ELIXIR: data for molecular biology and points of entry for marine scientists

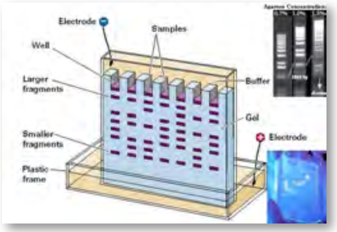
Guy Cochrane, EMBL-EBI

EuroMarine 2018 General Assembly meeting

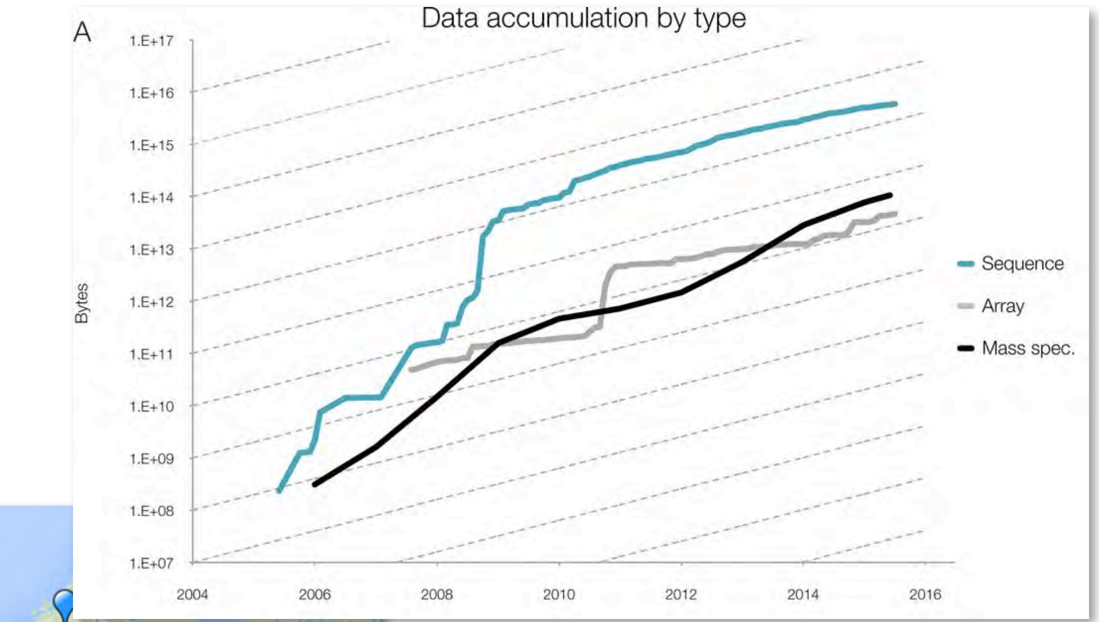
17-18 January 2018

www.elixir-europe.org

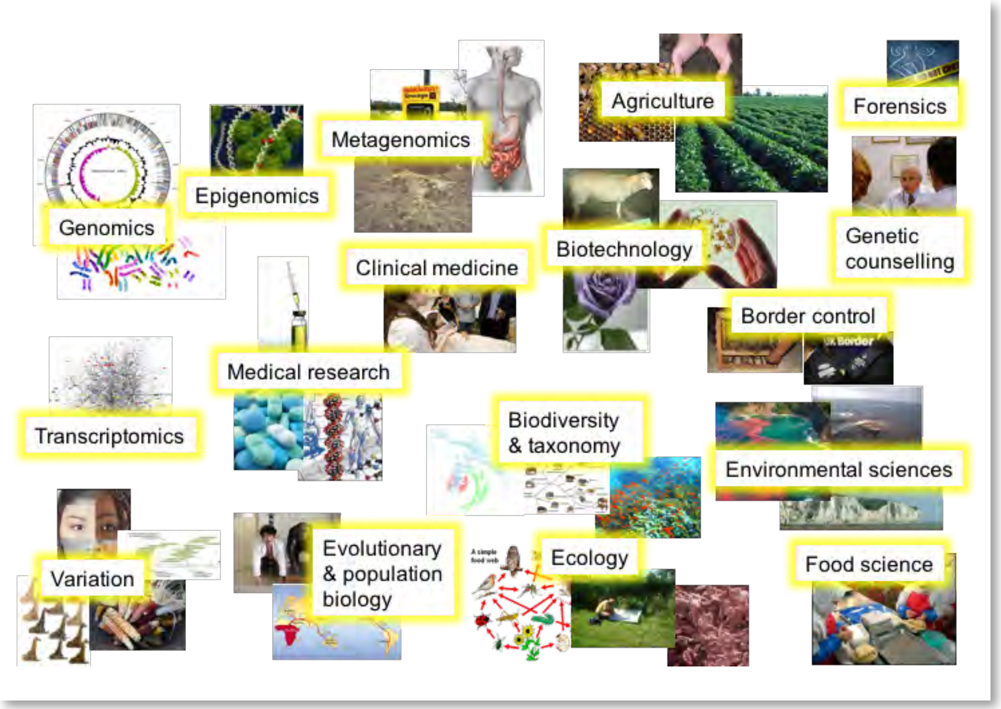
Scales of molecular biology data

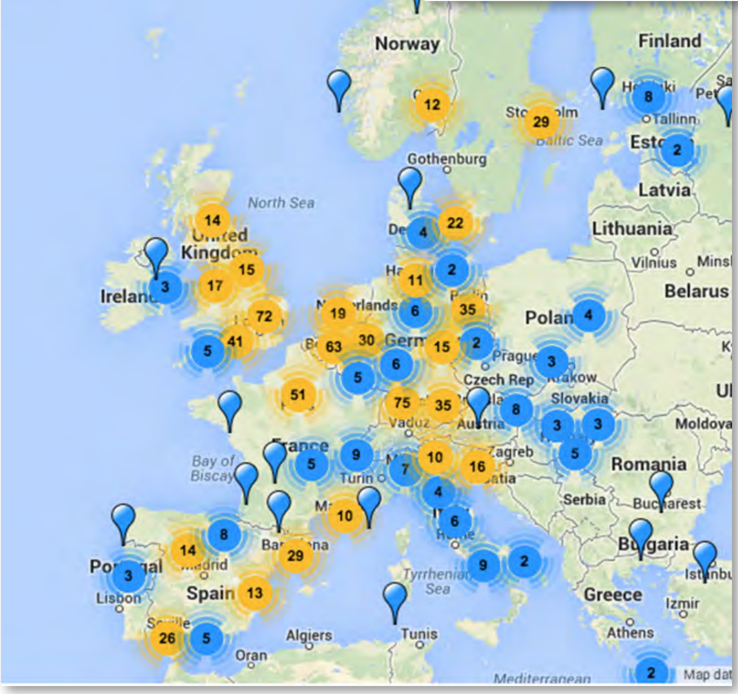


Credits:
Applied Biosystems
Oxford Nanopore Technologies
https://universe-review.ca/R11-16-DNAsequencing.htm



