ELIXIR : Infrastructure for life science data

Niklas Blomberg

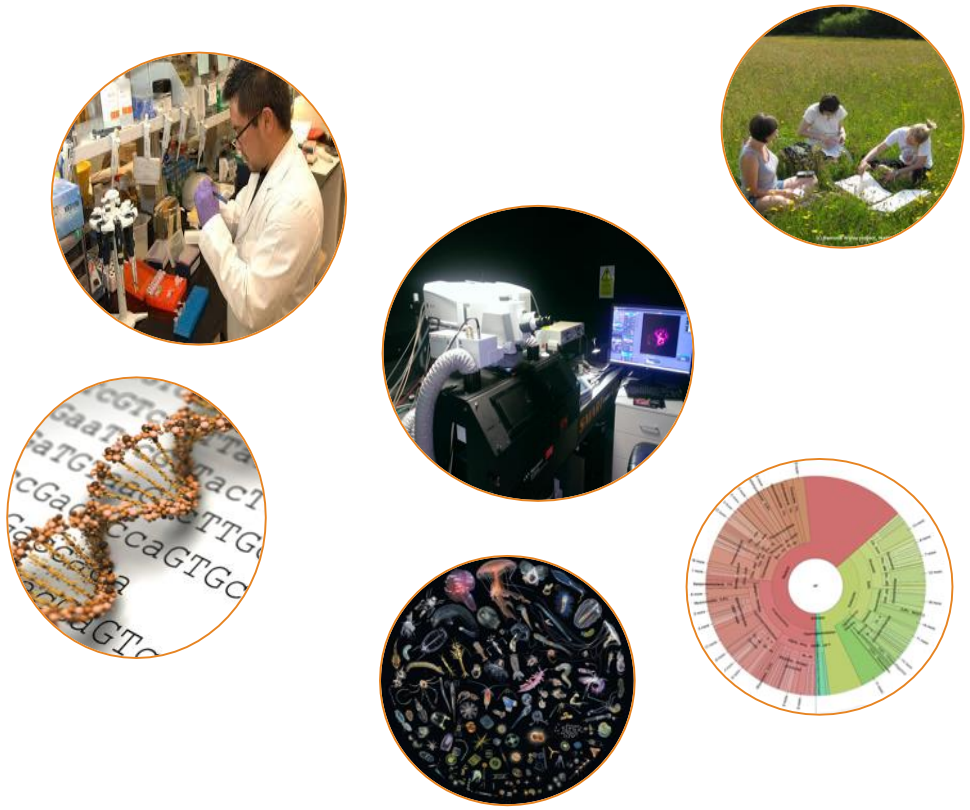
ELIXIR Greece Launch

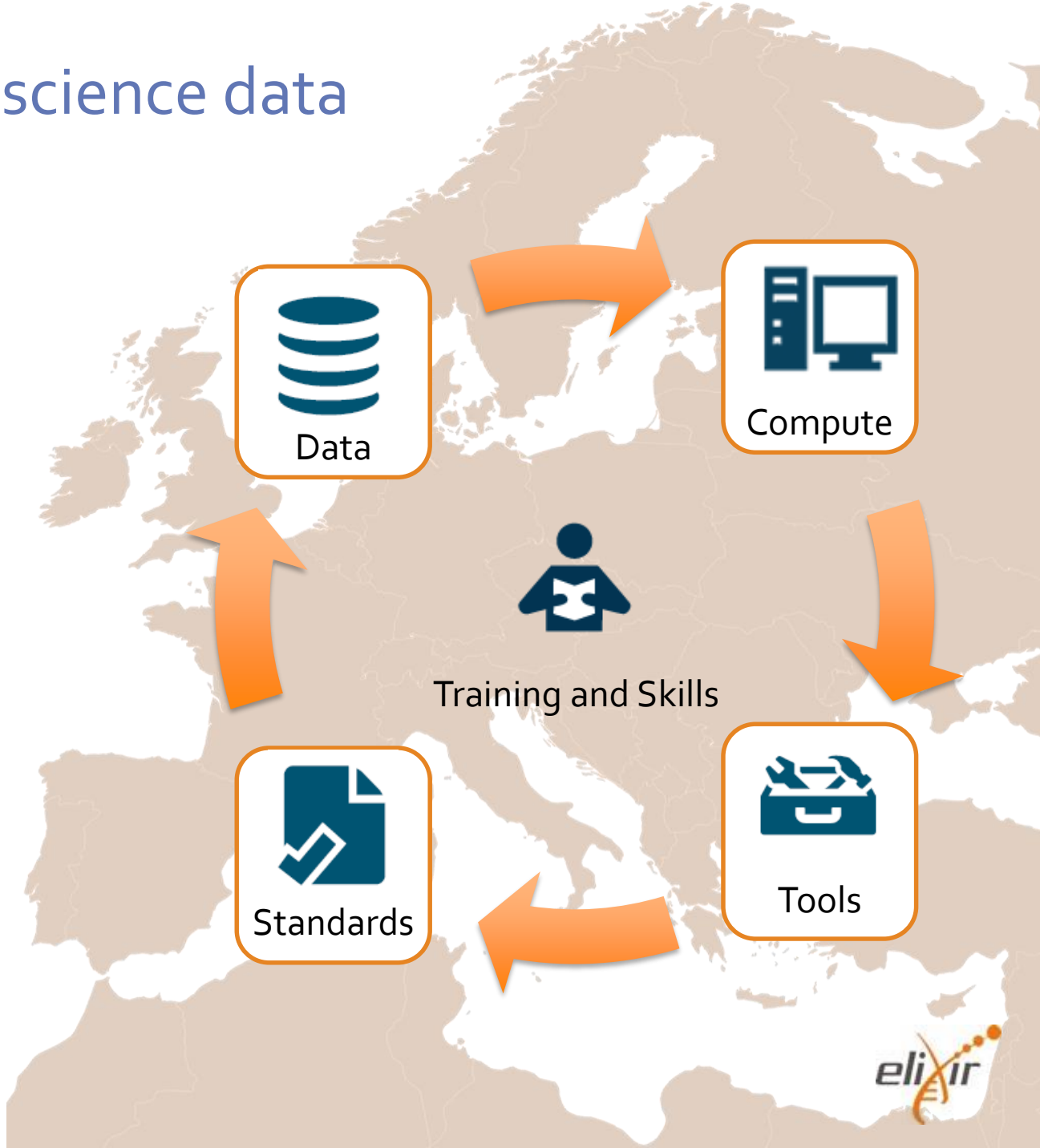
www.elixir-europe.org

Data challenges & Opportunities

- Geographically distributed data production
- Secure access and governance of human data
- Open data mandates of National and European funders

