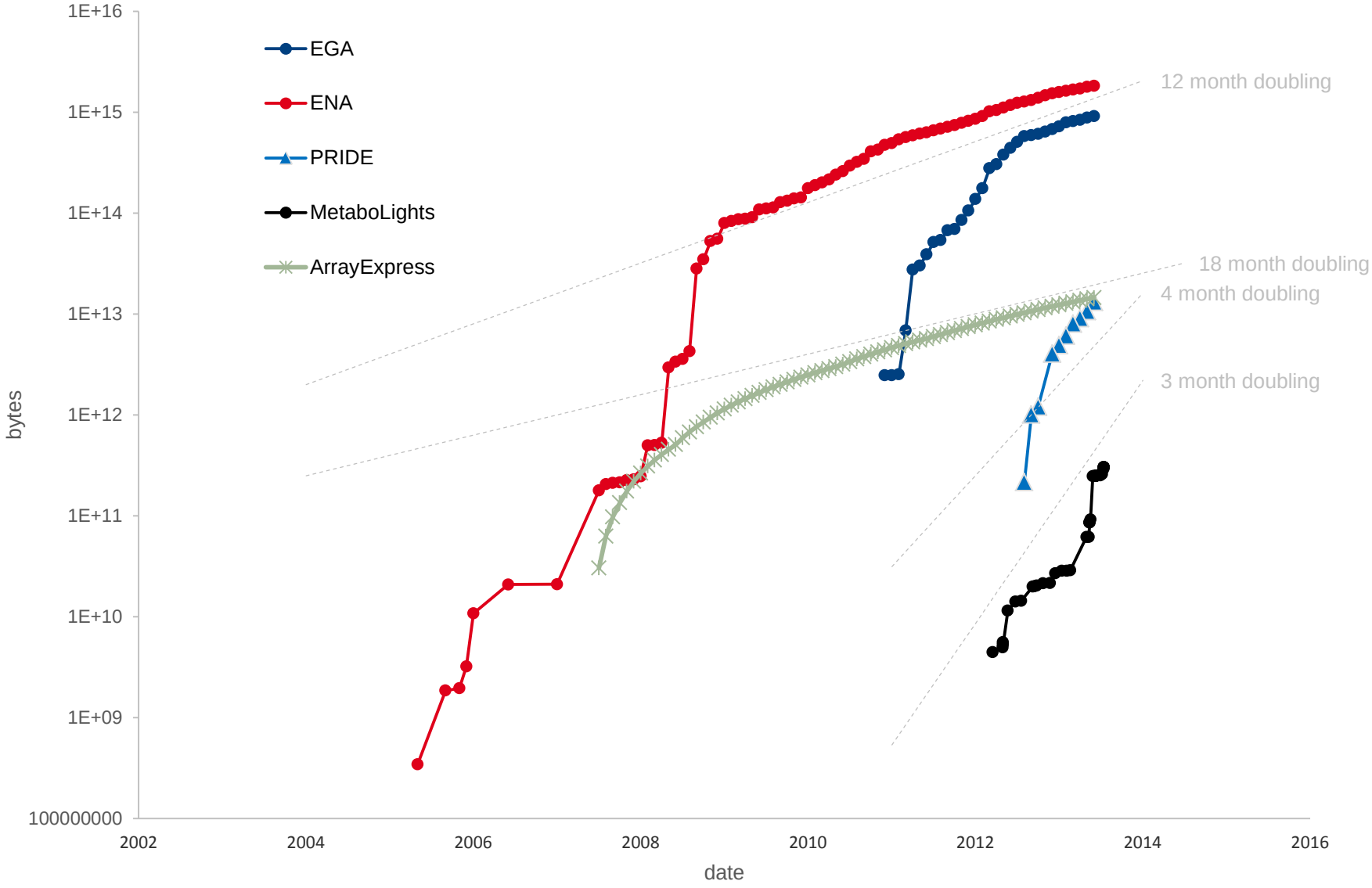
Matching the treatment to the cancer

- One in 10 women in the EU-27 will develop breast cancer before the age of 80.
- If we can identify patterns of genes that are active in different tumours, we can diagnose and treat cancers earlier.

The challenges

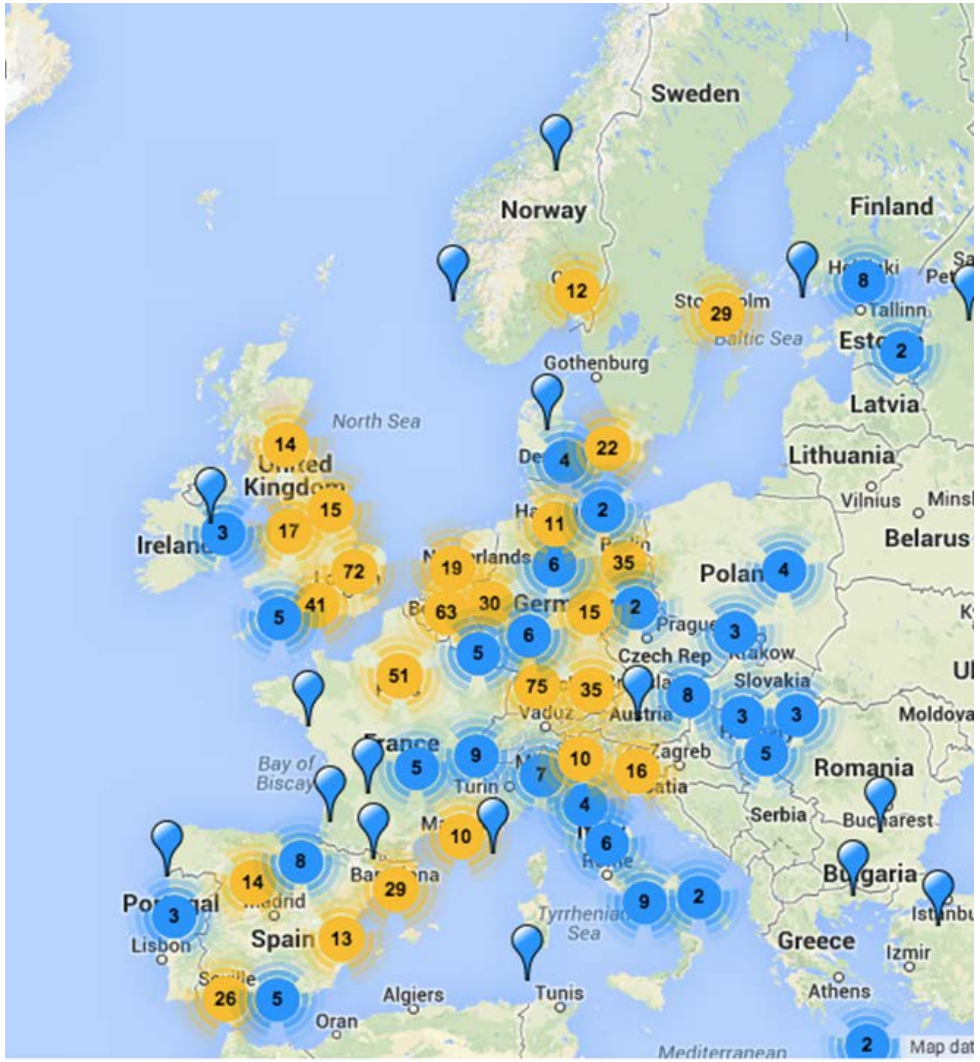
European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

