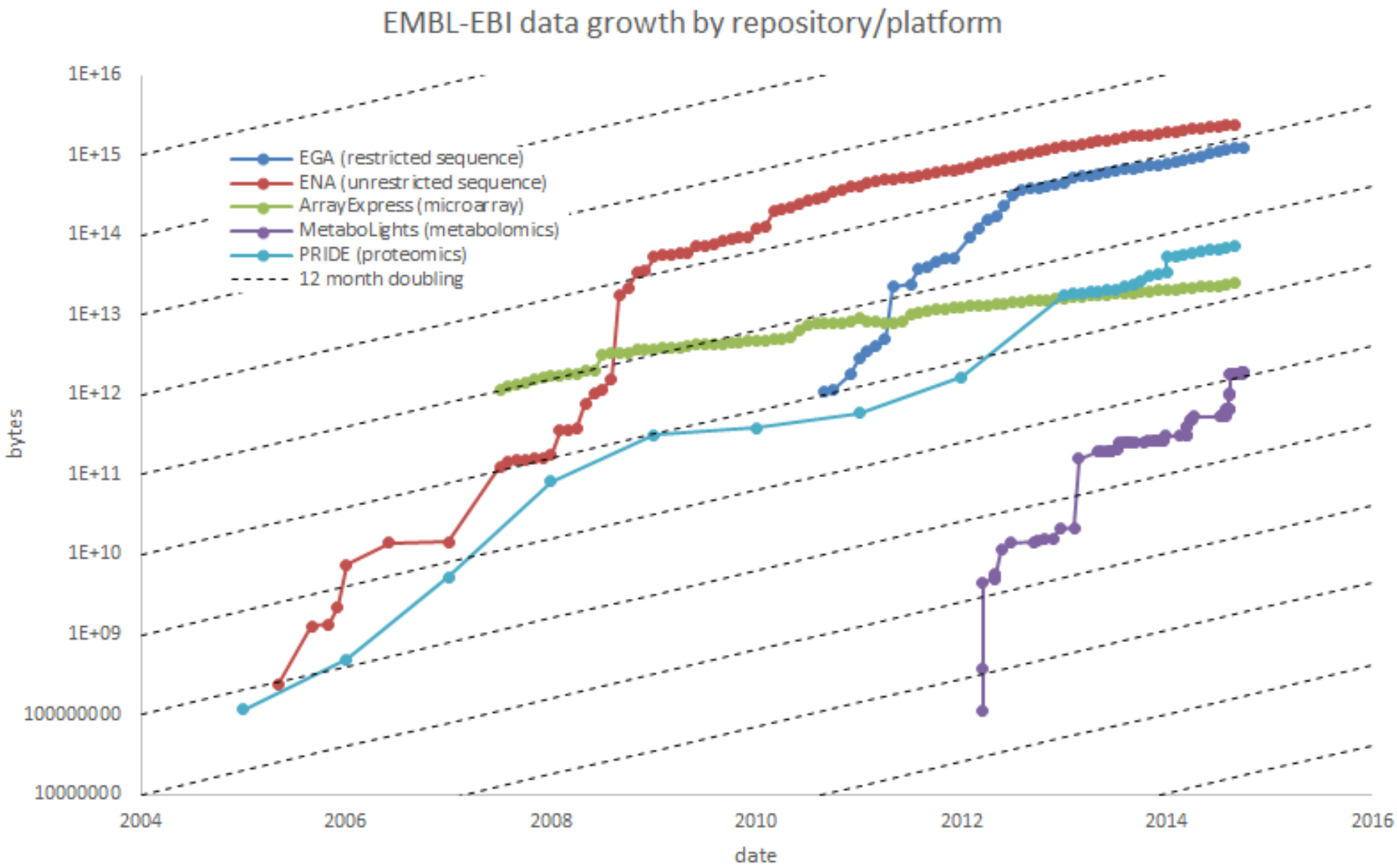
Good practices in enabling the re-use of research

Andy Smith

RECODE conference, Athens, 15 January 2015

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

Data growth in the life sciences



Data resources in life science

- Genomics Databases (non-vertebrate) (17.9%)
- Protein sequence databases (12.9%)
- Human Genes and Diseases (9.8%)
- Structure Databases (9.7%)
- Metabolic and Signaling Pathways (9.3%)
- Nucleotide Sequence Databases (8.8%)
- Human and other Vertebrate Genomes (7.1%)
- Plant databases (7.1%)
- RNA sequence databases (4.9%)
- Microarray and other Gene Expression Databases (4.5%)
- Other Molecular Biology Databases (3.3%)
- Immunological databases (1.8%)
- Organelle databases (1.6%)
- Proteomics Resources (1.2%)
- Cell biology (0.2%)