European Infrastructure for life-science data

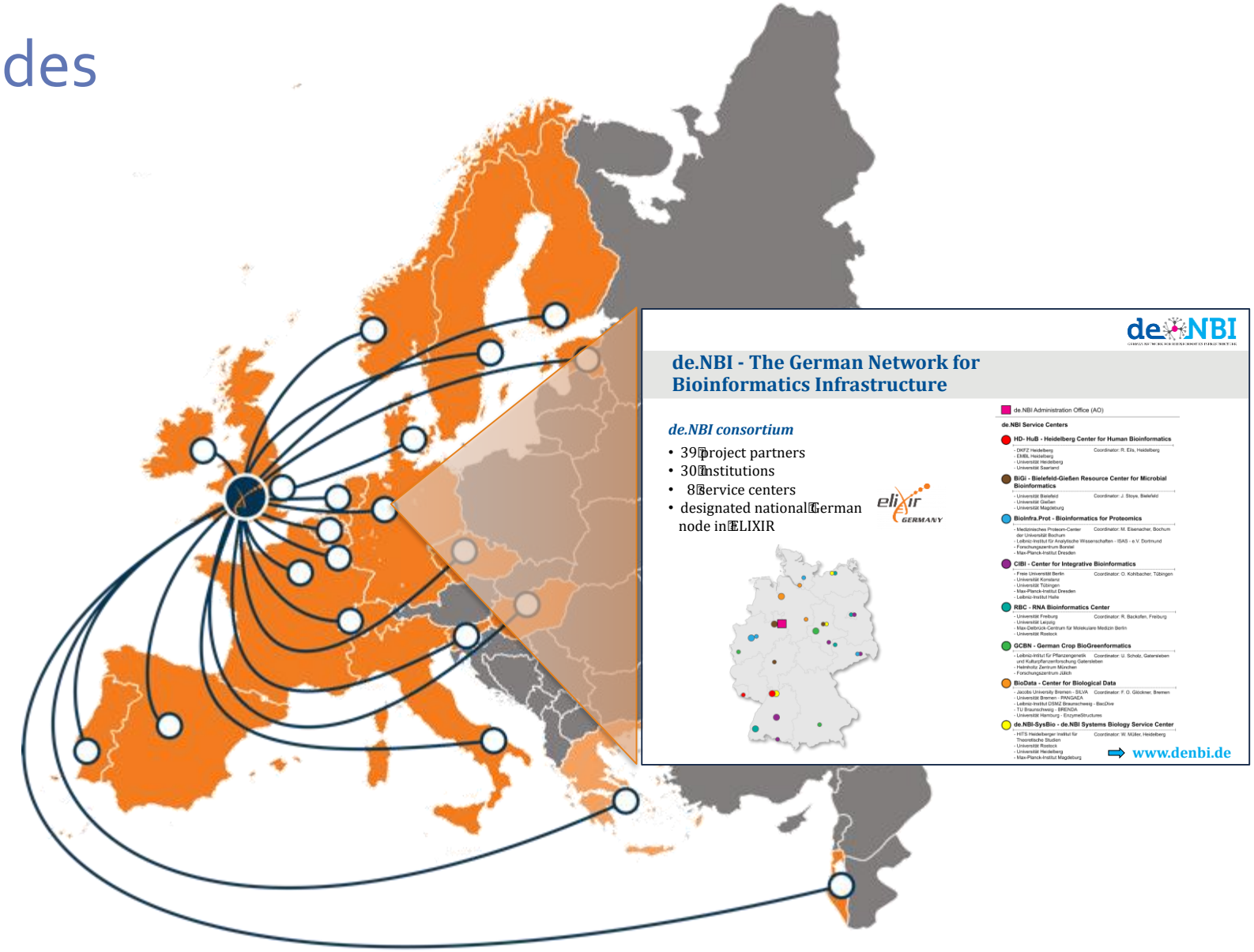




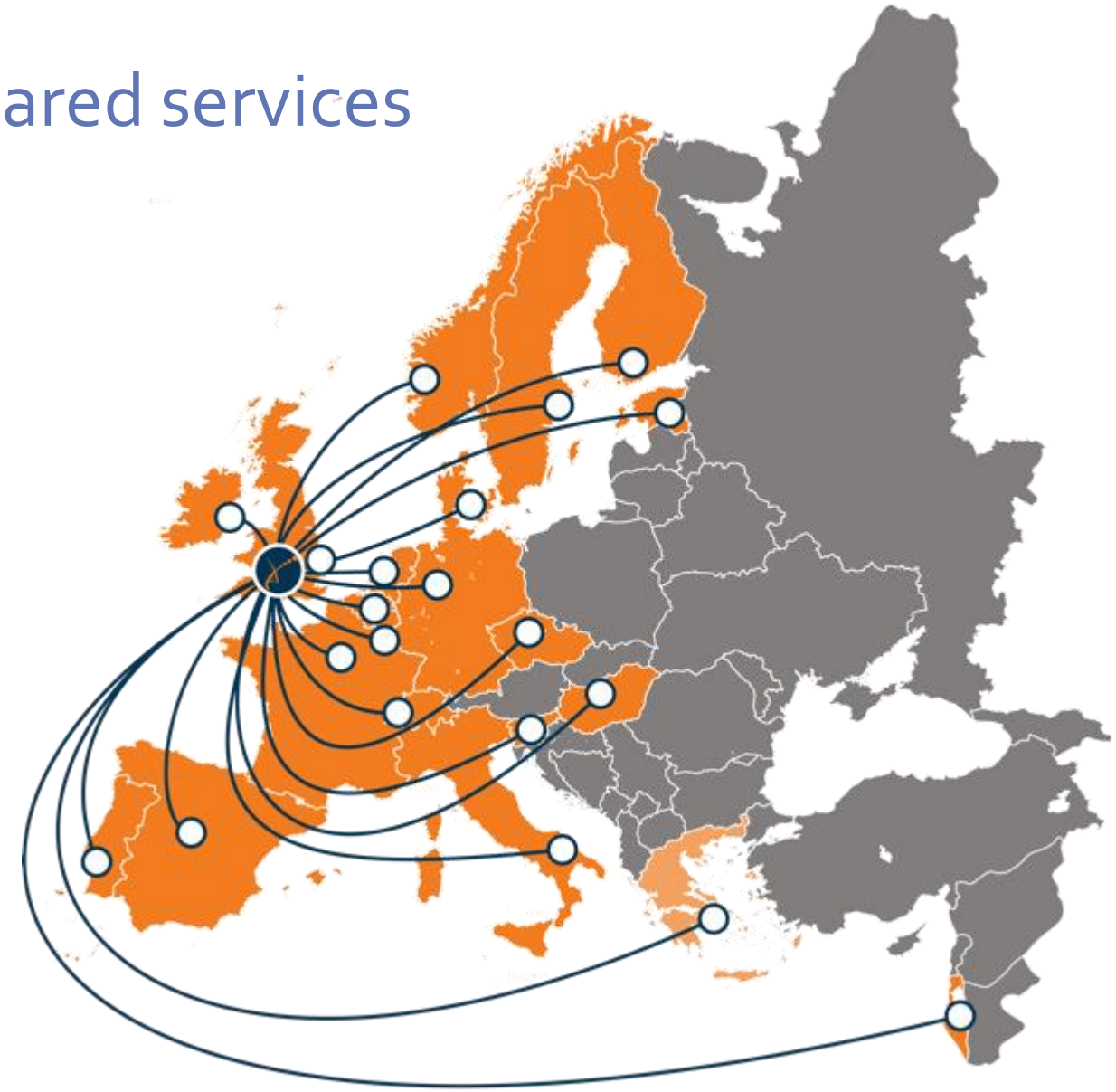
...delivered in partnership with research communities

A network of data Nodes

- ELIXIR Nodes are funded nationally
- ELIXIR Nodes build on national strengths and priorities
- ELIXIR Nodes provides a national framework for long-term resource management



Distributed infrastructure with shared services



ELIXIR's Platforms of shared services

- Accessed internationally
- Operated by Nodes
- Funded by national and international schemes
- Connected through ELIXIR's technical programme

ELIXIR

Technical Platforms that connect Compute, Data, Tools, Interoperability and Training services between Nodes

ELIXIR Communities that connect domain experts between Nodes

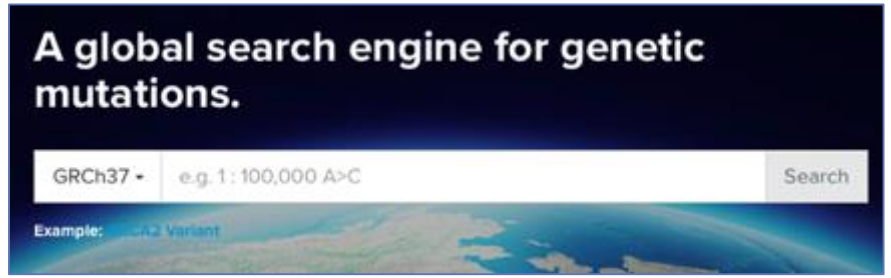
Plant sciences

- Mission: facilitate genotype-phenotype analyses for crop and tree species
- Actions:
  - Develop standards for representation of genotypic and phenotypic data
  - Make data discoverable and interoperable through common APIs
  - Annotate and submit key exemplar datasets to relevant public archives
  - Develop reusable modules for visualisations
  - Disseminate best practices and tools to national projects

Paul Kersey (EBI), Cyril Pommier (FR) & Celia Miguel (PT)

ELIXIR partners at leading plant research centres

Public data discovery web-service



Do you have information about the allele "C at position 32,936,732 on chromosome 13?"

Yes / No
(+optional metadata about the allele)

Do you have information about the allele "C at position 32,936,732 on chromosome 13?"

Beacon X: Yes
Beacon Y: No
Beacon Z: No
...

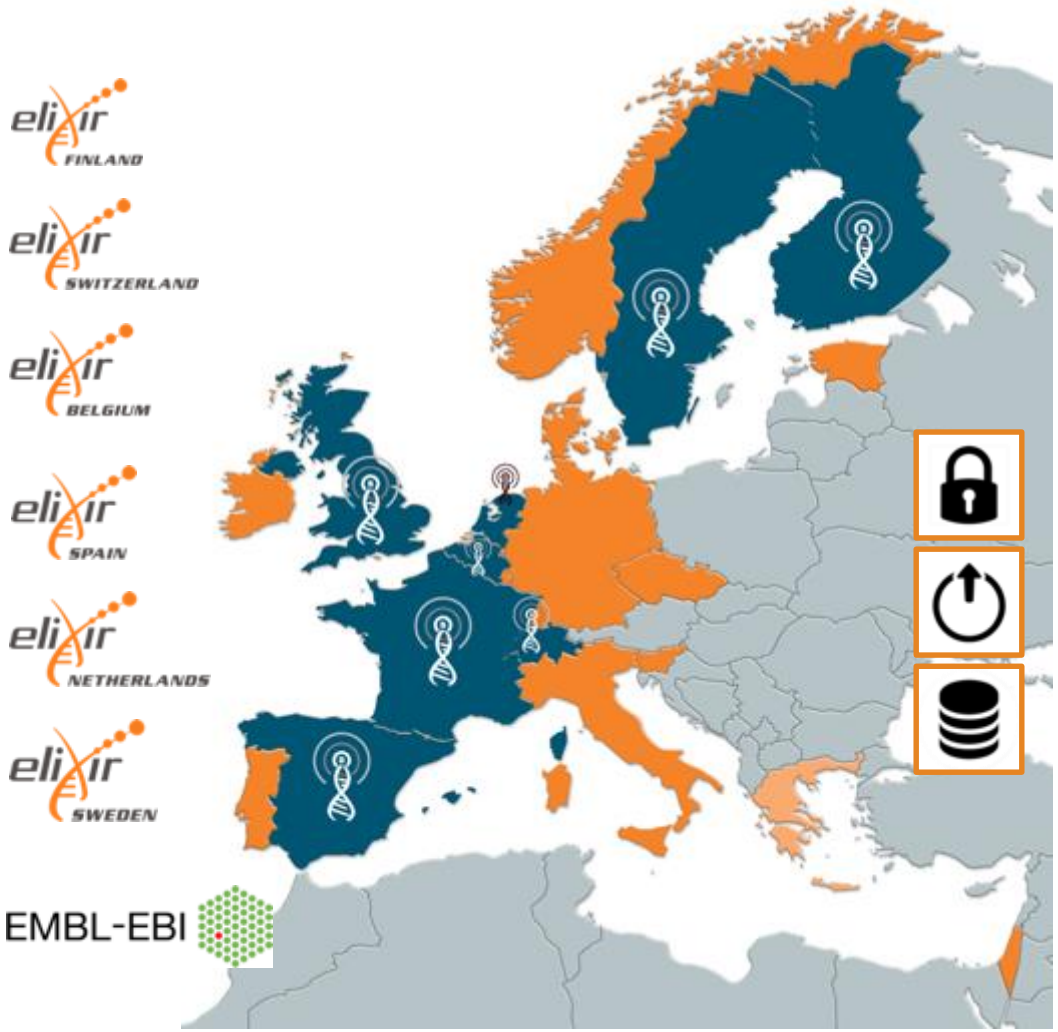
Beacon

beacon-project.io

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

elixir

ELIXIR and Beacons



- Funding to drive implementation of the Beacon technology within ELIXIR nodes
- ELIXIR Authentication and Authorization Infrastructure