Growing data



The data challenge: geography

- Data production increasing sites across Europe
- European Illumina sales up 20% 2013



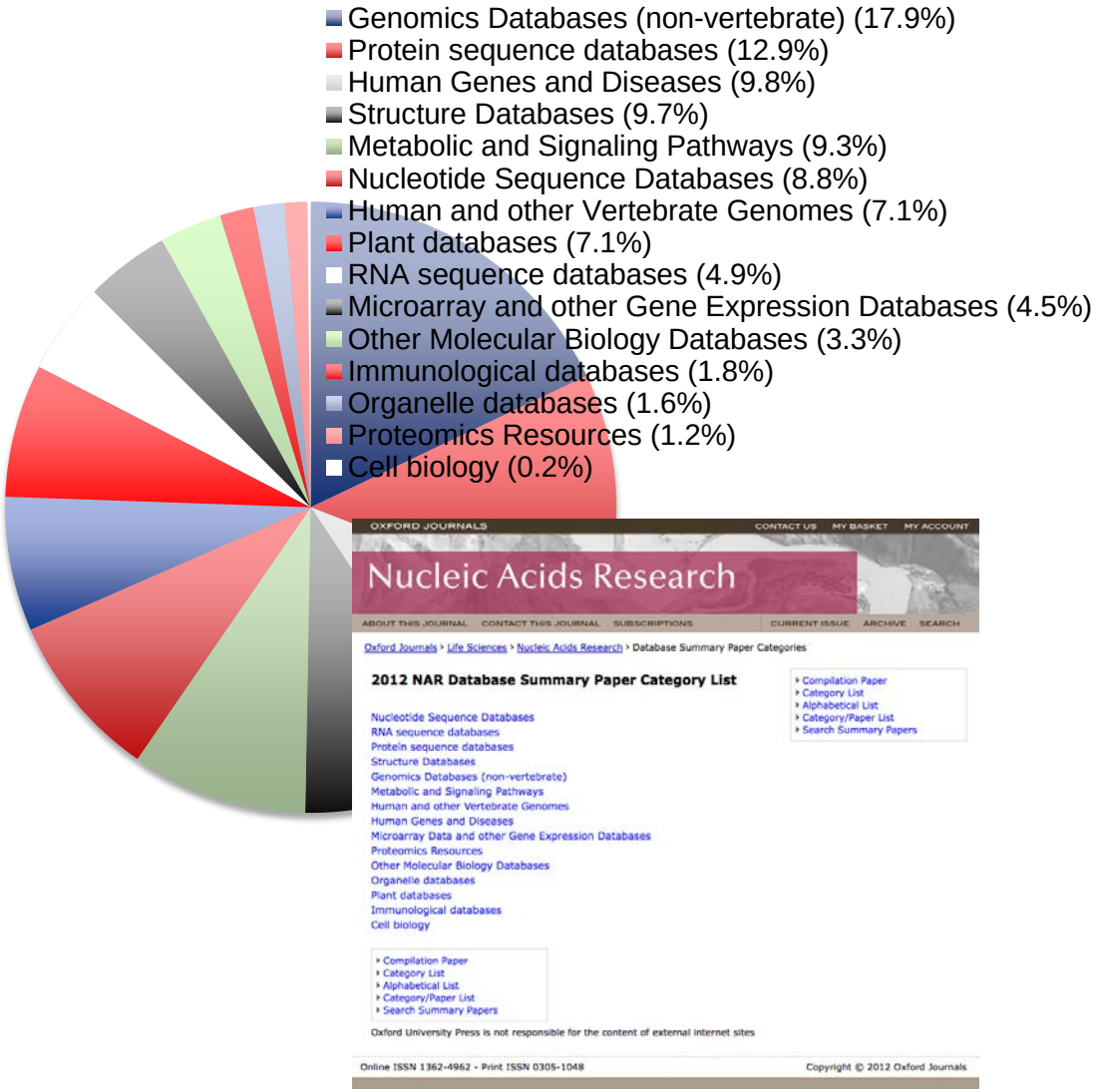
Source: http://omicsmaps.com

Data resources in life science

- Diverse
- Many
- Disperse

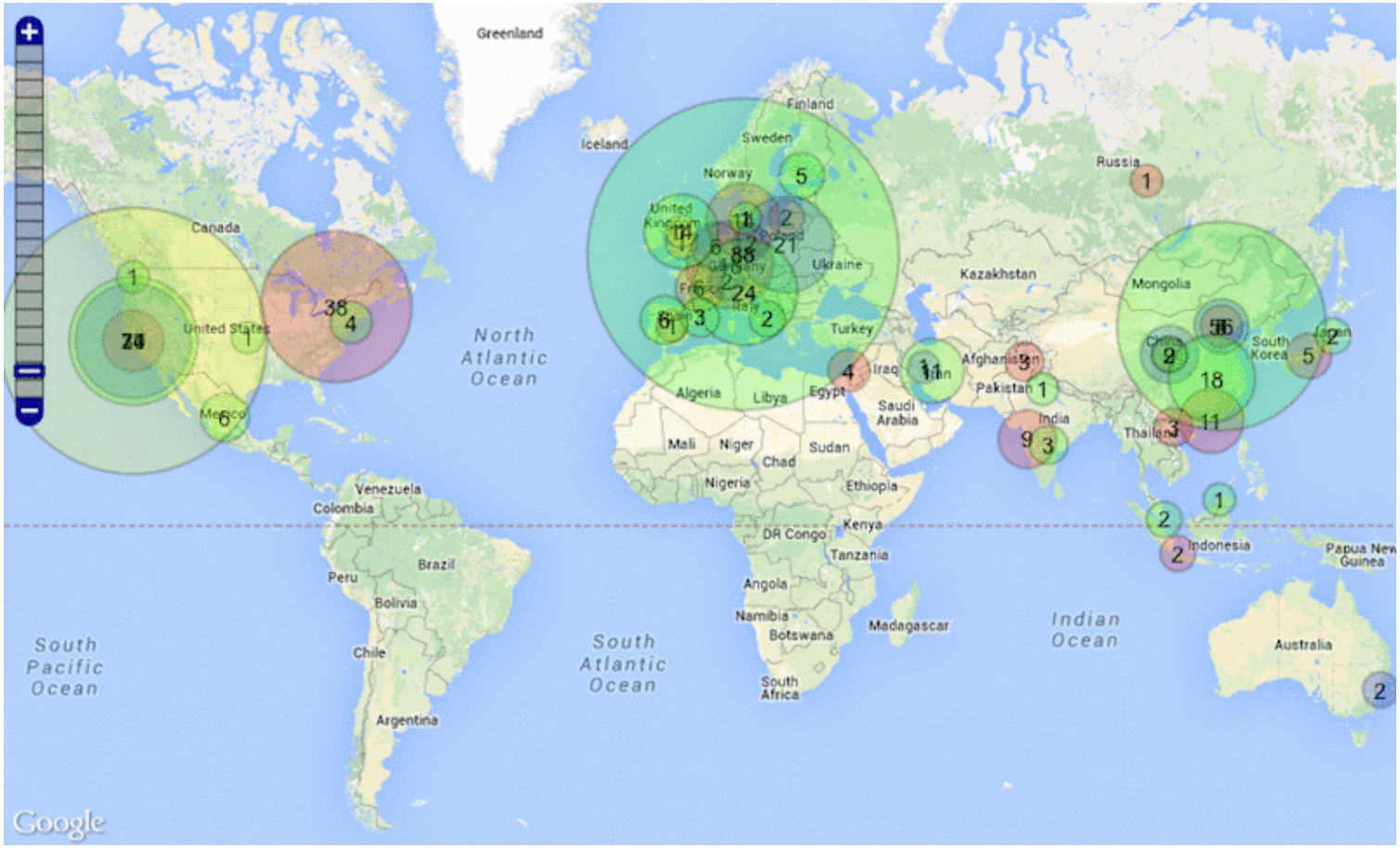
~1800

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

Users are global



Source: EMBL-EBI Live Data Map