Source: Charles E. Cook et al. *Nucl. Acids Res.* 2016;44:D20-D26

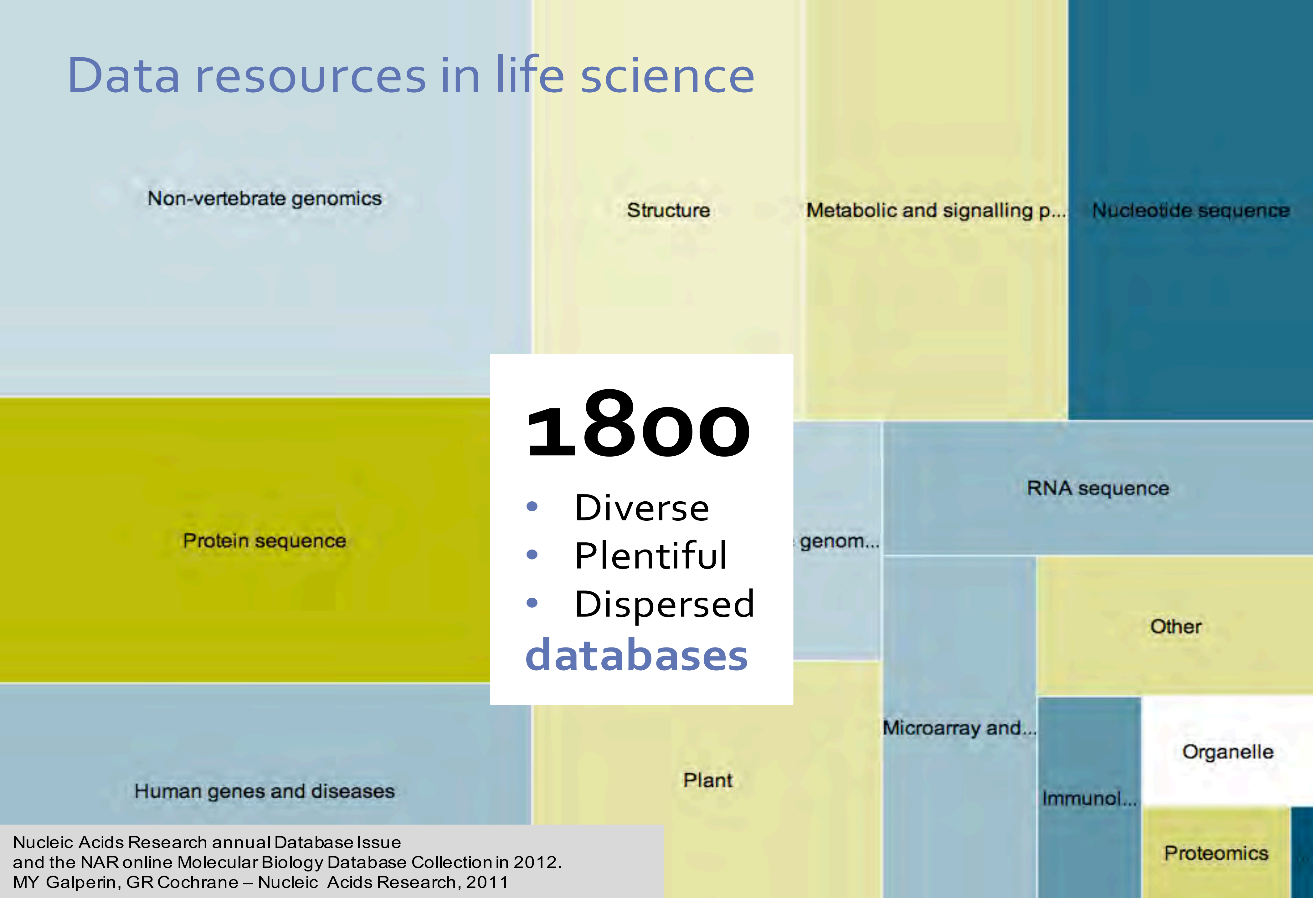




Source: http://omicsmaps.com

elixir

Data resources in life science



1800

- Diverse
- Plentiful
- Dispersed

databases

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

ELIXIR Members

- Belgium
- Czech Republic
- Denmark
- EMBL
- Estonia
- Finland
- France
- Germany
- Hungary
- Ireland
- Italy
- Israel
- Luxembourg
- Netherlands
- Norway
- Portugal
- Slovenia
- Spain
- Sweden
- Switzerland
- United Kingdom

ELIXIR Observers

- Greece

ELIXIR: European infrastructure for biological information

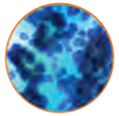
Data infrastructure for Europe's life-science research:

- *Data*
- *Interoperability*
- *Tools*
- *Compute*
- *Training*

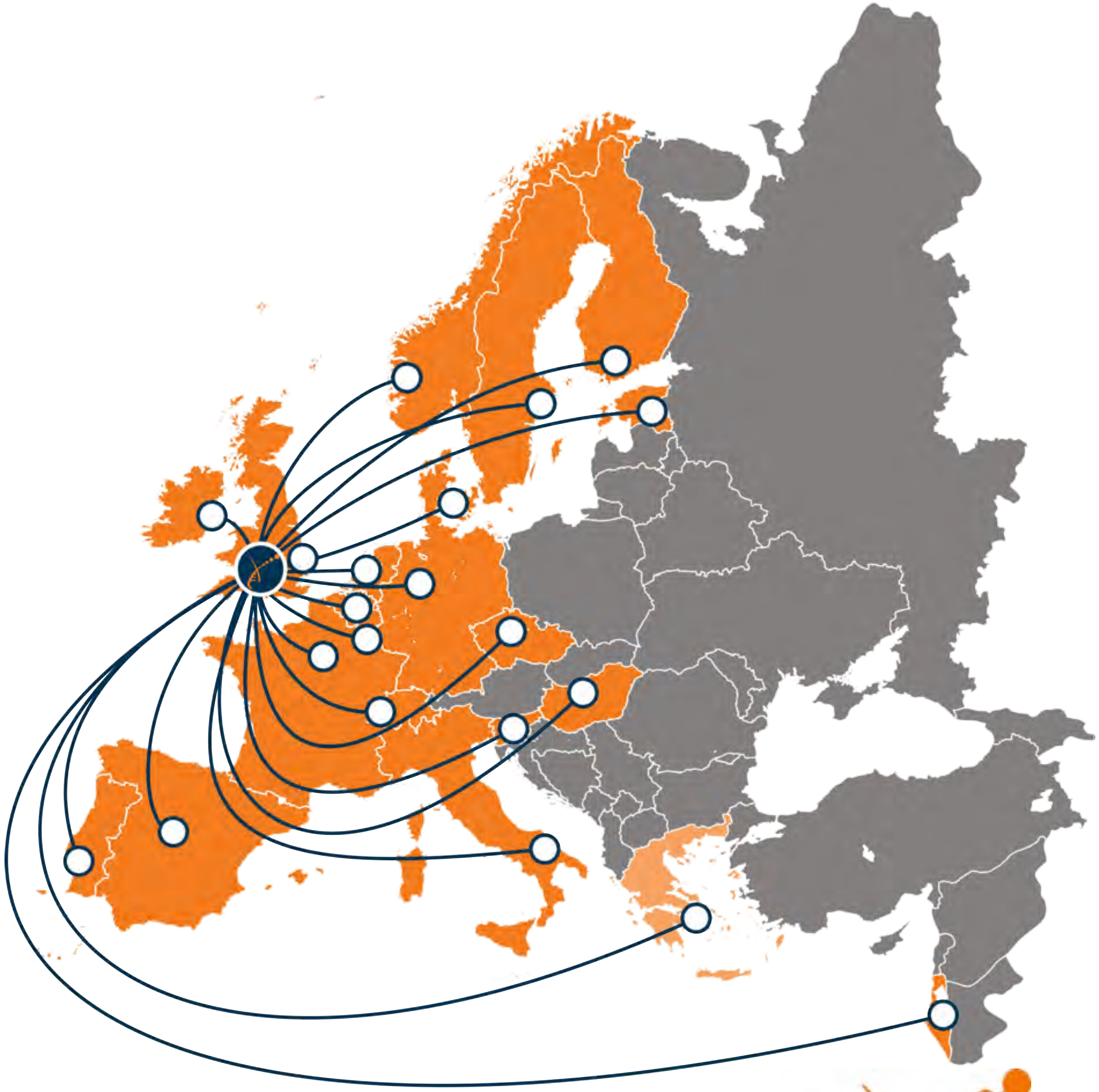
Marine metagenomics

Crop and forest plants

Human data



Rare diseases



@ELIXIREurope www.elixir-europe.org

ELIXIR Services

Data deposition:

ENA, EGA, PDBe, EuropePMC, …

Data management:

Genome annotation
Data management plans

Added value data:

UniProt, Ensembl, OrphaNet, …

Data Interoperability:

BioSharing, identifiers.org and OLS

Compute:

Secure data transfer, cloud computing, AAI

Bioinformatics tools:

Bio.tools

Industry:

Innovation and SME programme
Bespoke collaborations

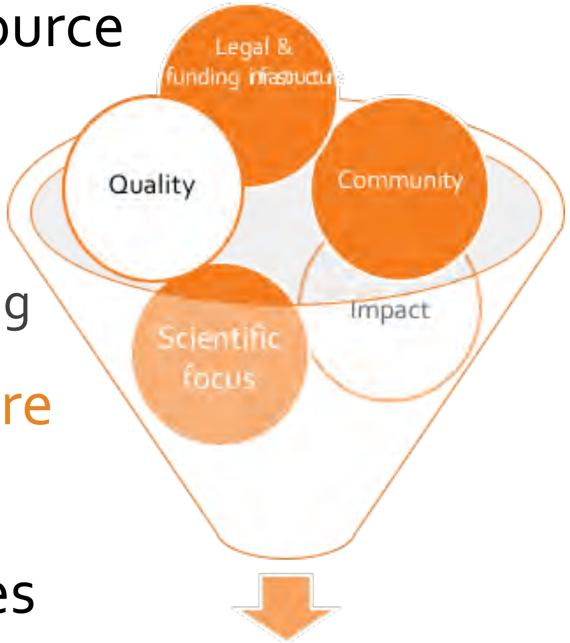
Training:

TeSS, Data Carpentry, eLearning

elixir

Core Data Resources – a data infrastructure

- Scientific focus and quality of science
  - Curation level, benchmarking
- Community served by the resource
  - Web statistics
- Quality of service
  - Uptime, user support and training
- Legal and funding infrastructure
  - Institutional support, use policy
- Impact and translational stories
  - Foundational role

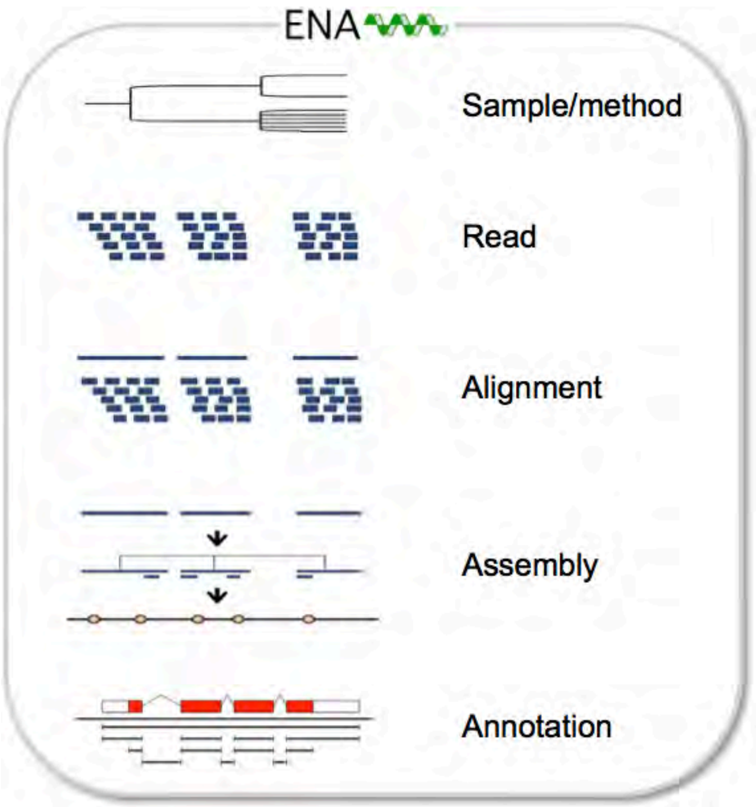


Initial set of Core Data Resources

- ArrayExpress
- ChEBI
- ChEMBL
- EGA
- ENA
- Ensembl
- Ensembl Genomes
- Europe PMC
- Human Protein Atlas
- The IMEx Consortium (IntAct and MINT)
- InterPro
- PDBe
- PRIDE
- STRING db
- UniProt

Durinx C, McEntyre J, Appel R *et al.* Identifying ELIXIR Core Data Resources *F1000Research* 2016, **5**(ELIXIR):2422

Core resource example: European Nucleotide Archive (ENA)



ENA

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

- **Globally comprehensive scientific record** and European node of INSDC
- A **broad platform** for the management, sharing, integration and dissemination of sequence data
- Established in the early 1980s, extended for **new technologies and applications**
- **Connectivity** with broader EMBL-EBI resources
- Sequence data **foundation**
- **Sustained** within EMBL-EBI under EMBL funding with additional support from EC, UK Research councils, Wellcome Trust, etc.
- **Substantial scale:** 1 submission every 6 minutes, 1.3 petabase pairs across 1.5 million taxa, 2,000-5,000 active data providers, global consumer userbase
- Rich submission, discovery and retrieval **software, tools and services**

EMBL-EBI elixir

Use Case: Marine metagenomics

- Mission: develop sustainable metagenomics infrastructure to enhance research and innovation
- Actions:
  - Develop standards for the marine domain
  - Develop databases specific to marine metagenomics
  - Implement tools and pipelines for metagenomics analysis
  - Develop search engine for interrogation of marine metagenomics datasets
  - Develop and deliver training for end users

https://www.elixir-europe.org/use-cases/marine-metagenomics

Nils P. Willassen , NO & **Rob Finn**, EBI

Tara Oceans expeditions: facts and figures

35,000 samples

Scientists on the expedition helped collect, pack and ship around 35,000 samples of plankton and water.

7012 datasets

One of the richest molecular sample collections in the public domain - freely available to everyone.