Public	Accessible to internet users
Registered	Accessible to bona fide researchers
Controlled	Authorized – signed agreement needed

www.elixir-europe.org/beacons

EUROPEAN GENOME-PHENOME ARCHIVE

Beacon

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

elixir

Federation of human genome data: *localEGA*

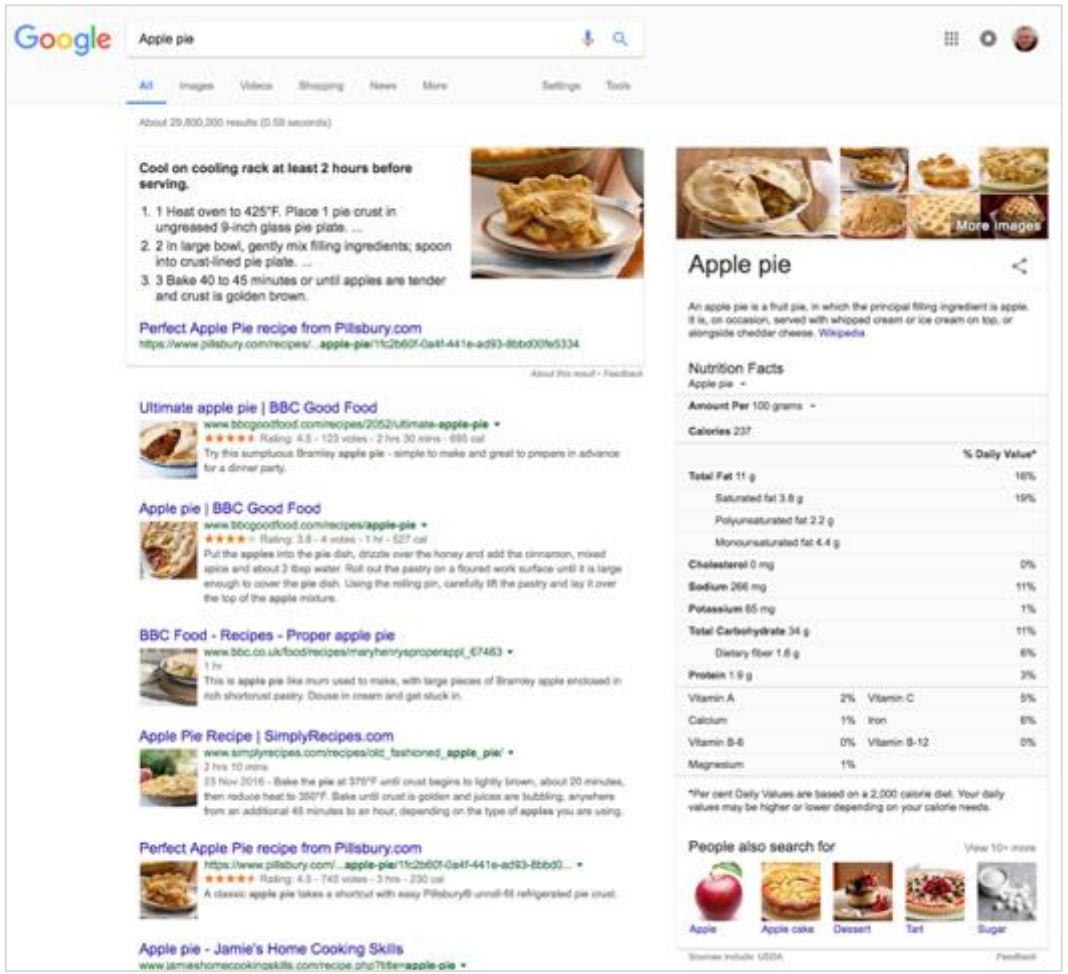
- Many national datasets from human research participants needs to be stored locally
- ELIXIR-EXCELERATE developing "localEGA" – shared metadata (FAIR) and local data store (secure)
- Linking local EGA to national clouds – and international access (ELIXIR-AAI)

How do you find a needle in a federated haystack?

Bioschemas

"schema.org markup for life sciences – minimum properties needed for finding data"

http://bioschemas.org



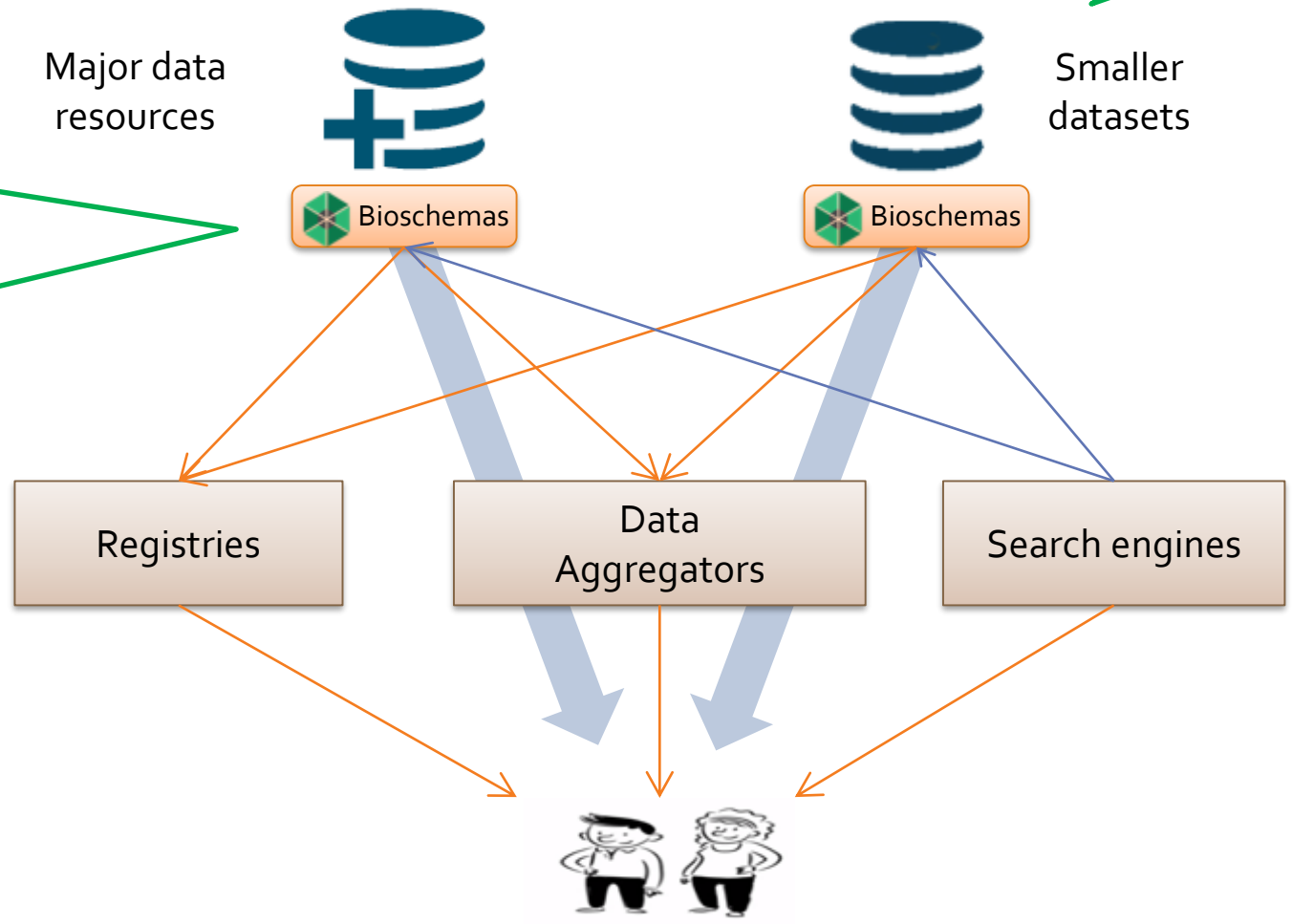
Bioschemas.org

*A community initiative built on top of Schemas.org to improve **Findability** and **Accessibility** in Life Sciences*

- Rapid markup
- Exposed to harvesting
- Find

- Standardised metadata
- Metadata publish and harvest without APIs or special feeds

- Feed bio registries and aggregators



Bioschemas adopted in EOSC

In November we presented early adopters…

now 16 live deployments!