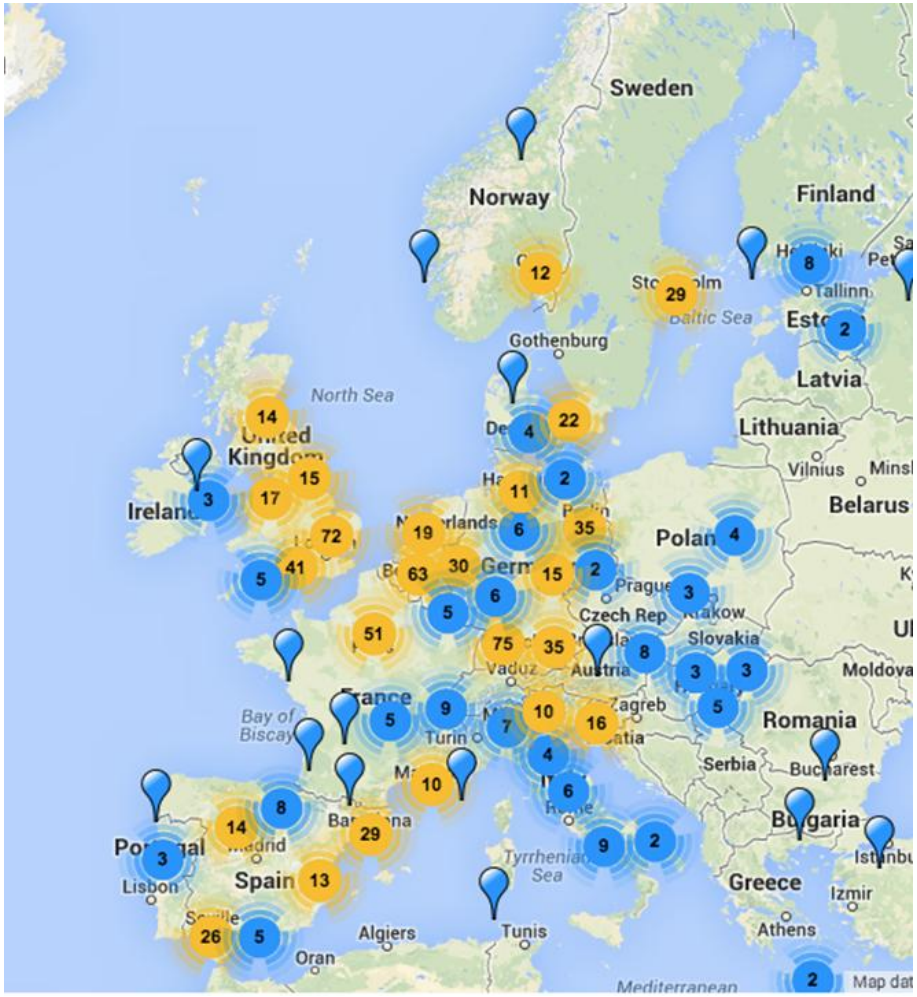
molecular biology data resources

~1800

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

The data challenge: Geographic spread

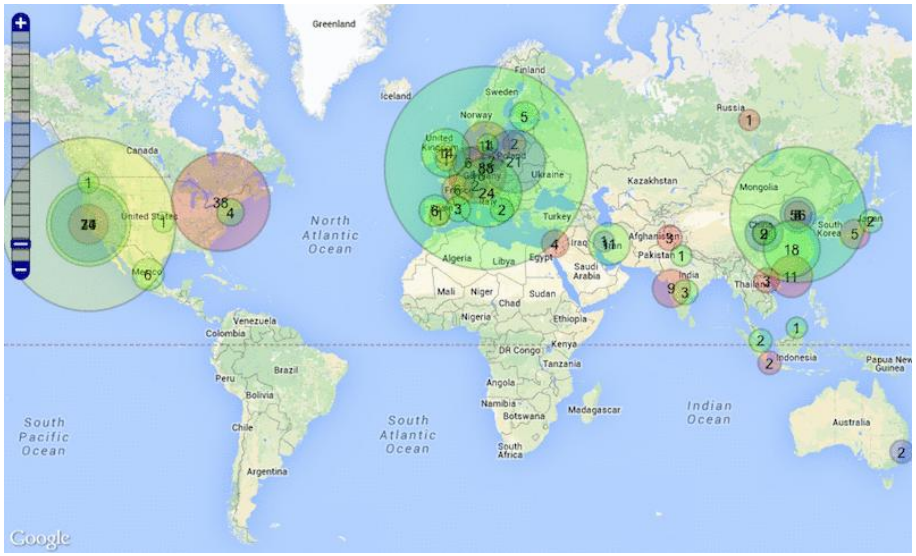
- Data production sites increasing across Europe
- European Illumina sales up 20% 2014