11,535 gigabytes

Size of the *Tara* datasets in the European Nucleotide Archive as of May 2015. This represents 12,581 gigabases - roughly equivalent to 135 fully sequenced human genomes.

Unlimited

Potential to discover new things about life in the world’s oceans.

40 million genes

Largest-ever DNA sequencing effort for ocean science.

Genetic sequences collected could represent tens of thousands of new species and ecosystem interactions.

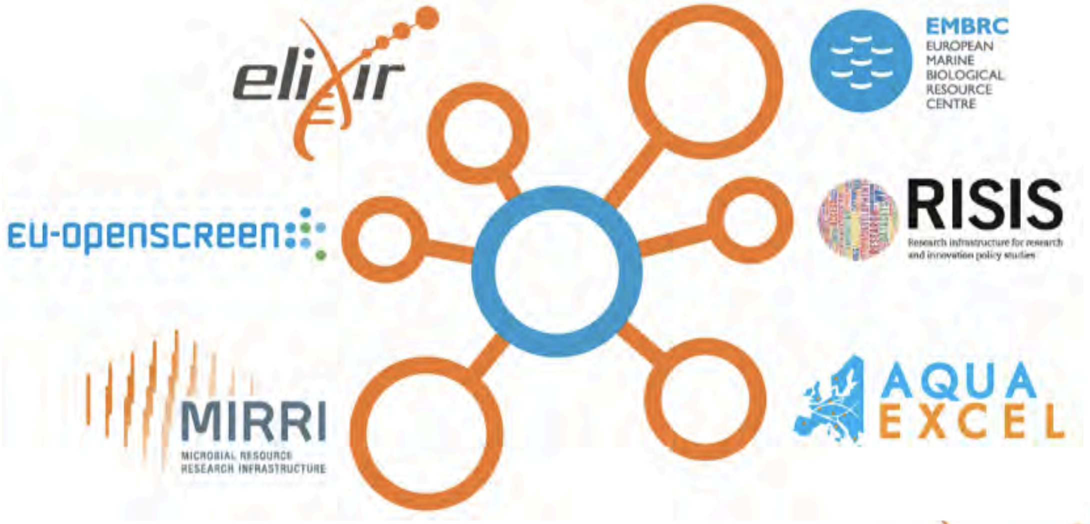
Considering the size of the world’s oceans, there is much, much more to discover.