- Generic standard for data discovery
- ”Research schemas” as Emerging federation architecture in EOSC (Part of EDMI)
- Created by the Bioschemas community, funded by the ELIXIR Budget

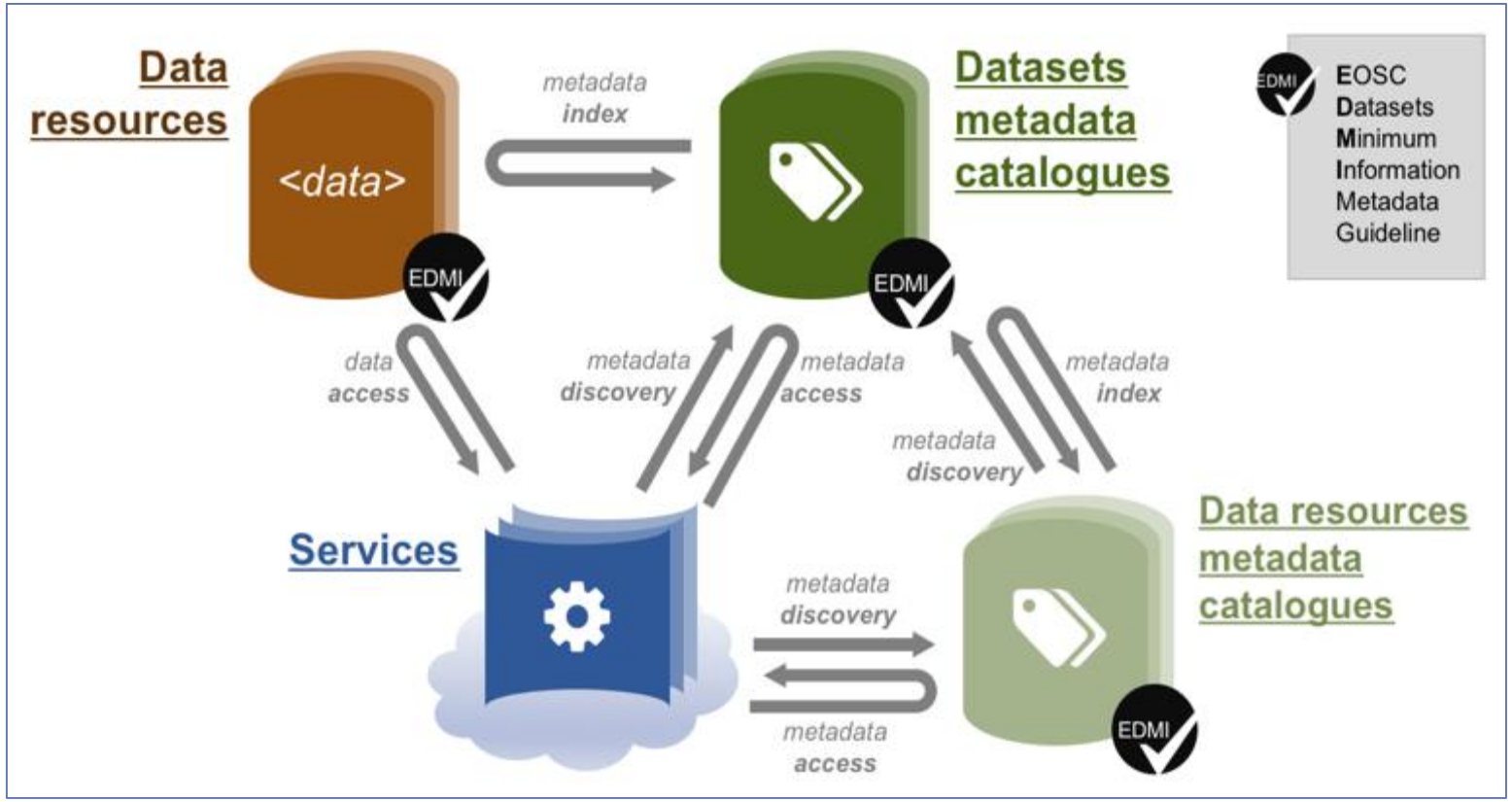
Services/sites implementing Bioschemas's markup

Name	Bioschema Profile	Profile Version	Structured Data Testing Tool
Identifiers	DataCatalog	0.1	visualise
Fairsharing	DataCatalog	0.1	visualise
Gigadb	DataCatalog	0.1	visualise
Human Protein Atlas	DataCatalog	0.1	visualise
EGA	DataCatalog	0.1	visualise
Isaexplorer	DataCatalog	0.1	visualise
IUPHAR/BPS	DataCatalog	0.1	visualise
MobiDB	DataCatalog	0.1	visualise
EGA Dataset	Dataset	0.1	visualise
MobiDB	Dataset	0.2	visualise
DataMed	Dataset	0.1	visualise
Biosamples	Sample	0.1	visualise
Pscan	Tool	0.1	visualise
PscanChIP	Tool	0.1	visualise
Cscan	Tool	0.1	visualise
BAR 3.0	Tool	0.1	visualise

http://bioschemas.org/liveDeploys/

EOSC pilot
The European Open Science Cloud for Research Pilot Project

European Open Science Cloud is build around domain registries



ELIXIR: Gateway for User access and mechanism for exposing life-science services (via *ELIXIR Registries)*

ELIXIR Authentication and Authorization Infrastructure AAI

Reliable electronic identification of users (ELIXIR ID) is needed to access the key services and capacities of ELIXIR.

- **You can link existing user accounts to create your ELIXIR ID today at www.elixir-europe.org** ELIXIR AAI allows Users to continue using their federated academic, corporate or social media identity by linking it to a personal ELIXIR ID.
- The ELIXIR service providers connected to ELIXIR AAI benefit from a centralised user identity and access management services.
- Protocols SAML2, OpenIDConnect.
- https://www.elixir-europe.org/services/compute/aai

- 359 Home Organisation IdPs enabled for login (via eduGAIN)
- 987 ELIXIR users
- 155 groups created in ELIXIR AAI
- 61 registered Resource Providers

ELIXIR Cloud & Compute

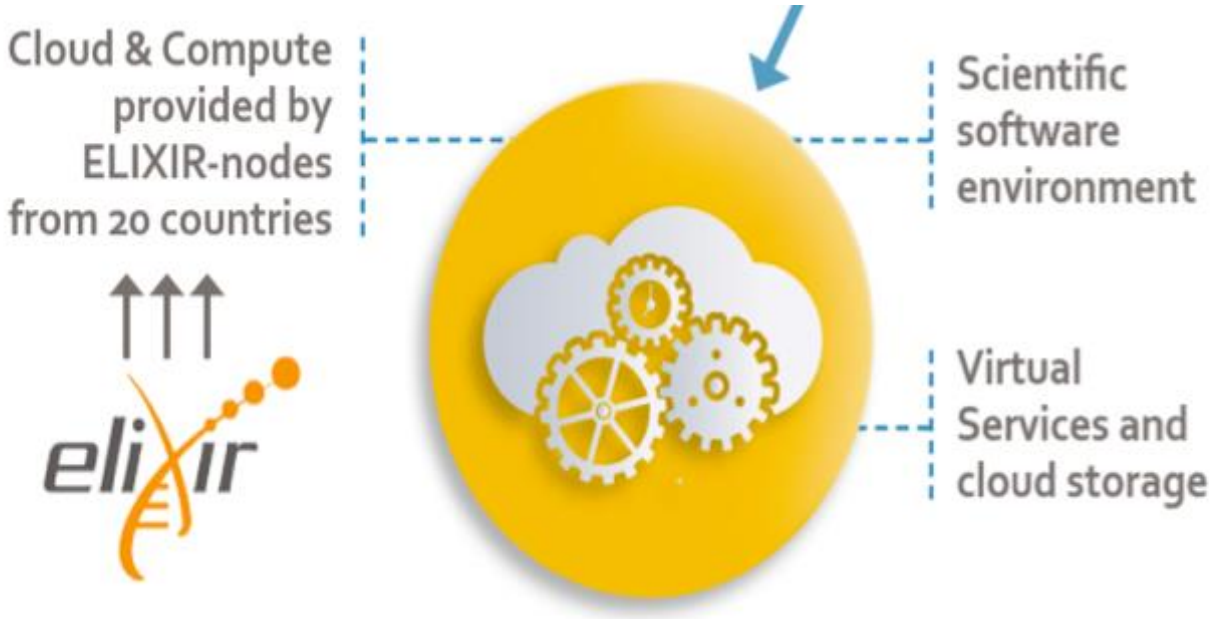
ELIXIR Cloud capacities surveyed here

DK, DE, EBI, FI, FR, CH confirmed capacity

> 60.000 compute cores

> 24.000 TB of storage

> 3.000 compute users



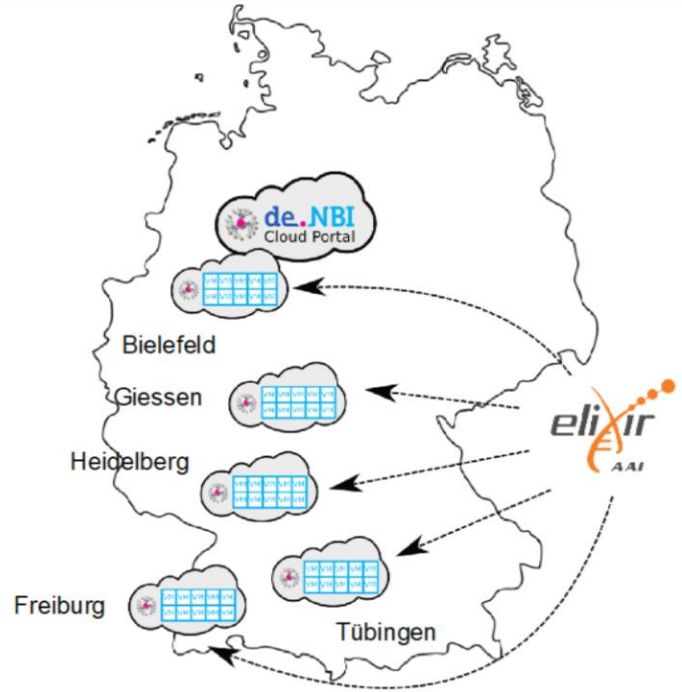
AAI as foundation for Transnational cloud access

de.NBI Cloud introduction

- Started 2016
- 5 compute centers
- >10000 Cores
- >5 PB Storage Capacity
- >100 TB RAM
- In collaboration with