Source: http://omicsmaps.com

Global data users

- 8 million requests a day on EMBL-EBI website

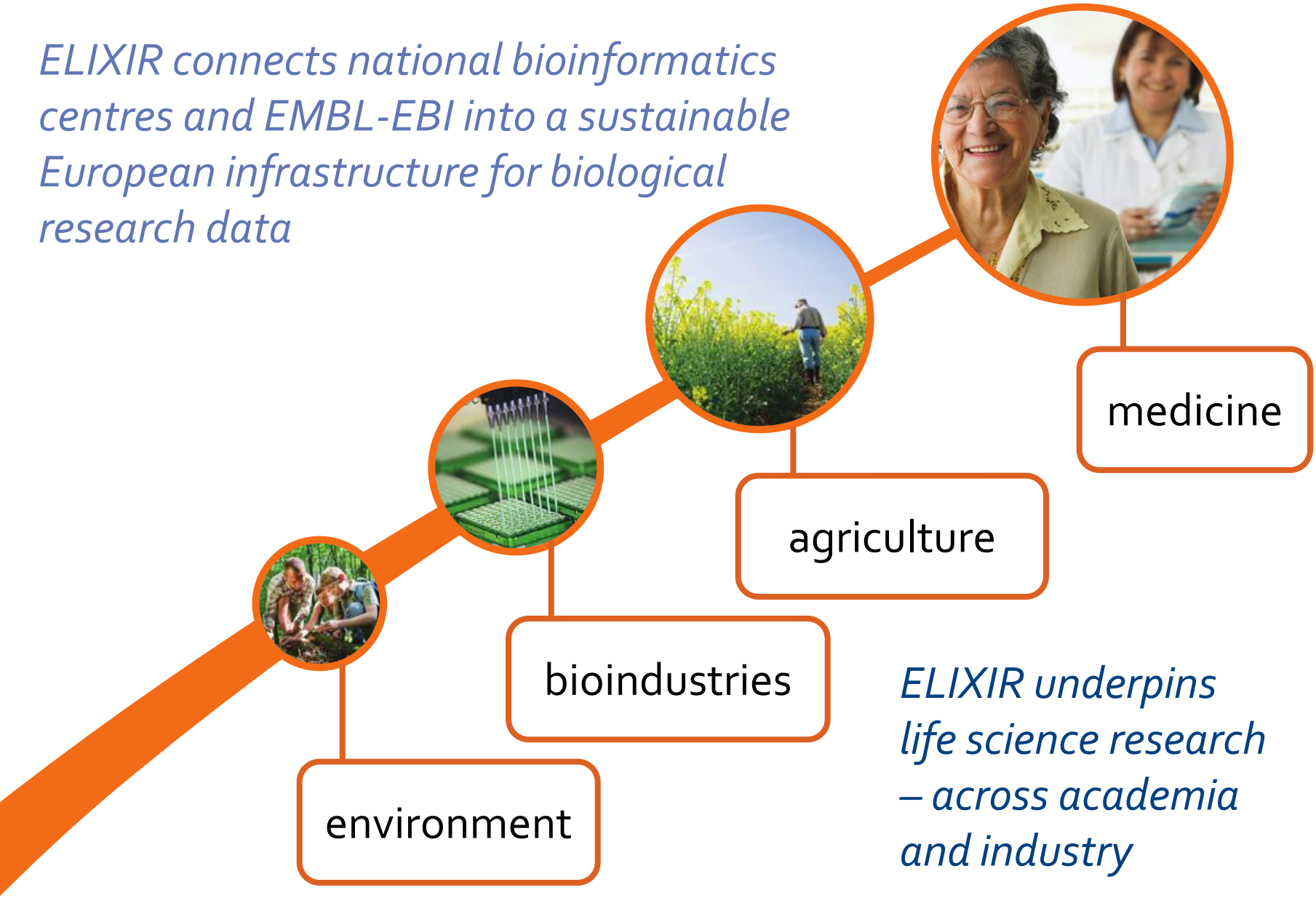


- UniProt has 800,000 daily requests
- Human Protein Atlas - over 1,000 citations globally and more than 750,000 visits during 2013, 60% from outside of Europe