Tara data: www.ebi.ac.uk/services/tara-oceans-data

EMBL-EBI

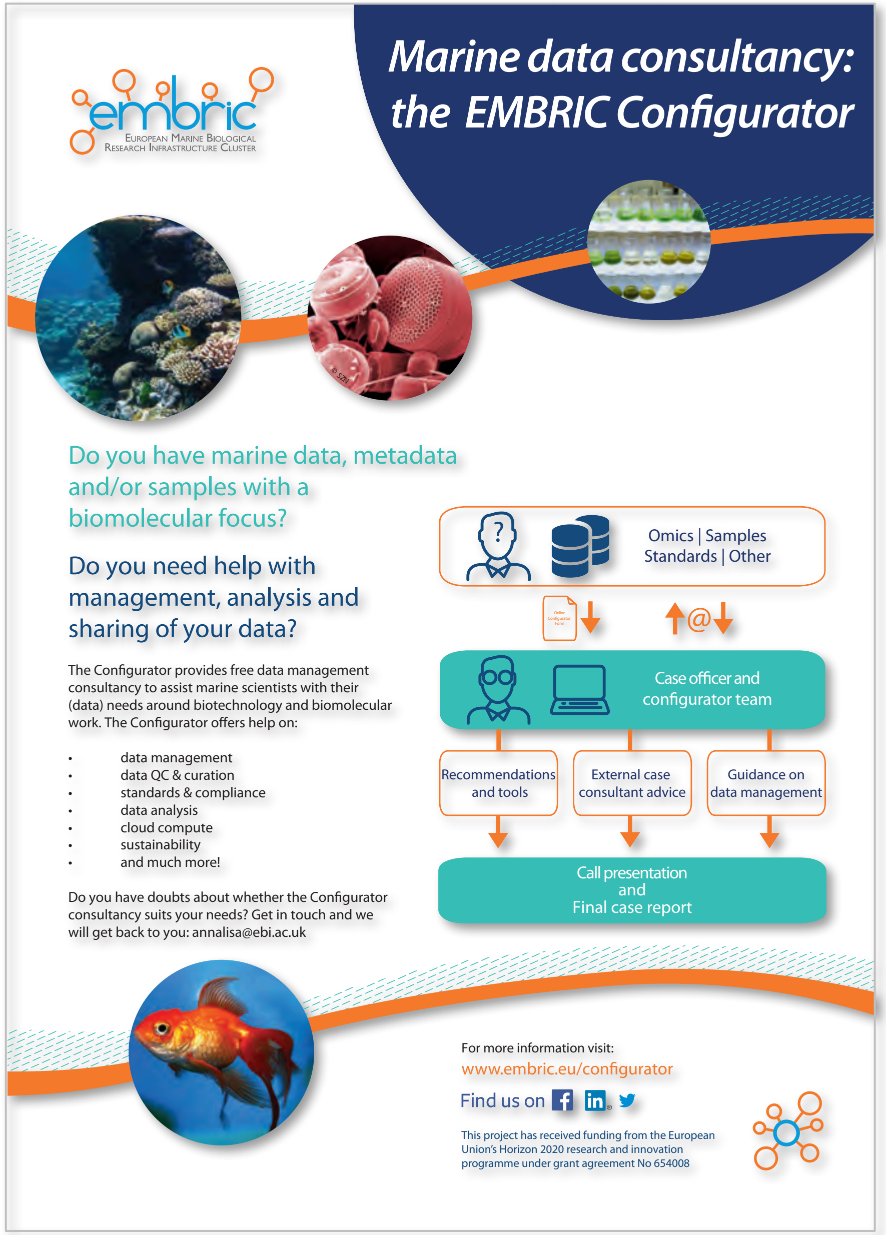
EMBRIC

- **European Marine Biological Research Infrastructure Cluster to promote the Blue Bioeconomy**
  - Objectives
    - Coherent chains of high quality services across RIs
    - Strengthen connection between science and industry
    - Defragment communities
    - TNA - 15 February 2018 opening of the 2nd call (http://www.embric.eu/access/TA)
  - Consortium
    - Led by UPMC – Bernard Kloareg
    - 27 partners
    - INFRADEV-4 'RI cluster' call


elixir
EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE
eu-openscreen
RISIS
Research infrastructure for research and innovation policy studies
MIRRI
MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE
AQUA EXCEL

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°654008.