Elixir

⇨ Single Sign-On



23.10.2017 7

Name	Description	ELIXIR Node
Data Archiving	ELIXIR Luxembourg provides integrated storage and archiving for curated Translational Medicine data. Data is stored on tranSMART servers or other suitable servers.	ELIXIR Luxembourg
AAI - REMS	CSC develops and hosts an open source tool REMS to assist a DAC to manage Data Access Applications and access rights to sensitive datasets. REMS is a key component in ELIXIR AAI.	ELIXIR Finland
Authentication and Authorization Infrastructure (AAI)	The ELIXIR AAI (Authentication and Authorisation Infrastructure) is the ELIXIR service portfolio for authenticating users and helping relying services to manage users' access rights in the services.	ELIXIR Czech Republic, ELIXIR Finland
Computerome	Computerome is the Danish National Supercomputer for Life Sciences. It serves all life science research groups within Denmark and is also open for international collaboration.	ELIXIR Denmark
Consulting/Expertise	Expertise on very diverse subjects of data processing, analysis and curation as well as connections to existing infrastructure RI's like ELIXIR. This expertise can be made available to research projects and organisations.	ELIXIR Netherlands
CSC Chipster	Chipster is a user-friendly analysis software for high-throughput data. It contains hundreds of analysis tools for next generation sequencing (NGS), microarray, proteomics and sequence data.	ELIXIR Finland
CSC Cloud	The CSC Cloud service is targeted for High Performance Computing (HPC), allowing customers to run virtual machines with exclusive access to up to 16 cores. This service includes cPouta and ePouta.	ELIXIR Finland
de.NBI cloud	The upcoming de.NBI cloud will provide an analytics infrastructure for bioinformatics. It will consist of computing power and storage capacity as well as flexible workflows and analysis tools.	ELIXIR Germany
Embassy Cloud	This is an OpenStack platform co-located with EMBL-EBI's services and data resources. Access to the Embassy Cloud is available for researchers outside EMBL if they have a collaboration with staff at EMBL.	EMBL-EBI
Help Desk	Under Development in collaboration with SURFsara and Lygature, based on prior experience deploying tools for end users / life scientists.	ELIXIR Netherlands
OpenConext	Open source software underlying the technology for SURF's single sign-on and facilitating national and international online collaboration in education and research.	ELIXIR Netherlands
sciCORE	sciCORE provides a high-performance computing infrastructure, large-scale storage resources, scientific software and databases, server infrastructures and user support. It also provides expertise to scientific research groups.	ELIXIR Switzerland
Vital-IT	Vital-IT supports and collaborates with life scientists in Switzerland and beyond. It provides expertise, training and maintains a high-performance computing (HPC) and storage infrastructure.	ELIXIR Switzerland

FAIR Data

- **F**indable
  (Citable)
- **A**ccessible
  (Trackable and countable)
- **I**nteroperable
  (Intelligible)
- **R**eusable
  (Reproducible)


SCIENTIFIC DATA

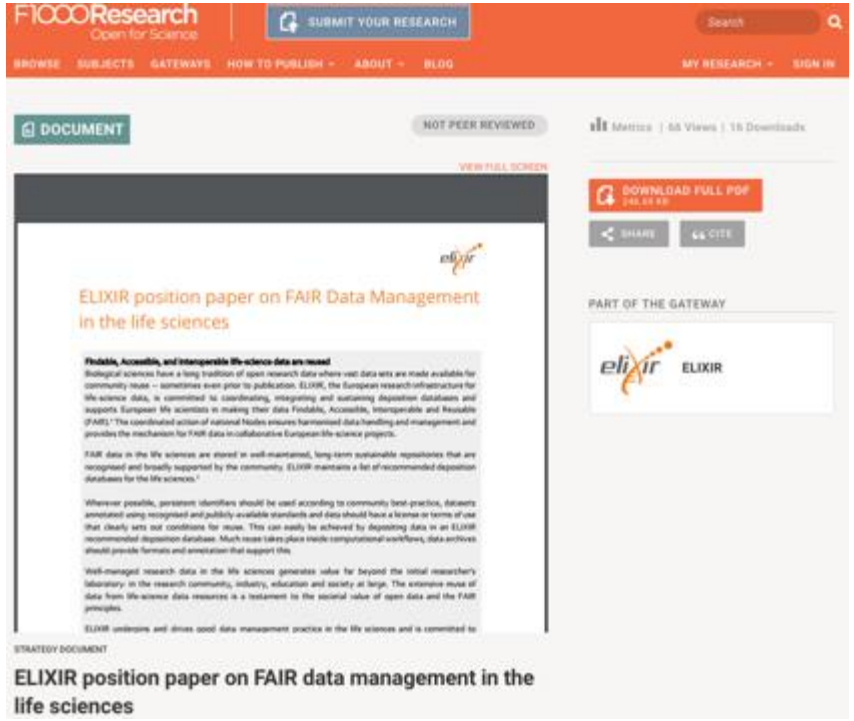
OPEN

Comment: The FAIR Guiding Principles for scientific data management and stewardship

picture credit: dtl.nl

ELIXIR Position Paper on FAIR data management in the life sciences

1. Open sharing of research data is a core principle
2. Data Management is crucial to science
3. Data should be submitted to deposition databases
4. All data submitted to Open Data archives should align with community-defined standards
5. ELIXIR Nodes implement FAIR for their respective nations
6. Professional skills, adequate resources and appropriate funding are needed for Data Management and infrastructure



Blomberg N and ELIXIR Consortium. ELIXIR position paper on FAIR data management in the life sciences. ***F1000Research*** **2017**, **6**(ELIXIR):1857 (document) (doi: 10.7490/f1000research.1114985.1)

"Whenever possible, biological research data should be submitted to the recommended community deposition databases"

ELIXIR Deposition Database list

Deposition Database	Data type	International collaboration framework [1]
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.	European Bioinformatics Institute and the Centre for Genomic Regulation
ENA	Nucleotide sequence information, covering raw sequencing data, contextual data, sequence assembly information and functional and taxonomic annotation.	International Nucleotide Sequence Database Collaboration
IntAct	IntAct provides a freely available, open source database system and analysis tools for molecular interaction data.	The International Molecular Exchange Consortium
MetaboLights	Metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.	
PDBe	Biological macromolecular structures.	
PRIDE	Mass spectrometry-based proteomics da… protein expression information (li… values) and the suppor…	

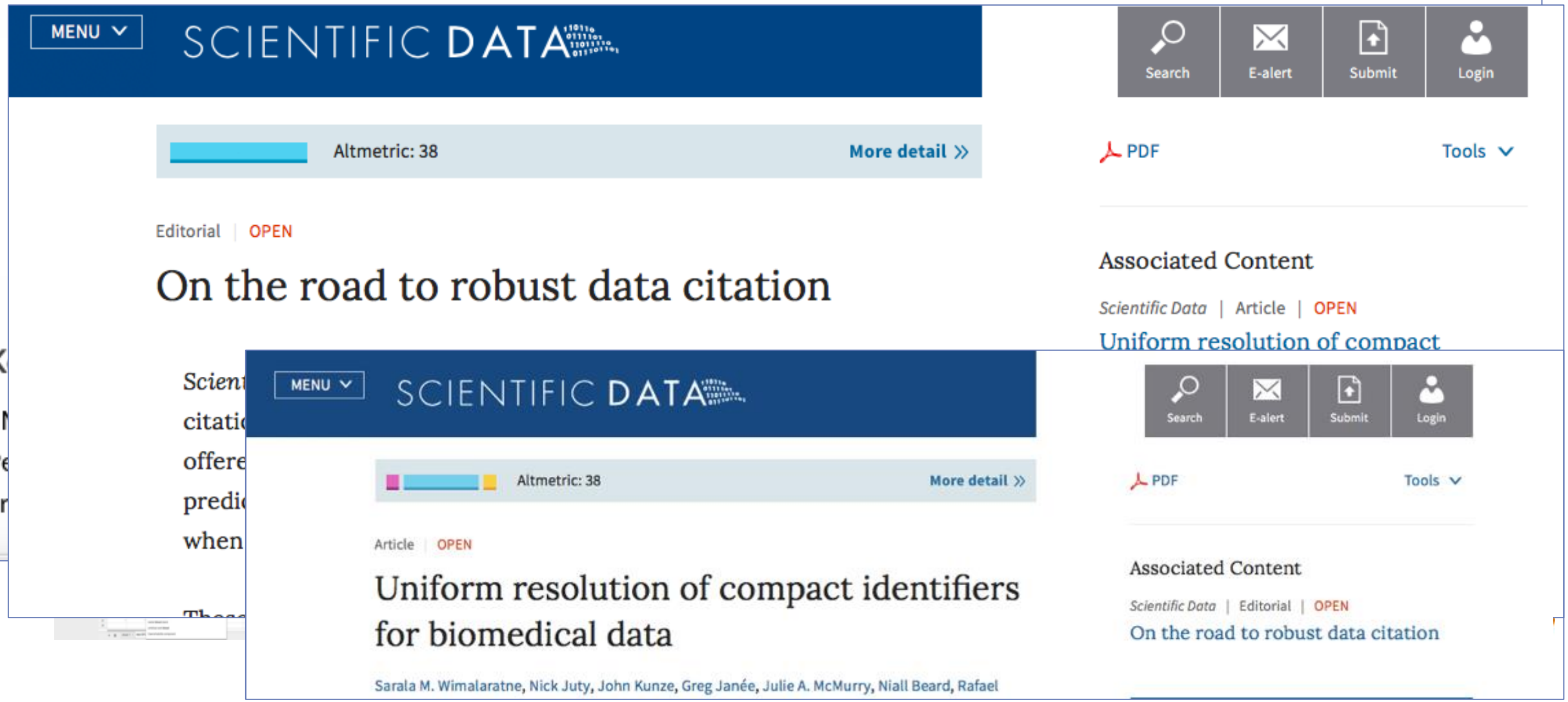
- The ELIXIR Deposition Databases meet the technical quality and governance criteria expected of ELIXIR Core Data Resources
- Agreed collectively by 21 Heads of Nodes
- International collaborative effort