ELIXIR's response

www.elixir-europe.org

ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data



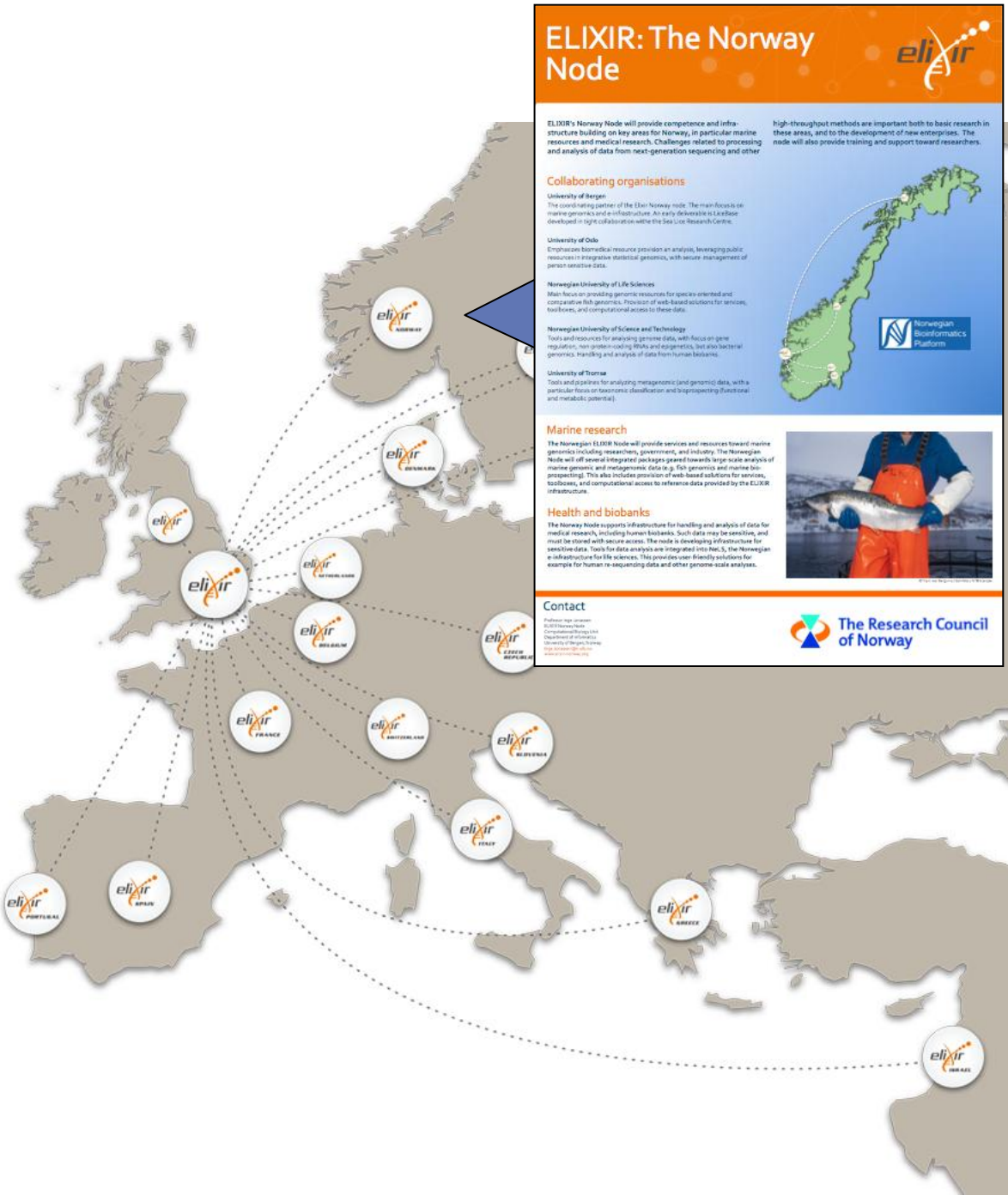
ELIXIR underpins life science research – across academia and industry

ELIXIR's activities

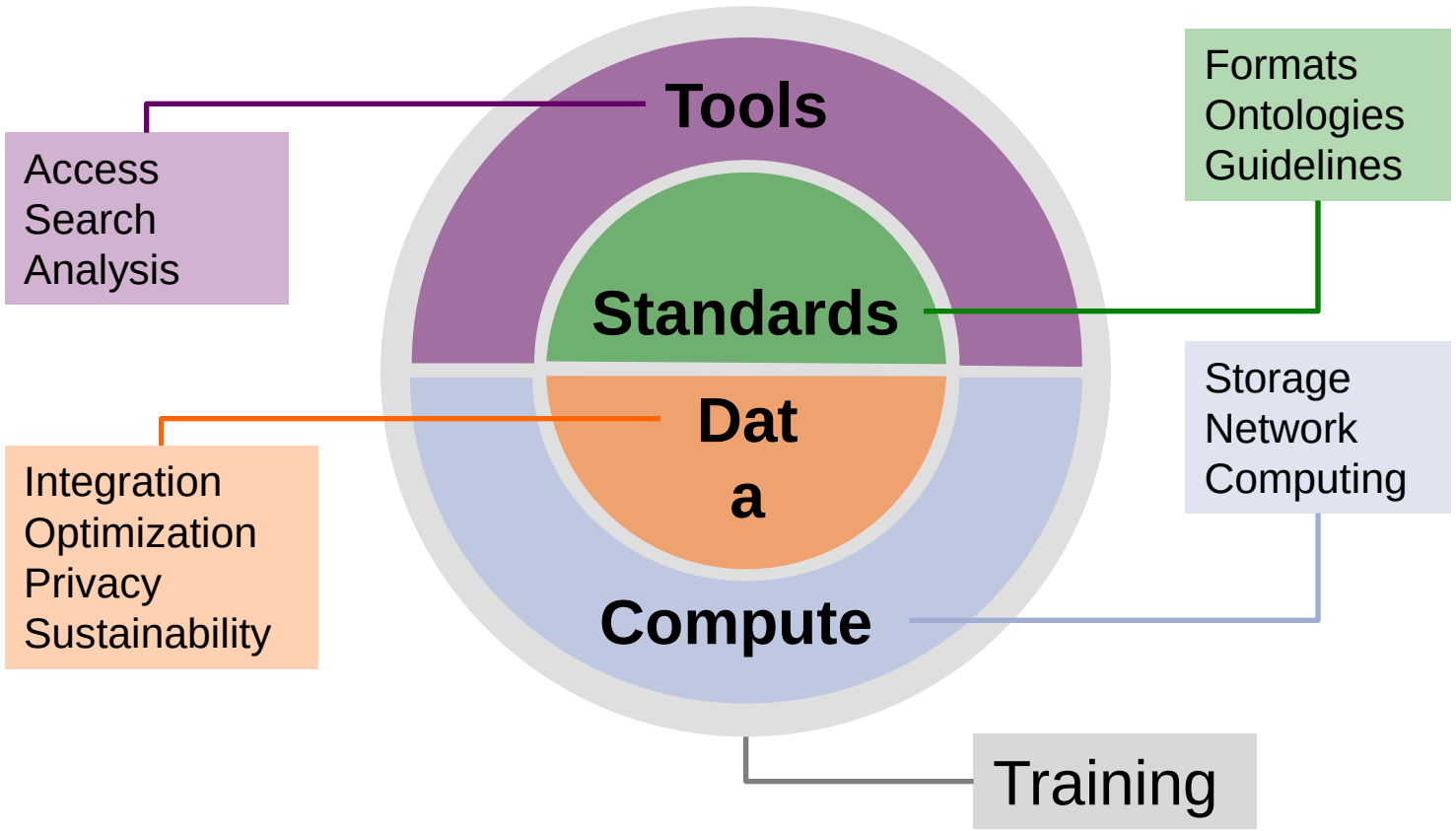
ELIXIR is the data infrastructure for Europe's life science research