The policy drivers: Open Access to data

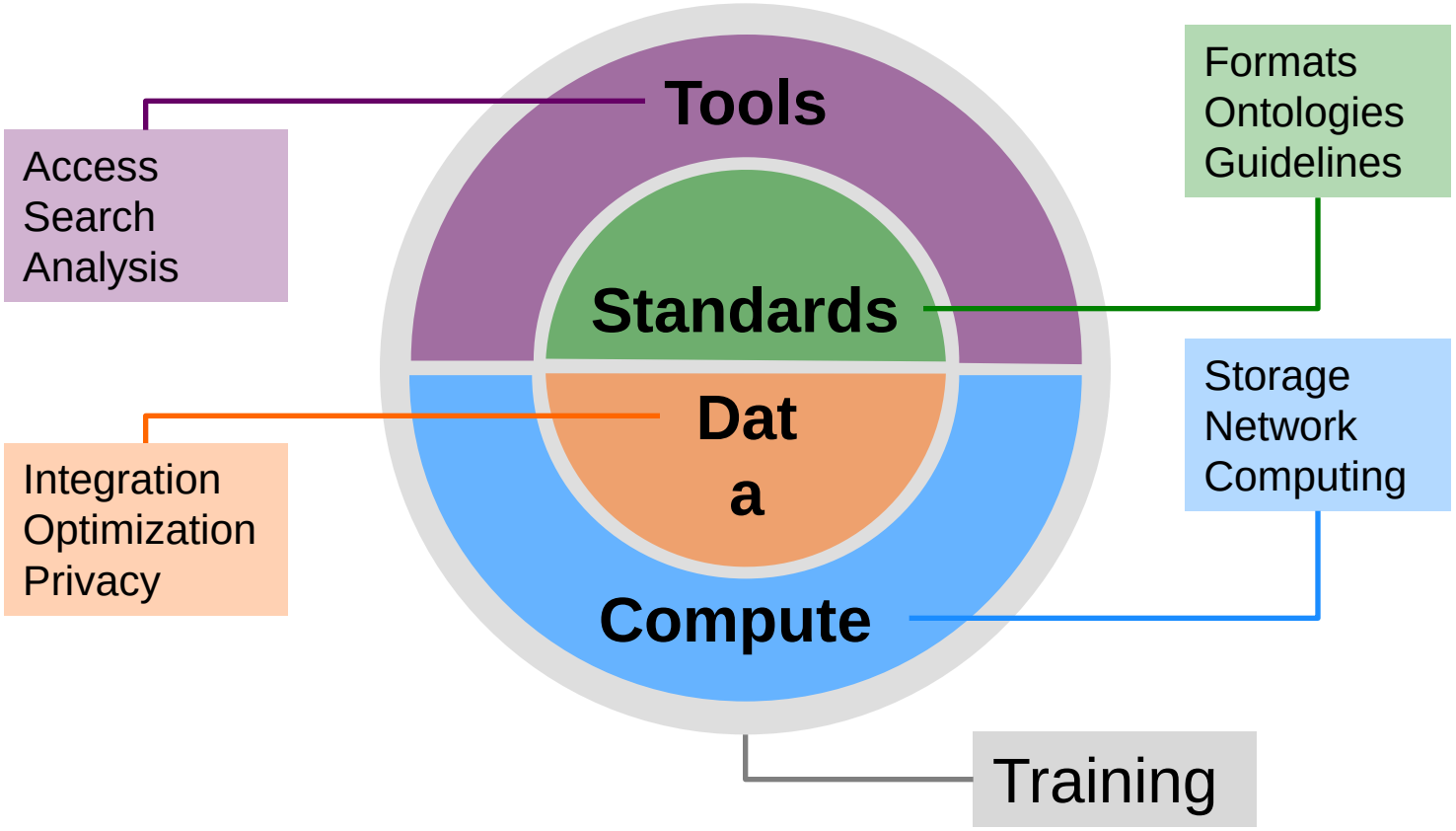
- Open access to life science data is essential for advances in many areas of research
- It provides a **valuable path to discovery**, one that in many other areas of research is limited by commercial confidentiality

- National funders increasingly require researchers to make data open
- EC's H2020 pilot on Open Research Data and Data Management Plans