https://elixir-europe.org/platforms/data/elixir-deposition-databases

"All data submitted to Open Data archives must be annotated in accordance with
community-defined standards"

https://elixir-europe.org/platforms/interoperability

Identifiers.org and N2T.net help establish global standards for citation of biomedical data



ELIXIR FAIR-CMM Model

Level	Process	Datasets and Linksets
1. Initial	Processes are disorganized. Success is dependent on specialised, heroic and one-off efforts, considered unrepeatable, because processes are not sufficiently defined and documented to be replicated.	Datasets are disorganised and may well be unstructured. No Linksets (i.e. explicitly published mappings between datasets).
2. Repeatable	Basic processes are established, created and maintained. Successes could be repeated, because the processes are defined, and documented.	Basic levels of FAIR are implemented by the dataset. Linksets are implied as the links are intermingled with the data. Descriptions of the links are not available.
3. Defined	An organization has its own process through greater attention to documentation, standardization, and integration.	Datasets have limited metadata and access capabilities. Linksets are identified, i.e. there are descriptions of the datasets that are linked to.
4. Managed	Organization monitors and controls its own processes through data collection and analysis	Datasets have further metadata and access capabilities. Linksets are explicitly managed (but can be idiosyncratic).
5. Optimizing	Processes are constantly being improved through monitoring feedback from current processes and introducing innovative processes to better serve the organization's particular needs	Datasets are fully annotated with metadata and access capabilities, i.e. they fully satisfy all the FAIR principles. Linksets are managed as first class objects, i.e. regarded as datasets in their own right, and accessible to mapping services.

"ELIXIR Nodes implement
a harmonised FAIR Data Management
programme for the life sciences"

Support via Implementation studies

- Genomics Data Management with EGA / TraIT (NL, EBI, ES)
- Data Management Implementation study

ELIXIR Core Data Resources – fundamentally important to life-science research

ELIXIR Core Data Resource list

Core Data Resource	Data type
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments.
CATH	A hierarchical domain classification of protein structures in the Protein Data Bank.
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds.
ChEMBL	Database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties and abstracted bioactivities.
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.
ENA	Nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
Ensembl	Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.
Ensembl Genomes	Comparative analysis, data mining and visualisation for the genomes of non-vertebrate species.
Europe PMC	Europe PMC is a repository, providing … books, patents and clinical …

https://www.elixir-europe.org/platforms/data/core-data-resources

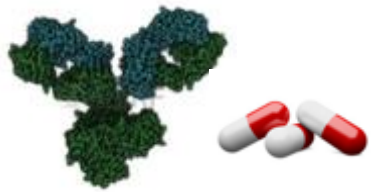
- 16 Core Data Resourced Nominated
- ELIXIR is committed to Open Access as a core principle for publicly funded research.
- Discussions on-going with Nodes, Resources and funders on high-quality, non-Open Access resources
- ELIXIR Core Data Resources should reflect this commitment and have terms of use or a license that enables the reuse and remixing of data.
- See "Identifying ELIXIR Core Data Resources"
- Agreed collectively by 21 Node directors

Open access life science data is part of the bioeconomy infrastructure

2010-2015:

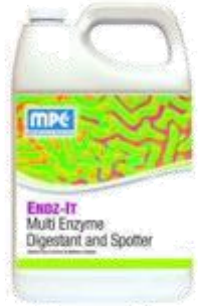
30 771 patents used bioinformatics repositories to identify genes, enzymes, SNPs, …

Vaccines



Pharmaceuticals

Beauty care



Industry enzymes …

Bousfield D, McEntyre J, et al. *Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources.*

Open Data as an Innovation Driver – ELIXIR SME Programme

- 4 Innovation and SME Forums
- 2 events in 2018 (Cambridge & Frankfurt)
- Growing ELIXIR industry intranet group (23 people) and mailing list (550)
- External events: BioVision, London Festival of Genomics, ENVRI Plus, IMI stakeholder event

ELIXIR Industry strategy

April 2016

Background

Industry usage of many key bioinformatics resources within Europe is high. Users of ELIXIR services range from large multinationals to micro-SMEs and cover fields including pharma, biotech, food and agriculture and blue biotech. These industries are major employers globally, generating wealth and supporting transformation to a knowledge-based economy.

The biomedical sector is worth €60 billion annually[1]. The pharmaceutical industry alone directly employs over 700,000 people in the EU, generating three to four times more downstream jobs and contributing to a trade surplus of €82 billion[2]. As a sector, the bioinformatics industry is forecast to be worth $13.3 billion by 2020[3]. Europe is a rich breeding ground for biotech start-ups, with 1,799 healthcare biotech companies operating in Europe in 2012. The healthcare sector as a whole accounts for 8% of the total European workforce and for 10% of the EU's GDP[4].

In the food sector, the EU was the top world exporter of food and drink products in 2012[5], exporting $98.7 billion (20.5% of world total) and importing $85.9 billion (18.1%). Fish production in aquaculture is increasing on average 8.8% per annum globally[6], driven by technological developments and better molecular understanding of species and parasites.

Industry itself sees added value in ELIXIR in terms of reduced costs and decreased duplication of effort, access to common data and interface standards and much better public-private data integration. Many of these requirements are addressed through the general implementation of ELIXIR and are relevant for all users in industry and academia.

Given the number of research-intensive companies in Europe, and the increasing reliance within industrial R&D upon computational methods, stimulating innovation and supporting industry is a key priority for ELIXIR. Indeed, ELIXIR's Scientific Programme for the period 2014–2018 includes a Strategic Objective to "Support Innovation in 'big-data biology'".

The range of industry sectors that ELIXIR has the potential to support is also broad - from publishers to SMEs to HPC and cloud providers. ELIXIR's Industry Strategy aims therefore to be comprehensive in its scope, yet focussed and tailored enough to address the needs of each industry sector.