ELIXIR Nodes build local bioinformatics capacity throughout Europe

ELIXIR Nodes build on national strengths and priorities



ELIXIR activities



ELIXIR's data activities

- Identifying the Core Data resources (focus of sustainability efforts)
- ELIXIR Pilot Projects
  - EGA as a joint venture
  - AAI
  - Embassy cloud

- Data management policies
  - Principles of data management and sharing at European Research Infrastructures https://zenodo.org/record/8304#.VLdRPxZAubB

ELIXIR tools activities

- Discovery Portal will facilitate access to and analysis of the data

bio registry tools & data services | elixir | BioMedBridges

Partners | About | Home

Type search terms he

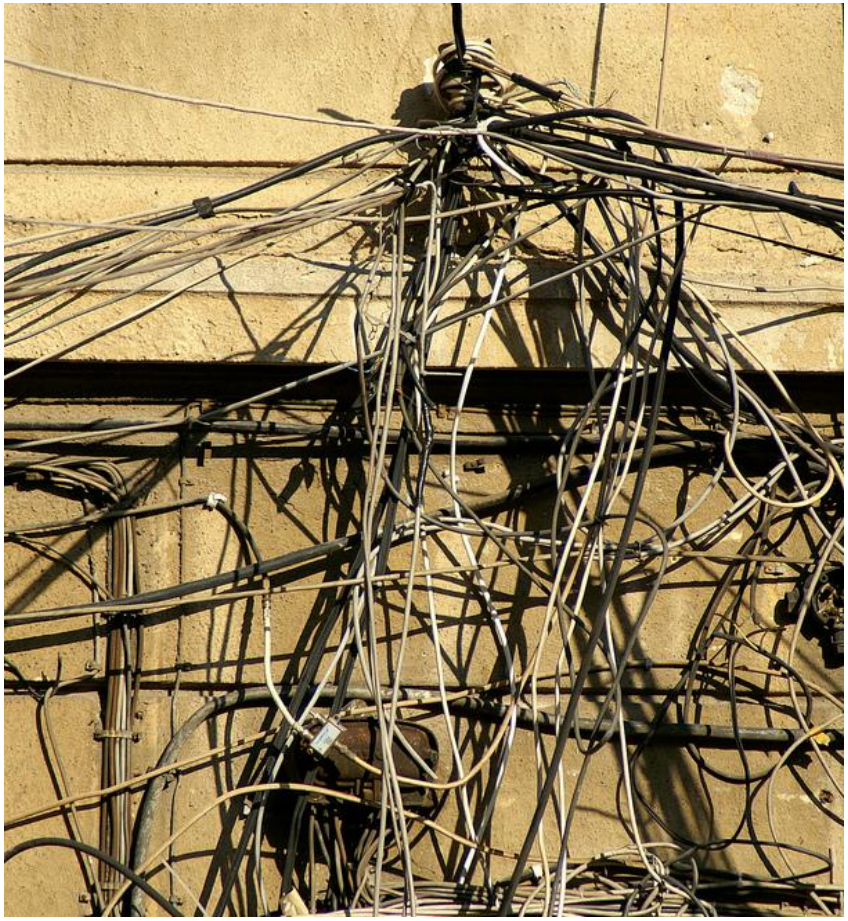
Show columns | Clear filter

Name	Type	Description	Topics	Web UI	REST A	Funding
	Filters			■	■	BMB
xnat.bigr.nl	Tool	TBD. Will serve the data that is currently only available for download from ergocar.bigr...		☐	☑	BMB
BioSD (BioSampl	Database	Unified storage and access to biological sample information		☑	☑	BMB
PiMS	Tool	A Laboratory Information Management System (LIMS) developed to support the unpredi...	Protein production LIMS ELN	☐	☑	BBSRC | CCP4 | BMB
UniChem	Database	Rapid cross-referencing of chemical structures, and their identifiers, between databases	Data search and retrieval Cheminformatics Ch...	☑	☑	BMB
xnat.bmia.nl	Tool	TBD. set up in collaboration with CTMM TraIT, which represents EATRIS/BBMRI in BioMe...		☐	☐	BMB
Tools Registry	Database			☐	☐	BMB
Mapping and regi	Database			☐	☐	BMB
Legal Tool TBN	Tool			☐	☐	BMB
ESFRI BMS Onlin	Database			☐	☐	BMB
MousePhenotype	Database	International Mouse Phenotyping Consortium portal: genotype-phenotype annotations fr...	Systemic phenotyping Genotype and phenotype	☑	☐	BMB