The response

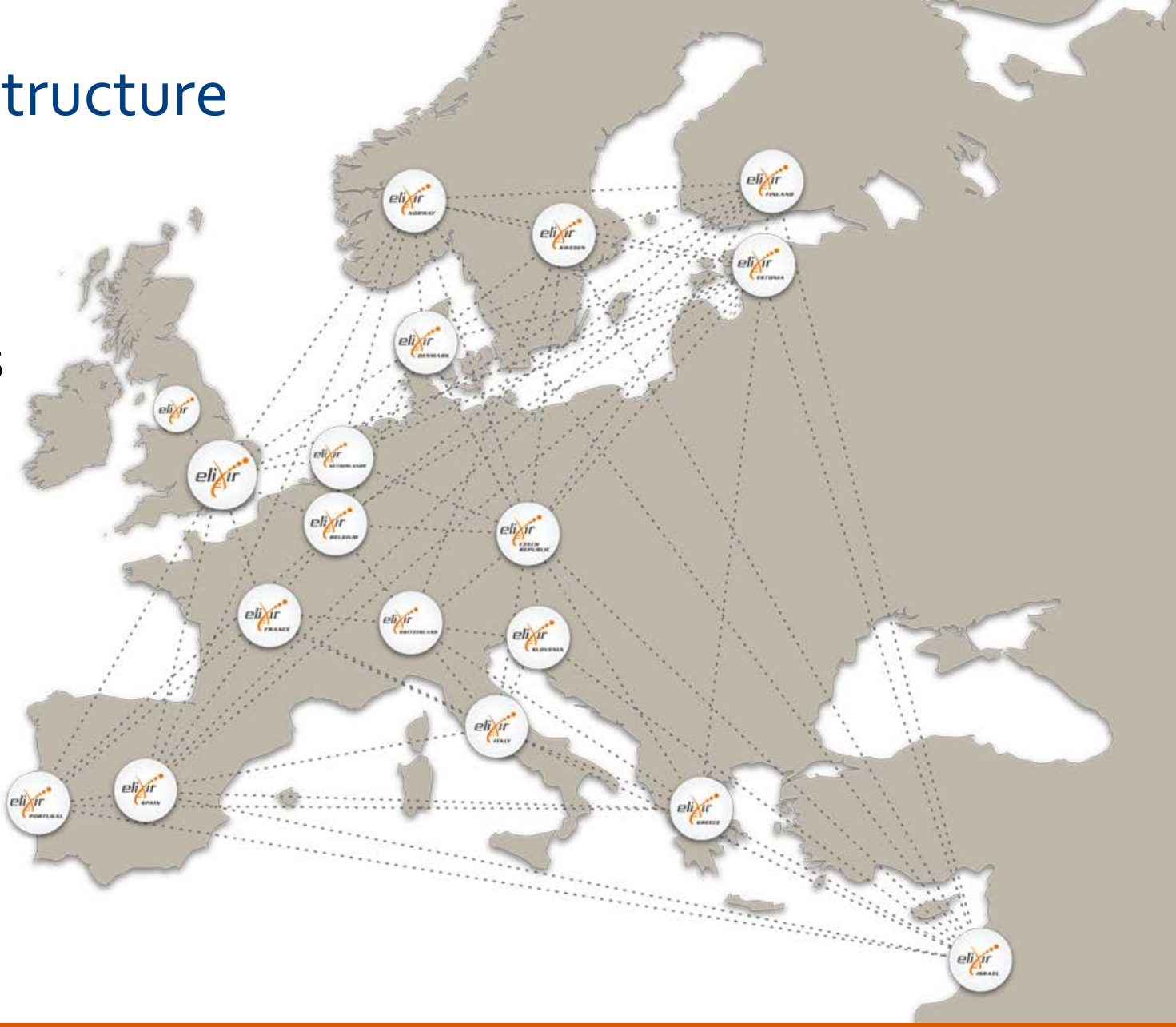
European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

Infrastructure for Life Sciences



ELIXIR's structure

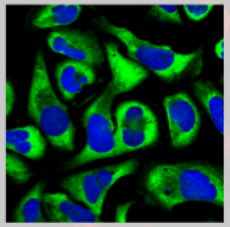
- Tools
- Standards
- Data
- Compute
- Training
- Industry

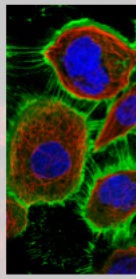


ELIXIR Nodes

ELIXIR: Swedish Node

The Swedish ELIXIR node will initially contribute with the Human Protein Atlas (http://www.proteinatlas.org).





Collaborating

Human Protein Atlas

The Swedish ELIXIR node will initially contribute with the Human Protein Atlas (http://www.proteinatlas.org), which contains: (1) Human Tissue Atlas, (2) Human Subcellular Atlas, (3) Human Cell Line Atlas, and (4) Human Cancer Atlas. Furthermore, RNA transcript data has been added for a majority of the tissues in the normal tissue atlas and the cells in the cell line atlas.

The Human Protein Atlas (HPA) programme is a scientific research programme led by Prof. Mathias Uhlén with the goal to explore the whole human proteome using an antibody-based gene-centric approach with the effort to map and characterise a representative protein for each protein-coding human gene. New data are released annually to the Human Protein Atlas.

The December 2013 release has data corresponding to 16,621 human genes (approx. 83% of the human protein-coding genes) and RNA expression data for more than 90% of the genes. We plan to extend the public service by providing four separate subparts of the Human Protein Atlas all within the framework of the ELIXIR infrastructure. Efforts within the Swedish ELIXIR node will focus on integration of the Human Protein Atlas with other ELIXIR resources. The protein atlas has over half a million visitors per year and the papers describing the work of the Human Protein Atlas have received over 1000 citations.

The Swedish ELIXIR node is coordinated by BILS (Bioinformatics Infrastructure for Life Sciences; http://bils.se).

Contact

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).

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

The Research Council of Norway

ELIXIR: Finland Node

The ELIXIR node in Finland, Biomedinfra.fi, provides compute cloud and storage resources for life sciences with integrated computational access to very large biological data resources. Services are for biomedical organisations and evolved with the development of the European e-Infrastructure (GÉANT for research network, EGI.eu and PRACE for computing and EUDAT for data). It is a flexible infrastructure component for bioinformatics service providers and can be used to host tools and build topical data service environments. It provides support to BBMRI example a secure virtualised platform for sensi- ank data.

edical data resources are collections and registries population that are being organised and digitised or Molecular Medicine Finland (FIMM) and the Na- r Health and Welfare (THL) and will become avail- . National biomedical data interoperation and in- the reference data of ELIXIR provides a use case for netic diagnostics of, for example, heart diseases, y.

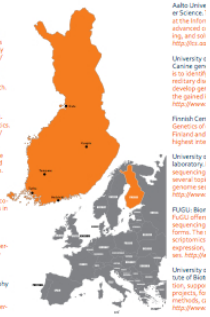
Legal, ethical and privacy requirements for enabling research on biomedical data require solutions for researcher authentication and authorisation. The Resource Entitlement Management System (REMS) enables coordination of resource access rights in collaboration between the data service providers and data access granting bodies. For user authentication, REMS relies on the technology and policies for data protection provided by the GÉANT eduGAIN inter federation service.