Understanding Industry's needs

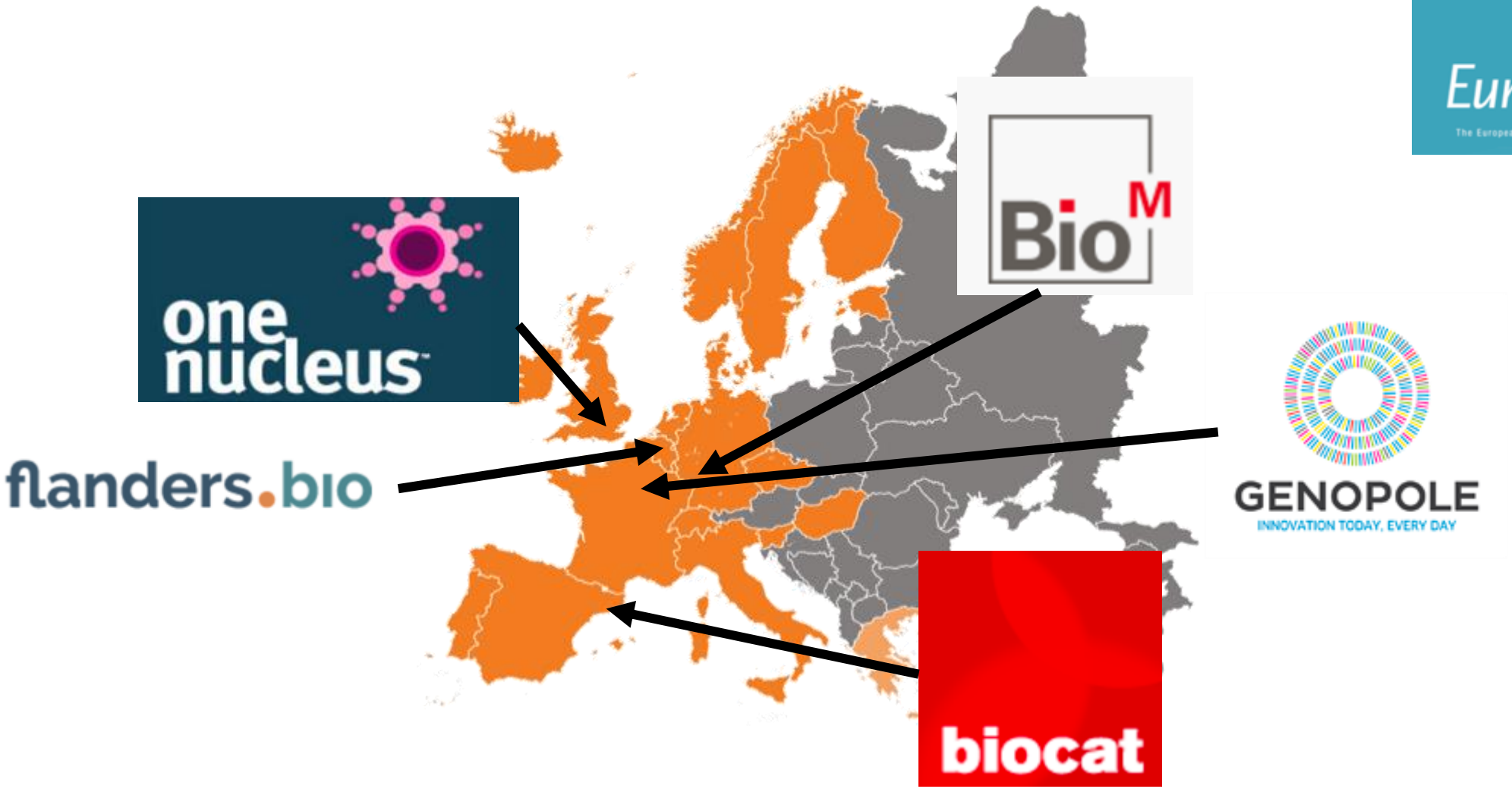
In order to understand the needs of industry and ensure that proposed activities are fit for purpose, ELIXIR's Industry Advisory Committee[8] provides high-level strategic advice on the activities included in this strategy. The ELIXIR Industry Advisory Committee's recommendation reports[9–11] map closely against the activities within this strategy.

ELIXIR's Industry Strategy responds to input from several other reports and analyses of industry needs. The ELIXIR Preparatory Phase industry report[12], and a report entitled "Developing ELIXIR interactions with industry", commissioned in 2013 and carried out by ConnectedDiscovery, clearly articulates a number of key value drivers for

1 Chakma et al. NEJM, 2014. doi:10.1056/NEJMp1402068
2 European Federation of Pharmaceutical Industries and Associations: http://www.efpia.eu/facts-figures
3 Bioinformatics Market by Sector (Molecular Medicine, Agriculture, Forensic, Animal, Research & Gene Therapy), Segment (Sequencing Platforms, Knowledge Management & Data Analysis) & Application (Genomics, Proteomics & Metabolomics) - Global Forecast to 2020": http://www.marketsandmarkets.com/Market-Reports/bioinformatics-ip.html
4 EC: Investing in health 2013: http://ec.europa.eu/health/strategy/docs/ev_investing_in_health.pdf
5 FoodDrinkEurope, 2014: http://www.fooddrinkeurope.eu/uploads/publications_documents/Presenting_an_EU_informed_policy_for_food_and_drink.pdf
6 Sustainable agriculture, forestry and fisheries in the bioeconomy: http://ec.europa.eu/research/bioeconomy/pdf/ki-na-13-005-en-c.pdf?view=fi%3Agenvote-none

7 ELIXIR Scientific Programme 2014–2018: https://www.elixir-europe.org/sites/default/files/documents/elixir_scientific_programme_final.pdf
8 ELIXIR IAC membership: https://www.elixir-europe.org/industry/understanding_industry_needs
9 ELIXIR IAC's Summary Report 2016: https://www.elixir-europe.org/system/files/ec_elixir_iac_public_summary_report_0.pdf
10 ELIXIR IAC's Summary Report 2015: https://www.elixir-europe.org/system/files/ELIXIR%20Industry%20Advisory%20Committee%20Report%202015.pdf
11 ELIXIR Preparatory Phase industry report: https://drive.google.com/drive/folders/0B4WQGu2wm1GSUFJaTjYxdkdzcUk

Extending the outreach via local Industry Clusters



Outcome from innovation events:

MediSapiens

Node - collaboration

CSC

Node - collaboration

BEEVA

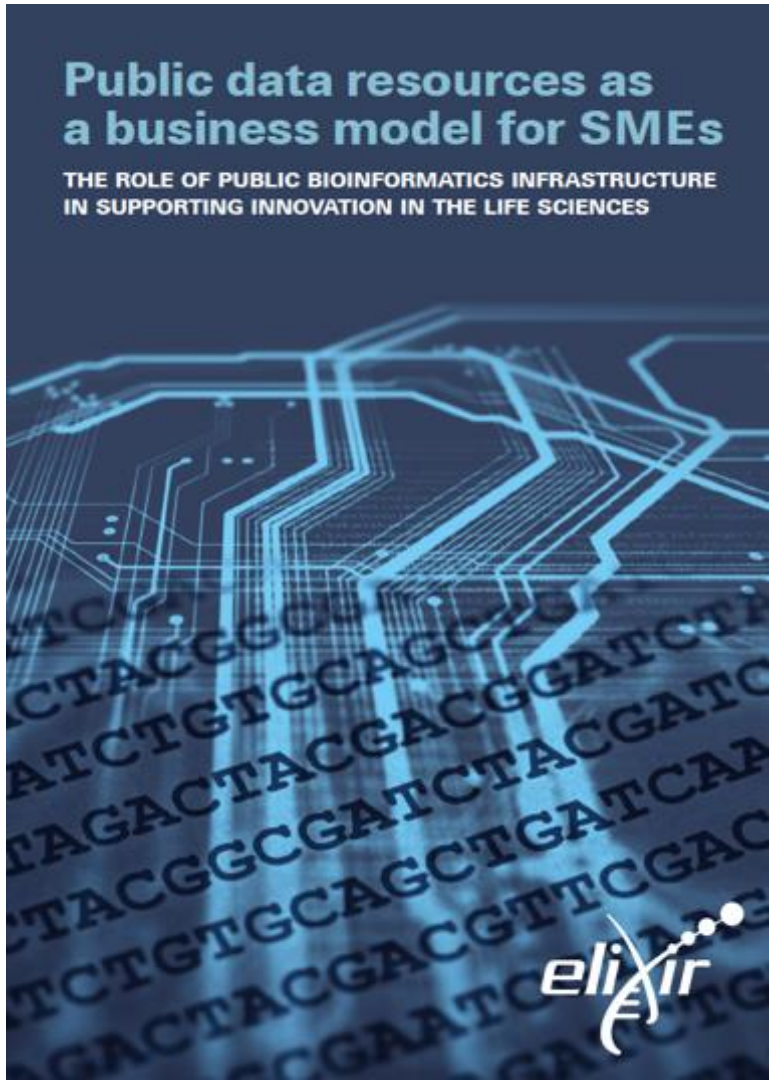
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Service - exchange