Ontology Tree Vi	Tool			☐	☐	BMB
MetalS2	Tool	Allows two metal-binding sites to be structurally aligned, independently of protein fold	Bioinorganic chemistry Structure comparison	☑	☐	BMB
WebMicroscope	Package	Platform for virtual microscopy		☑	☐	BMB
Shape Matching T	Tool	Searching and annotation of volume data from structural biology	Structural biology	☑	☐	BMB
PIMS	Tool	Laboratory Information Management System (LIMS) developed to support the workflows...		☑	☐	BMB
MOLGENIS compu	Tool	Workflow tool useful for core facilities to create pipelines that run on PBS or SGE cluster...	Data handling	☐	☐	BMB
MOLGENIS	Framework	For biologists MOLGENIS is a suite of web databases for genotype, phenotype, QTL and ...	Data management Data search and retrieval D...	☑	☐	BMB
MetalPDB	Database	Knowledge on metal sites in biological macromolecules, built on structural information in...	Bioinorganic chemistry Search and Retrieval	☐	☐	BMB
MaxOcc	Tool	Numerical assessments about the maximum percent of time that a conformer of a flexib...	Protein dynamics	☑	☐	BMB
FindGeo	Tool	Determine metal coordination geometry	Bioinorganic chemistry	☑	☐	BMB
Clinical Trials Inf	Tool		Data search and retrieval	☐	☐	BMB
BBMRI-NL catalog	Database	Database of collections of biomaterial and associated data subsumed under the umbrell...	Biobank Data search and retrieval	☐	☐	BMB
AMPS-NMR	Tool	NMR protein structure refinement	Structure determination Protein structure assig...	☑	☐	BMB
Phenotator	Tool	Use this tool to enter phenotypes descriptions and annotate these description with ontolo...	Annotation	☑	☐	BMB
GenoBridge	Tool	TBD Natalie is working on this as of late September 2013	Genotype phenotype	☐	☐	BMB

elixir

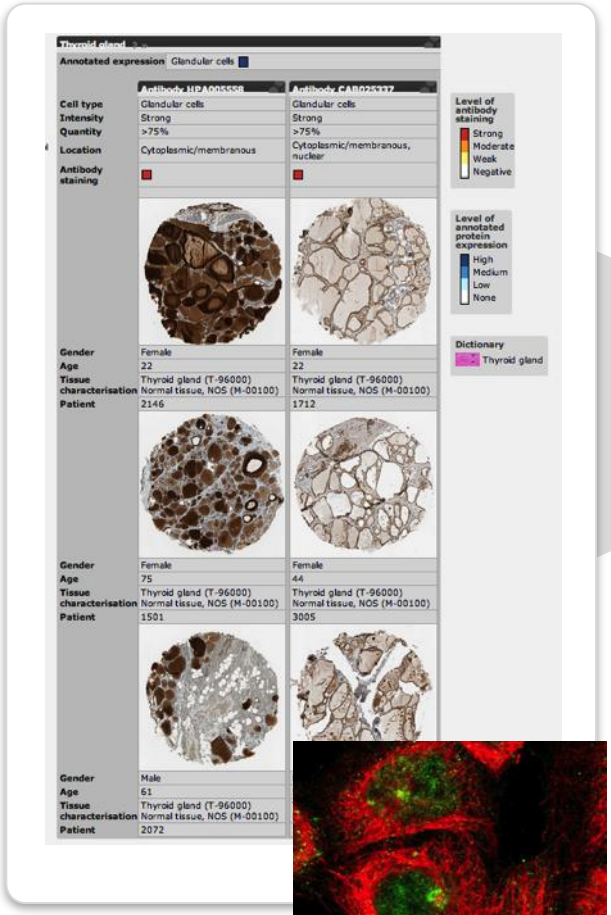
Interoperability and standards (FAIR)