The ELIXIR node is operated by the Finnish data centre provider CSC– IT Center for Science under Biomedinfra.fi consortium agreement with FIMM and THL. It provides opportunities, for example, to build reference genetic data res nome sequence variation for tra

The ICT hardware is hosted by C and university network (Funet) essary. The node organizes trai

The Nordic ELIXIR nodes (Denm nate their development in collab

ng organisations



Ministry of Education and Culture

ACADEMY OF FINLAND

Biome

ELIXIR: Danish Node

The Danish ELIXIR node on bioinformatics tool interoperability and integration will address the main problems in tool utilization experienced by the life sciences community.

Collaborating organisations

Technical University of Denmark	Aarhus University	Novozymes A/S
University of Copenhagen	Lundbeck A/S	Exiqon A/S
University of Southern Denmark	Novo Nordic A/S	Glostrup Research Institute, Glostrup Hospital

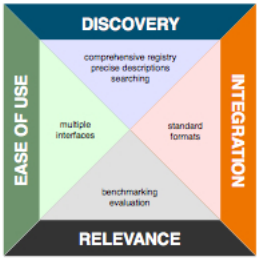
Key services in the Danish ELIXIR node:

Creation of a comprehensive tool registry

Creation of a comprehensive tool registry equipped with adequate search functionality based on the existing and emerging ontologies in the area of life sciences. This involves a comprehensive effort in tool description shared by the Node and several ELIXIR partners as well as the individual tool providers. The registry will continuously interchange content with other tool registries, existing and emerging.

Provision of tools

Provision of tools in the form preferred by users across academia and industry: the Node will encourage and assist tool providers in generation of access interfaces sought after by the community. In some cases the Node will host entire tools. This effort will involve constant monitoring of the tools area in general and the changing needs of the users specifically. The Node will actively promote programmatic access to tools and contribute to competence build-up in this area among the tool providers. The Node will also work on cloudification of tools and general mirroring in order to improve robustness.



Benchmarking, sustainability and renewal

Benchmarking, tool sustainability and renewal: the Node will coordinate the flow of information on the tool quality and relevance, and make it available to the users on regular basis. Automatic update of tools will also be promoted and the Node will actively encourage tool providers to do this.

Tool interoperability

Tool interoperability: promotion of standard data formats and schemas as well as evaluation and promotion of workflow engines and integrated workbenches. The effort will build upon the ongoing efforts in this area within the European countries behind ELIXIR.

Contact

DTU

AARHUS UNIVERSITET

bio registry tools & data services | elixir | BioMedBridges

Partners | About | Home

Genomics

Show columns | Clear filter

Name	Type	Description	Topics	Web UI	REST API
	Filters			■	■
ANNOVAR	Tool	variant annotation	Genomics	☐	☐
Arabidopsis nucleolar protein databas	Database	Comparative analysis of human and Arabidopsis nucleolar proteomes	Genome, proteome and ...	☐	☐
ArrayExpress	Database	A database of functional genomics experiments including gene expression w...	Functional genomics	☑	☑
ButterflyBase database of comparativ	Database	Gene database for butterflies and moths.	Invertebrates Comparati...	☐	☐
CDinFusion	Tool	A submission-preparation-tool for the integration of contextual data (CD) wi...		☐	☐
Comprehensive microbial genome res	Database	Information on all of the publicly available, complete prokaryotic genomes. ...	Prokaryotes and archae ...	☐	☐
Comprehensive yeast genome databa	Database	Information on the molecular structure and functional network of the entirel...	Pathways, networks and...	☐	☐
CREST	Tool	CREST (Classification Resources for Environmnetal Sequence Tags) is a colle...	Metagenetics Amplicon ...	☐	☐
Database of homologous sequences f	Database	Homologous genes from fully sequenced organisms. It allows selection of se...	Nucleic acid sequence al...	☐	☐
EBI metagenomics	Database	Pipeline for analysis of metagenome sequences	Metagenomics	☐	☐
Ensembl Genomes Bacteria	Database	Ensembl Genomes bacteria division	Genomics	☐	☐
Ensembl Genomes Fungi	Database	Ensembl Genomes fungi division	Genomics	☐	☐
Ensembl Genomes Plants	Database	Ensembl Genomes plants division	Genomics	☐	☐
Ensembl Genomes Protists	Database	Ensembl Genomes protists division	Genomics	☐	☐
Ensembl Genomes	Database	Extension of Ensembl genome database across the taxonomic space	Genomics	☐	☐
Ensembl Human	Database	Ensembl human	Genomics	☐	☐
Ensembl Metazoa	Database	Ensembl Genomes metazoa division	Genomics	☐	☐
Ensembl Mouse	Database	Ensembl mouse division	Genomics	☐	☐
Ensembl REST API service	Service endpoint	Retrieves Gene Tree dumps for a given Gene Tree stable identifier	Comparative genomics	☐	☑
Ensembl REST API service	Service endpoint	Retrieves the Gene Tree that contains the given stable identifier	Comparative genomics	☐	☑
Ensembl REST API service	Service endpoint	Retrieves a Gene Tree containing the Gene identified by the given symbol	Comparative genomics	☐	☑
Ensembl REST API service	Service endpoint	Retrieves homology information by ensembl gene id	Comparative genomics	☐	☑
Ensembl REST API service	Service endpoint	Retrieves homology information by symbol	Comparative genomics	☐	☑
Ensembl REST API service	Service endpoint	Lists all available comparative genomics databases and their data release	Information	☐	☑
Ensembl REST API	Service	Ensembl REST API endpoints	Genomics	☐	☑