EUROPEAN GENOME-PHENOME ARCHIVE

elixir

Public data resources as a business model for SMEs



eagle genomics

Mind the Byte | SaaS for Drug Discovery

Ribo On

ONTOFORCE

REPOSITIVE

General Bioinformatics

thehyve

Customising

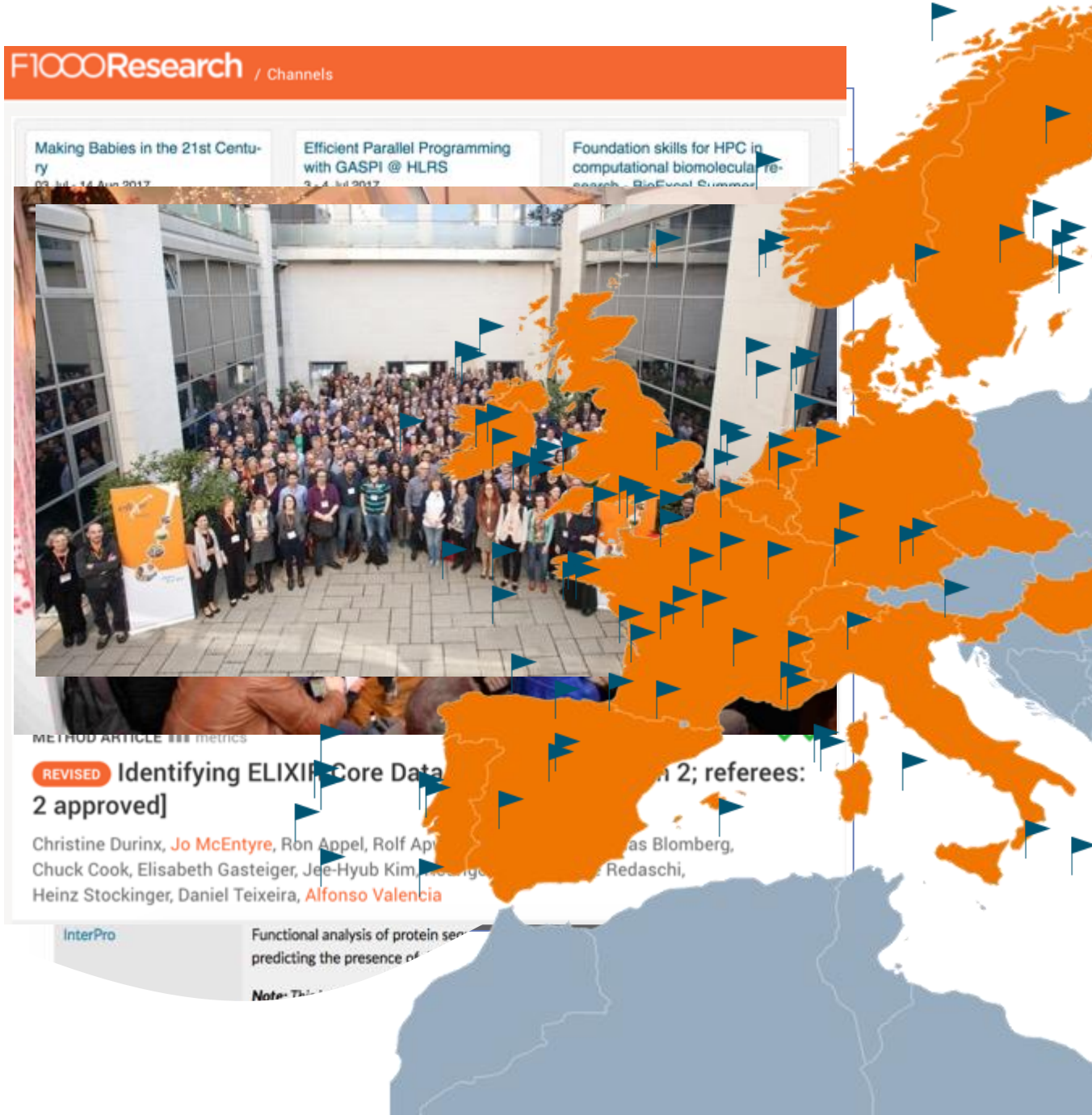
Aggregating

Enabling

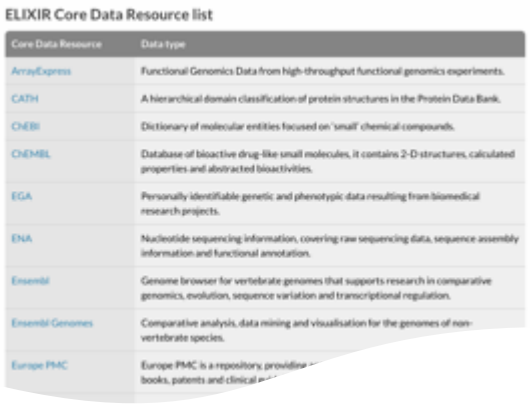
elixir

ELIXIR in numbers

- 21 Members and 1 Observer
- ~ 180 institutes involved
- 600+ staff
- 16 Core Data Resources
- 23 Implementation Studies ongoing or soon to start
- 20 papers in ELIXIR F1000 channel
- 264 live events in TeSS
- 350 companies attended Innovation and SME programme

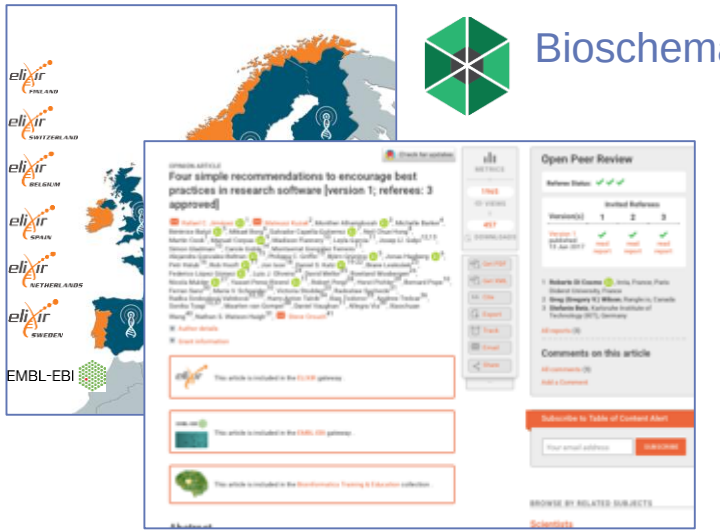


ELIXIR Outcomes


ELIXIR Core Data Resource list

Core Data Resource	Data type
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments.
CATH	A hierarchical domain classification of protein structures in the Protein Data Bank.
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds.
ChEMBL	Database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties and abstracted bioactivities.
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.
ENA	Nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
Ensembl	Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.
Ensembl Genomes	Comparative analysis, data mining and visualisation for the genomes of non-vertebrate species.
Europe PMC	Europe PMC is a repository, providing ... books, patents and clinical ...

Secure long-term haping data landscape and science policy

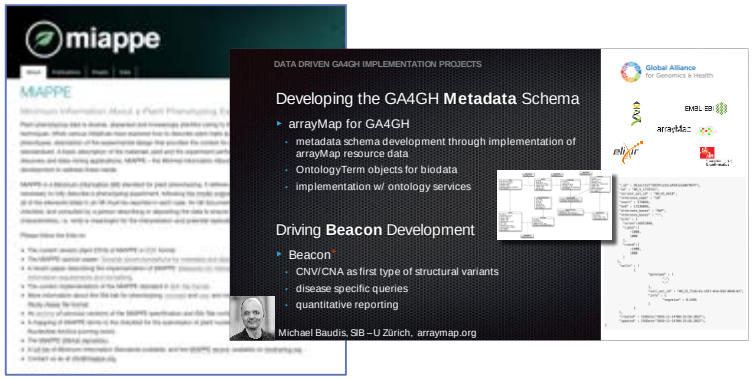


Bioschemas.org

Development of shared standards / conventions

Developed long-term, stable foundation services

Integration of node services



Community Standards

Tools


Søren Brunak


Alfonso Valencia

Data


Jo McEntyre


Christine Durinx

Compute


Tommi Nyrönen


Ludek Matyska


Steven Newhouse

Interoperability


Carole Goble


Chris Evelo


Helen Parkinson

Training


Patricia Palagi


Gabriella Rustici


Louisa Bellis


Celia van Gelder

Human data


Serena Scollen


Tom Keane


Jordi Rambla

Rare diseases


Ivo Gut


Marco Roos

Marine metagenomics


Nils Peder Willassen


Rob Finn

Plant sciences


Paul Kersey


Celia Miguel

Many thanks to Platform and Use Case leaders for the hard work in 2017!

ELIXIR's Coordination Groups

Training Coordinators Group

- Chair: Louisa Bellis

Technical Coordinators Group

- Chair: Rafael Jimenez

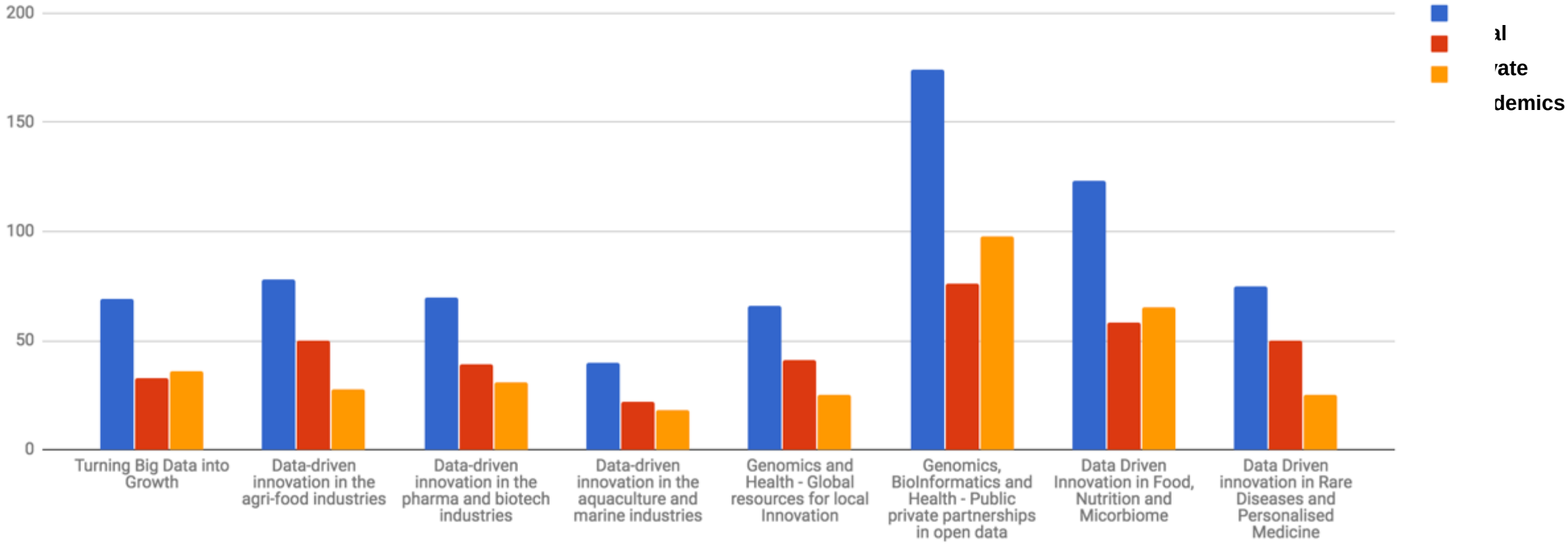
Thank you!

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe

ELIXIR Innovation and SME Forums connects Academia, Small Biotech and large companies



Copenhagen 2014	Wageningen 2015	Basel 2015	Oslo 2016	Helsinki 2017	Barcelona 2017	Brussels 2017	Paris 2017