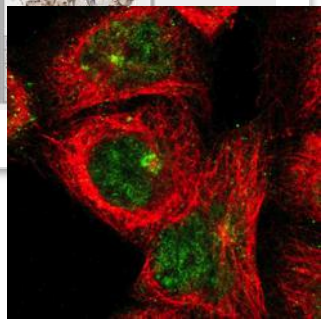
"… typically 40% of our effort in biomarker discovery is data integration"
J. Woijcek, CEO, Quartz Bio

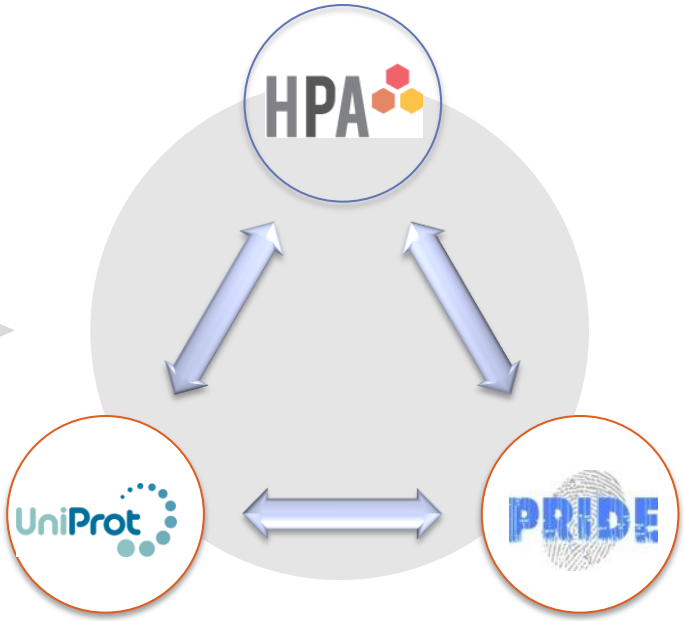


- Formats, Ontologies, Guidelines, BYOD
- International collaboration: ELIXIR NIH B2DK workshop on DOIs

Data interoperability – Human Protein Atlas







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.

Open Data as a driver of innovation

www.elixir-europe.org

elixir

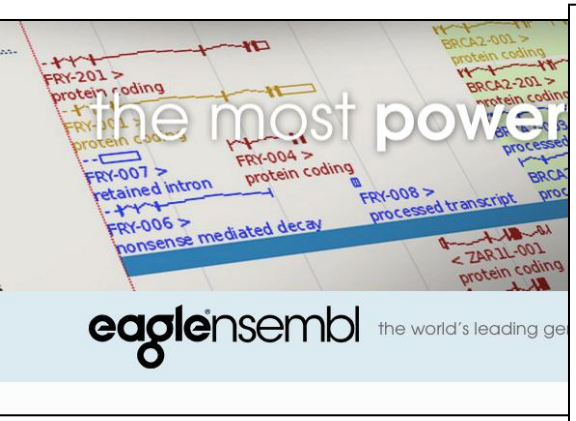
The value of Open Data

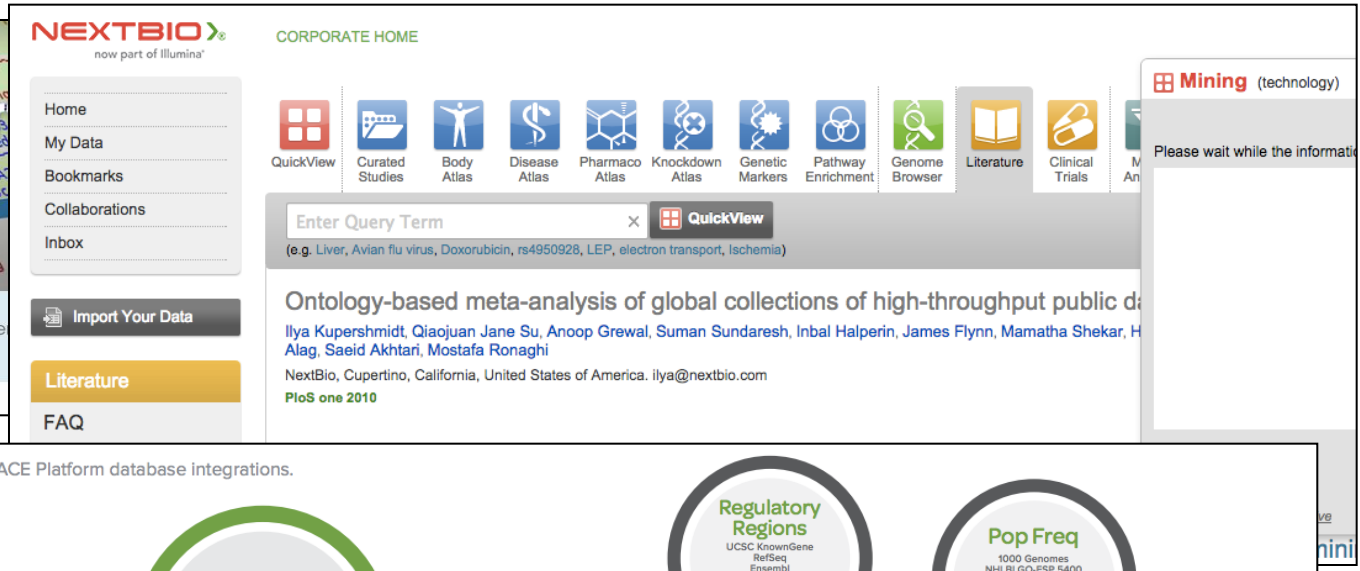
- Open access to life science data is essential for advances in life sciences:
  - understanding plant genomes in order to identify drought-, salt- and pest-resistant species
  - identifying patterns of genes that are active in different tumours
  - tracking transmission of diseases such as MRSA by identifying small variations in DNA sequence