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

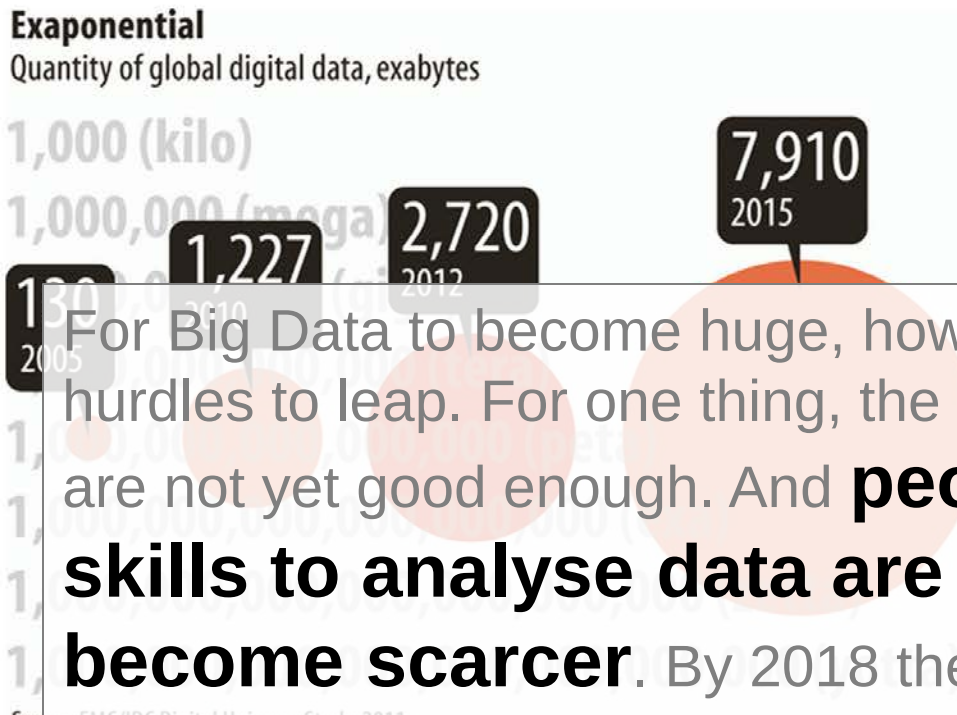
Welcome to the yotta world

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



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

Poland's agriculture: A golden age for

The enigma of flight 370: Dashed hopes

Newsbook - Mar 24th, 17:56

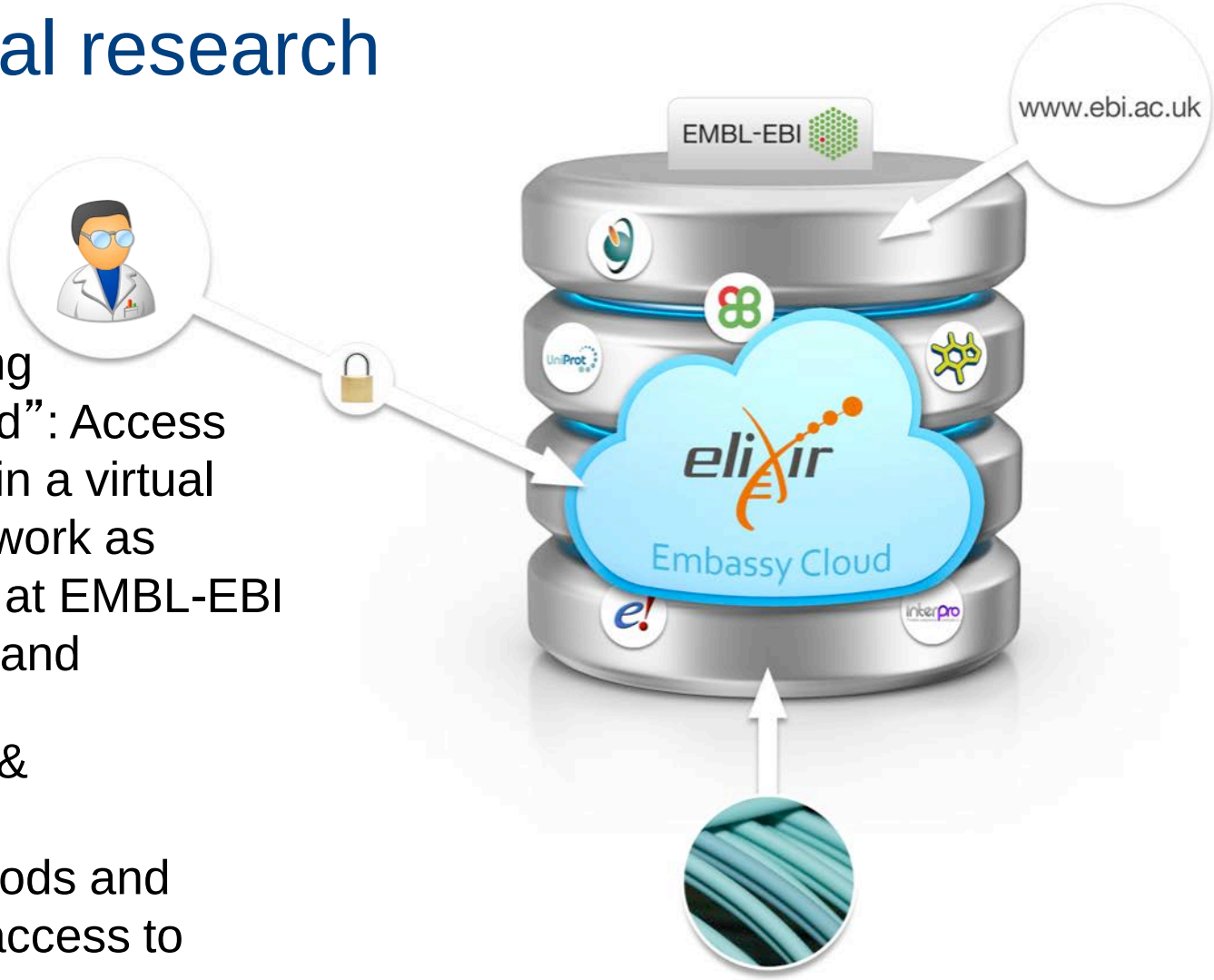
Feedback

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, …

ELIXIR pilots addressed key challenges in biomedical research

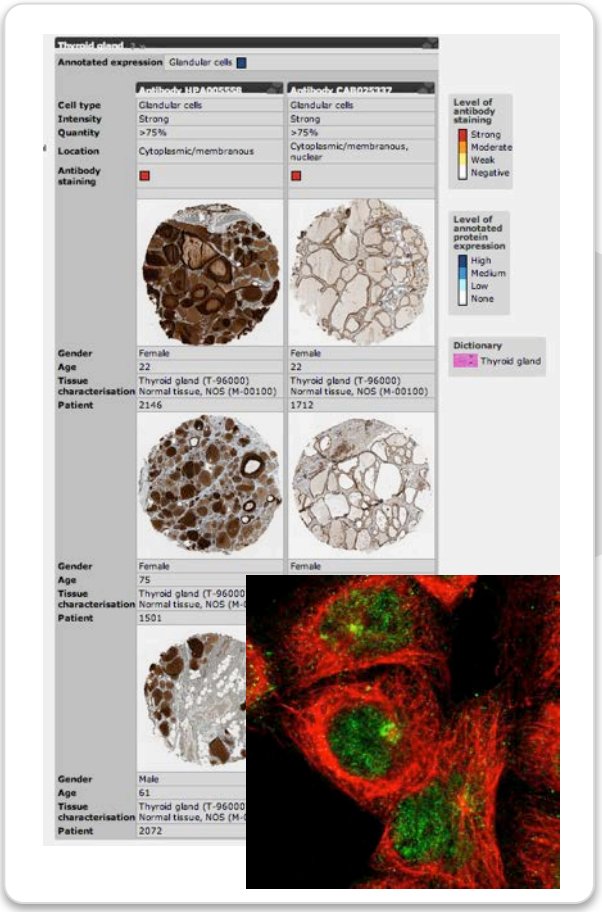
1. Cloud computing "Embassy cloud": Access reference data in a virtual environment – work as though you are at EMBL-EBI or SIB, Switzerland