EMBRIC Configurator



embric
European Marine Biological Research Infrastructure Cluster

Marine data consultancy: the EMBRIC Configurator

Do you have marine data, metadata and/or samples with a biomolecular focus?

Do you need help with management, analysis and sharing of your data?

The Configurator provides free data management consultancy to assist marine scientists with their (data) needs around biotechnology and biomolecular work. The Configurator offers help on:

- data management
- data QC & curation
- standards & compliance
- data analysis
- cloud compute
- sustainability
- and much more!

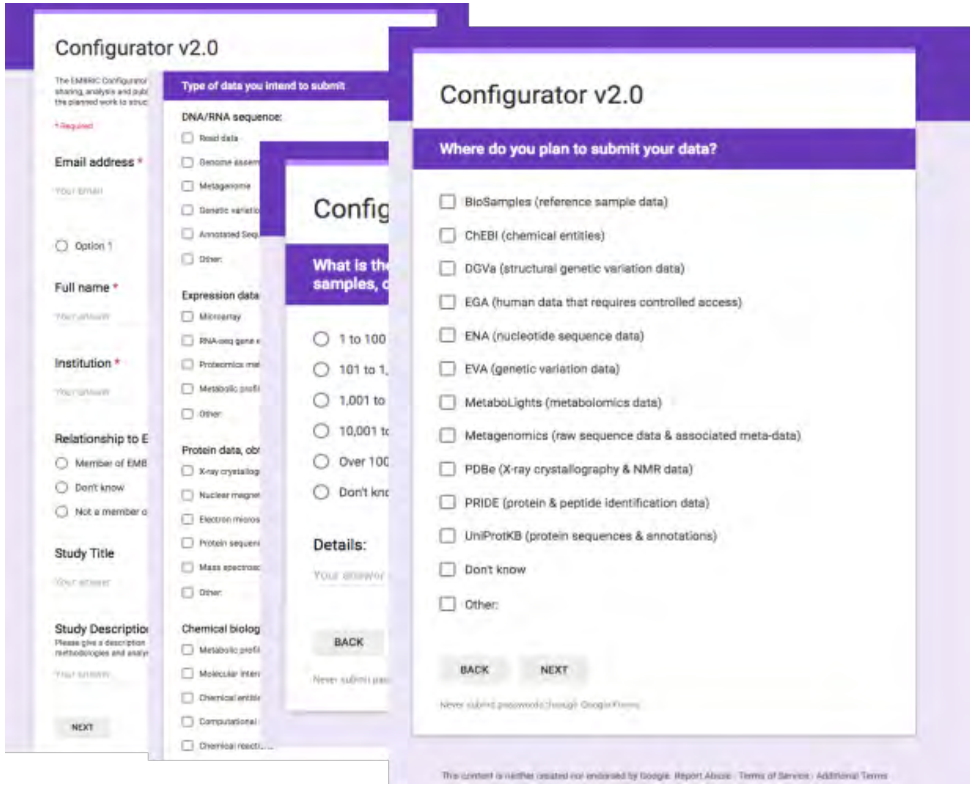
Do you have doubts about whether the Configurator consultancy suits your needs? Get in touch and we will get back to you: annalisa@ebi.ac.uk