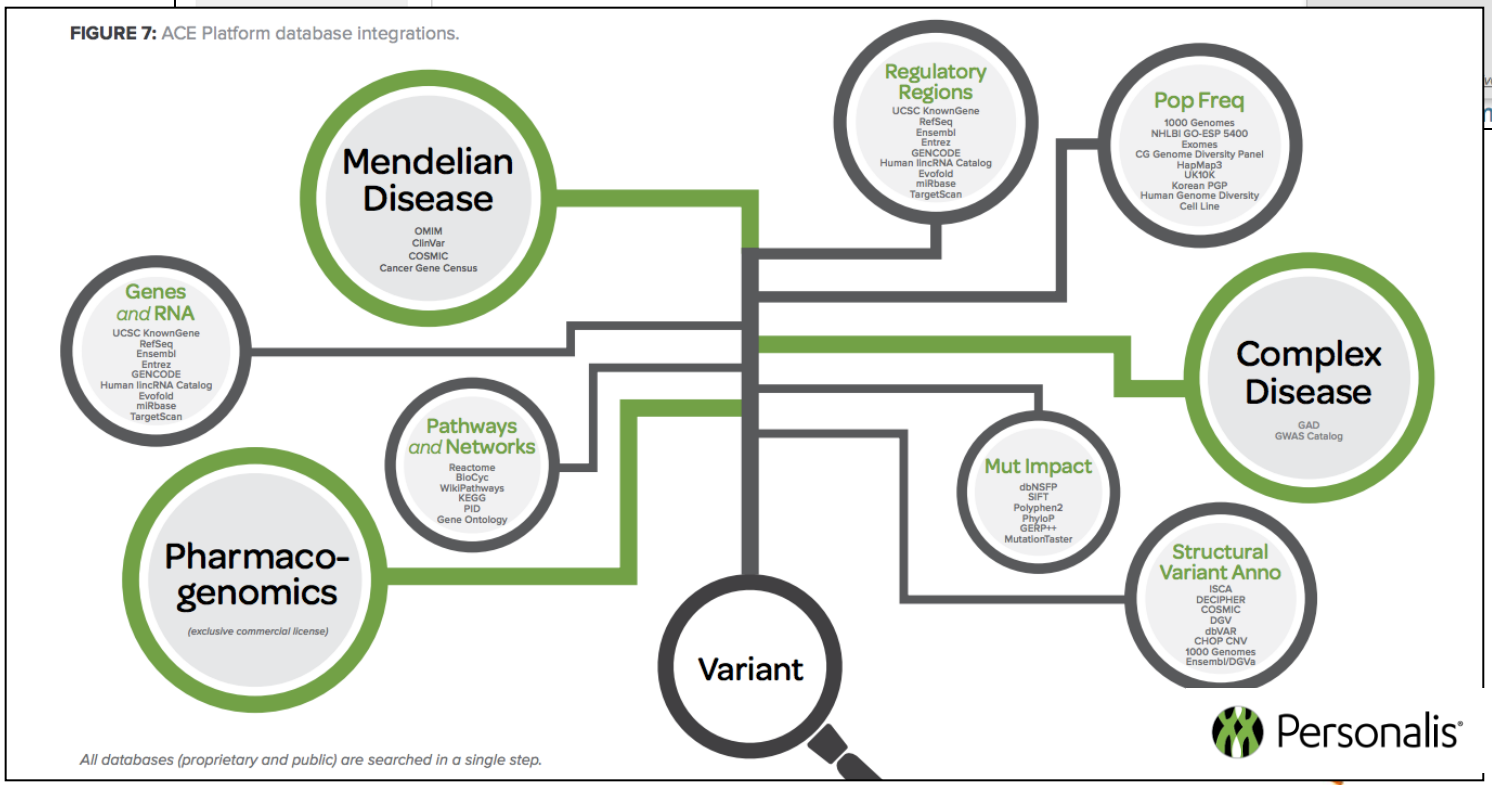
Value to industry

- Industry requires reliable, sustainable public data infrastructures
- 110 million hits from industry on EBI website; pharma, diagnostics, agri-food…
- Patents from public archives in 2014
  - 590 quoting UniProt
  - 197 quoting Ensembl
- New forms of Public Private Partnerships are being developed - CTTV
- Targeted support to SMEs

Public data: a foundation for innovation







Thank you!

Andy Smith

RECODE conference, Athens, 15 January 2015

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org