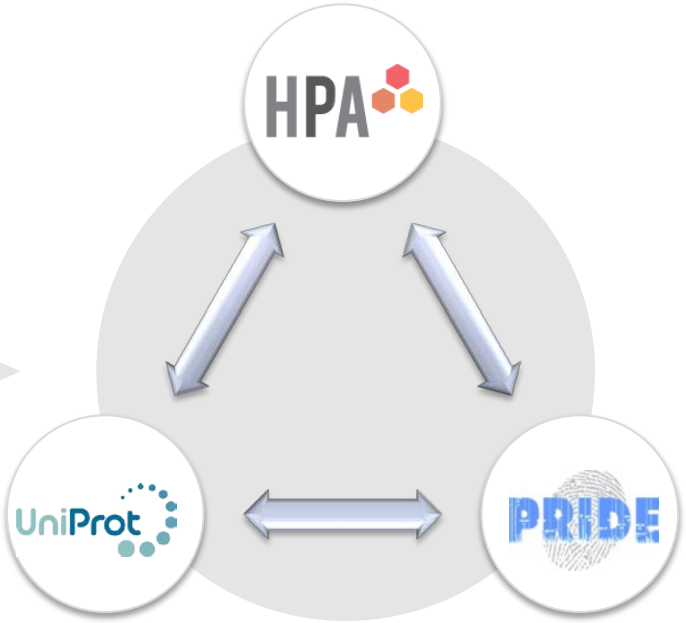
2. Authentication & Authorisation
   Improved methods and processes for access to clinical data



Identifying new drug targets

ELIXIR pilot: Interoperability of high-resolution protein data at EMBL-EBI and HPA, Sweden

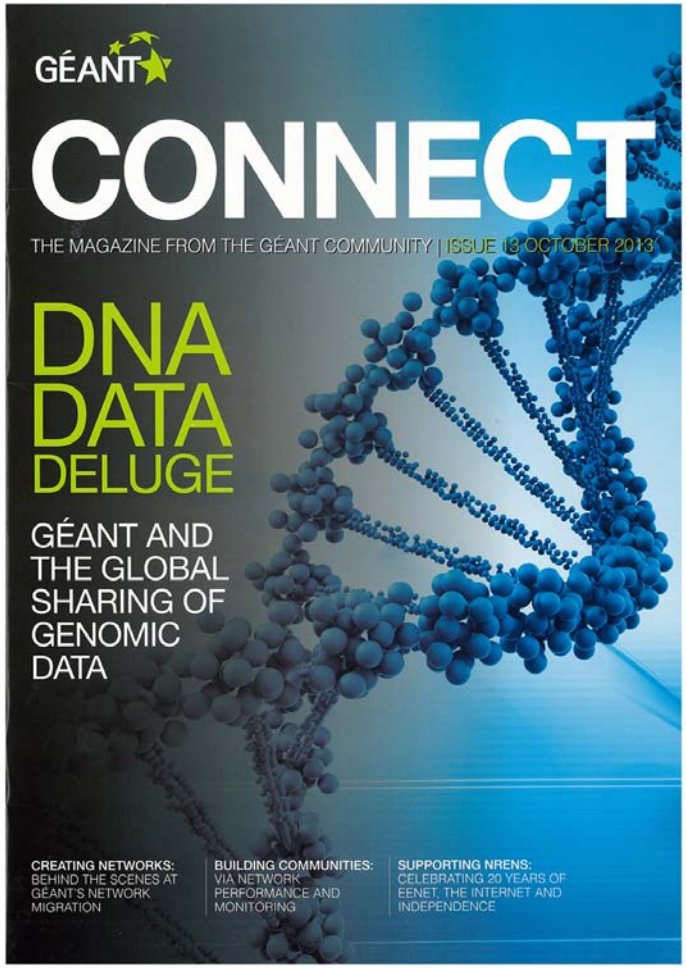




The Human Protein Atlas portal is a publicly available database with millions of high-resolution images showing the spatial distribution of proteins in 46 different normal human tissues and 20 different cancer types, as well as 47 different human cell lines.

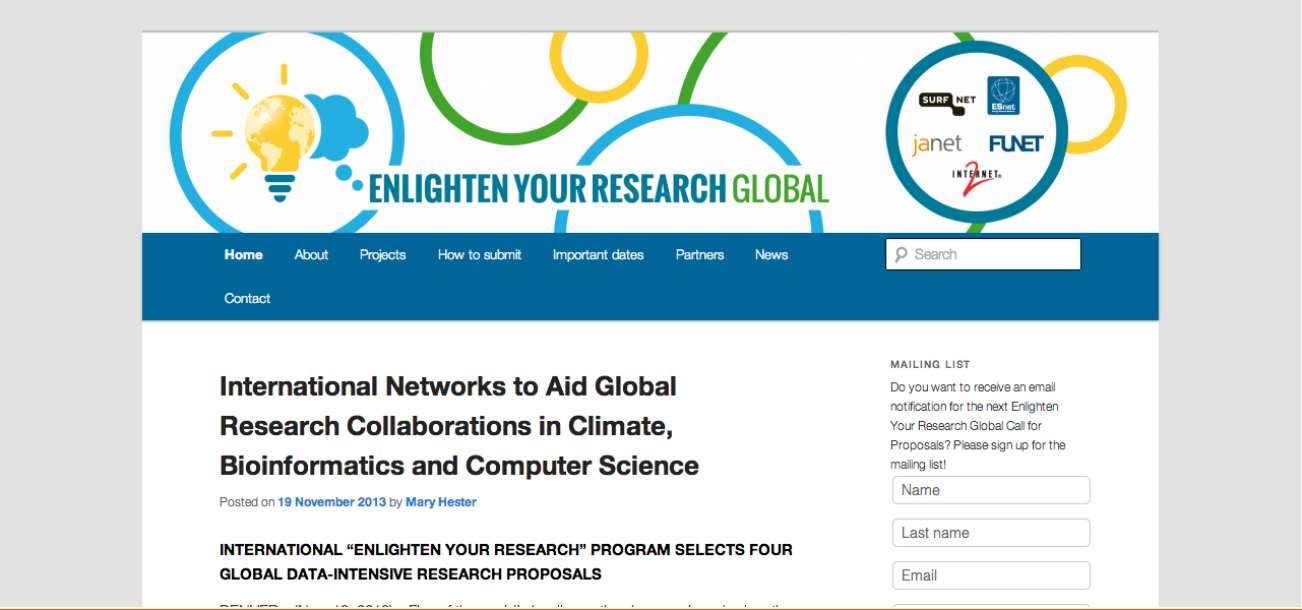
European ELIXIR Data - "LightPath" (EBI / CSC)