Omics | Samples
Standards | Other
@
Case officer and configurator team
Recommendations and tools
External case consultant advice
Guidance on data management
Call presentation and Final case report

For more information visit:
www.embric.eu/configurator

Find us on

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654008



Configurator v2.0

* Required

Email address *

Your email

○ Option 1

Full name *

Your answer

Institution *

Your answer

Relationship to E

○ Member of EMB
○ Don't know
○ Not a member o

Study Title

Your answer

Study Descriptio

NEXT

Type of data you intend to submit

DNA/RNA sequence:

□ Read data
□ Genome assem
□ Metagenome
□ Genetic variatio
□ Annotated Seq
□ Other:

Expression data

□ Microarray
□ RNA-seq gene
□ Proteomics me
□ Metabolic profi
□ Other:

Protein data, obt

□ X-ray crystallog
□ Nuclear magne
□ Electron micros
□ Protein sequen
□ Mass spectrosc
□ Other:

Chemical biolog

□ Metabolic profi
□ Molecular inter
□ Chemical entiti
□ Computational
□ Chemical reacti...

Config

What is the samples, c

○ 1 to 100
○ 101 to 1,
○ 1,001 to
○ 10,001 to
○ Over 100
○ Don't kno

Details:

Your answer

BACK

Configurator v2.0

Where do you plan to submit your data?

□ BioSamples (reference sample data)
□ ChEBI (chemical entities)
□ DGVa (structural genetic variation data)
□ EGA (human data that requires controlled access)
□ ENA (nucleotide sequence data)
□ EVA (genetic variation data)
□ MetaboLights (metabolomics data)
□ Metagenomics (raw sequence data & associated meta-data)
□ PDBe (X-ray crystallography & NMR data)
□ PRIDE (protein & peptide identification data)
□ UniProtKB (protein sequences & annotations)
□ Don't know
□ Other:

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Additional Terms

- 22 registered consultants
- 2 ongoing cases
- 4 complete cases
- Call for:
  - More consultants
  - More cases

embric
European Marine Biological Research Infrastructure Cluster

elixir

EMBRIC Configurator examples

- **Sustainability**: Long-term operation of domain-specific data resource on fish species
- **Data dissemination**: *Pseudo-nitzschia multistriata* genomics, transcriptomics and gene expression
- **Data access**: Small molecule activities with respect to strains and genes

Acknowledgements

elixir

www.elixir-europe.org

ELIXIR-EXCELERATE is funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559.

embric
EUROPEAN MARINE BIOLOGICAL
RESEARCH INFRASTRUCTURE CLUSTER

www.embric.eu/
www.embric.eu/configurator

EMBRIC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654008

ENA

www.ebi.ac.uk/ena/

TARA OCEANS

www.ebi.ac.uk/services/tara-oceans-data

BBSRC
bioscience for the future

EMBL

GORDON AND BETTY
MOORE
FOUNDATION

HORIZON 2020
THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION

wellcometrust

Oxford
NANOPORE
Technologies

elixir

Talk to us

Annalisa Milano
Data Coordination
Biocurator

Jeena Rajan
Sequence Data
Biocurator

Guy Cochrane
Team Leader