- 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
- 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 life sciences in Europe

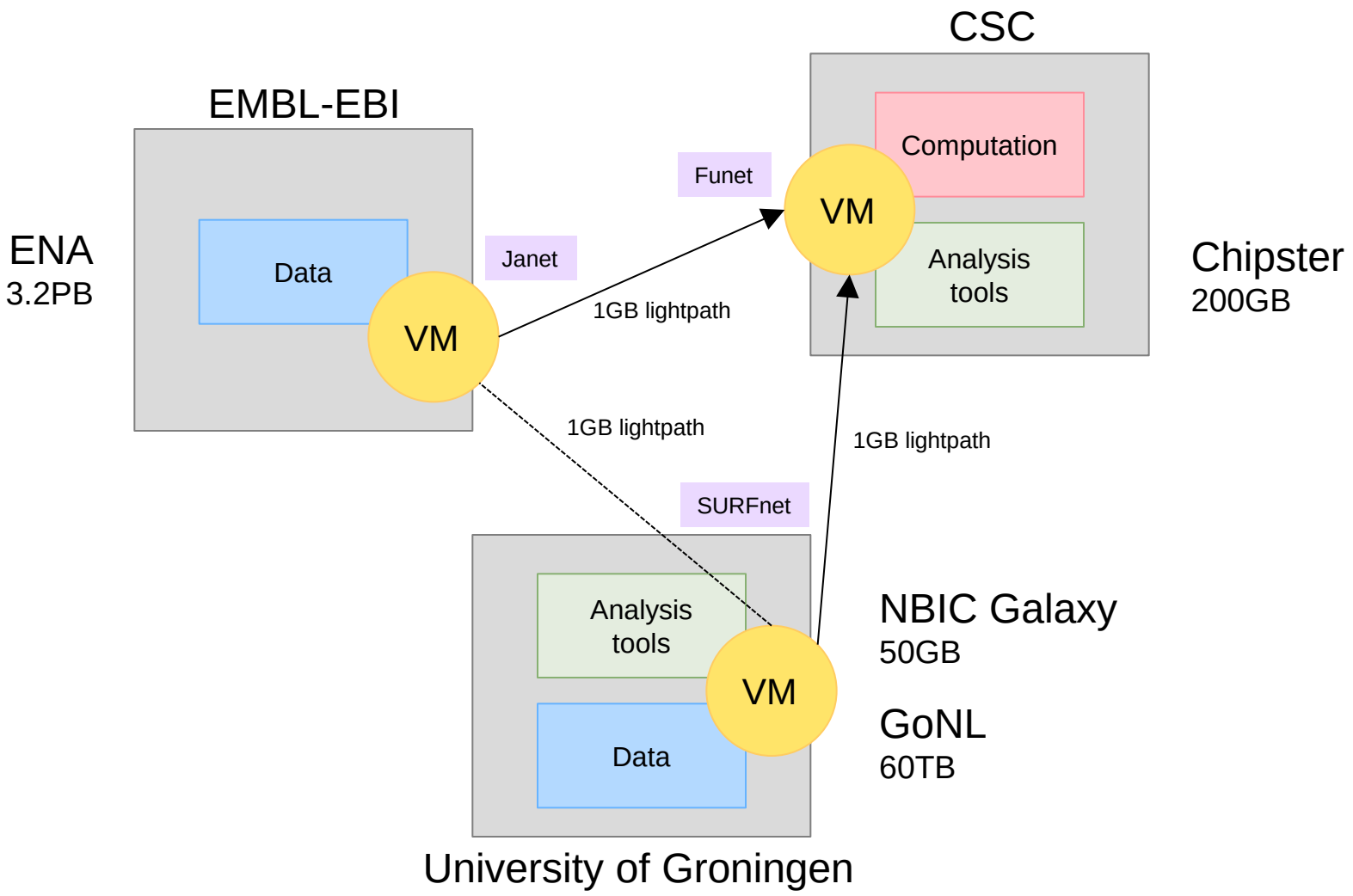


Cross-site VM Operation - pilot

- 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



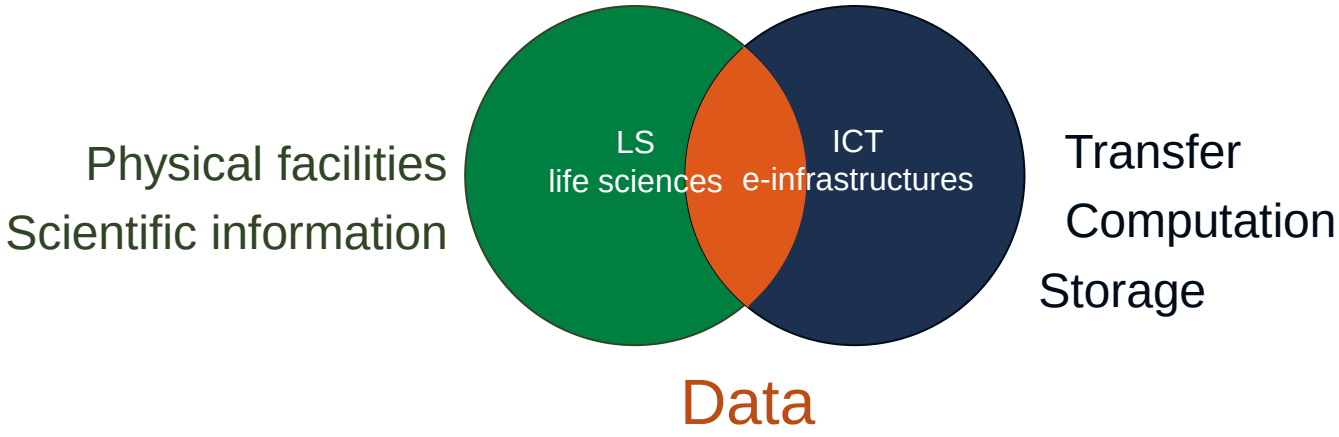
Cross-site VM Operation



European Research Infrastructures

Knowledge exchange workshop

- Discussion of big data challenges in life sciences
  - Focus on few representative domains
  - Looking 5 years ahead
- Jointly identify potential solutions to our problems



Thank you

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