EOSC Research and Data Interoperability (WP6 – Task2)

Niklas Blomberg, ELIXIR
Niklas.Blomberg@ELIXIR-Europe.org
WP6 Overview

Partners involved:
- STFC, ELIXIR, CNRS, KIT, ARC, JISC, CNR, INFN, DESY, ICOS, GEANT, INAF, BGS, UMAN, PIN, CEA, CINECA

Tasks:
- **T6.1**: gap analysis & interoperability architecture [CNRS lead]
- **T6.2**: EOSC Research and Data interoperability [ELIXIR lead]
- **T6.3**: Interoperability pilots [INFN lead]

Effort: 176 PMs
Deliverables: e-Infra Gap Analysis, June 2017
Milestones: 6.1, June 2017
Scientific data require curation and maintenance

- Every 6 minutes a sequence is deposited in ENA
- Core biomolecular archives contain >10 PB of data
- Uniprot connects over 120 biomolecular resources

**Source:** Charles E. Cook et al. *Nucl. Acids Res.* 2016;44:D30-D36

**Nature news, 19 December 2013**
Domain repositories are curated & used

European Nucleotide Archive

DNA Variations (SNPs)

Inherited disease (OMIM)

Gene Expression Studies

DOIs (‘long tail’)

Protein Data Bank

Kafkas S, Kim JH, and McEntyre JR Database Citation in Full Text Articles (May 2013) PLoS One 10.1371/journal.pone.0063184
Framework to Identify and sustain our core infrastructure of data resource

1) **Scientific focus** and quality of science
   e.g. curational effort, benchmarking

2) **Community** served by the resource
   e.g. web statistics

3) **Quality of service**
   e.g. uptime, user support and training

4) **Legal and funding infrastructure**
   e.g. institutional support, use policy

5) **Impact** and translational stories
   e.g. foundational role

---

Durinx C, McEntyre J, Appel R et al. Identifying ELIXIR Core Data Resources
(doi: [10.12688/f1000research.9656.1](https://doi.org/10.12688/f1000research.9656.1))
Open access science data is intensively reused

Biosimulation market worth $1bn/yr (2015)

http://www.marketsandmarkets.com/Market-Reports/biosimulation-market-838.html
Open access 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 access science data is part of the bioeconomy infrastructure.

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.
# FAIR data in EOSC(pilot)

## Findable
- simple, lightweight schemas for data discovery and data indexing
- No single entity owns the registry - open, accessible and extensible by communities

## Accessible
- Community and generic repositories – models to support existing repo’s
- Engage publishers!

## Interoperable
- Community knowledge hubs – overview of the landscape
- Domain infrastructure

## Reusable
- Citation practices, literature links, ...
- Describe societal value from data reuse
Findability of datasets and samples in the EOSC based on schema.org
Align with global standards - Findability of datasets via standard API/web services

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
…
EPOS vision for data interoperability

National RIs & International organizations generate data & services, ensuring quality control, standardization.
Helping users make the right decision

My funder’s data policy recommends the use of established standards, but which ones are widely endorsed and applicable to my toxicological and clinical data?

What are the mature standards and standards-compliant databases we should recommend to our authors?

Am I using the most up-to-date version of this terminology to annotate cell-based assays?

I understand this format has been deprecated; what has it been replaced by and is it appropriate for this data?
In Summary:

- RI have significant operational investments in data resource, interoperability and compute access
  - EOSC must facilitate data and service integration at the RI level

- EOSC should provide access to computational resources for the elaboration on new data products
  - Standards, interfaces, market for storage & compute

- EOSC needs to provide a framework (resources, governance) for metadata preservation, integration,…

End-users need to access their own RI capabilities and seamlessly, elastically, find themselves using EOSC when their own RI runs out of capability.
Which naturally links to RDA...

- Forum for broad community input
- Neutral ground for discussions w vendors & industry stakeholders (e.g. publishers)
- Broad consultation with Funders & Policy makers on value
- Beyond Europe
Thank you

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe
Cost & Value of research data infrastructure little known

Box B: Budget allocation to RDM and RDI unknown to many

Survey question to Research Infrastructure Providers, Libraries and Universities:
‘What percentage of the total budget of your organisation is allocated for RDM and RDI?’

Possible responses: 0 - don’t know, <10%, <20%, <30%, <40%, <50%, <60%, <70% or <80%

Interpretation: The lack of clarity about allocations to RDM and RDI is shown clearly: a substantial share of respondents cannot specify a budget allocation.

Percentage of the total budget allocated for RDI

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - Don’t know</td>
<td>2%</td>
</tr>
<tr>
<td>&lt;10%</td>
<td>9%</td>
</tr>
<tr>
<td>&lt;20%</td>
<td>13%</td>
</tr>
<tr>
<td>&lt;30%</td>
<td>20%</td>
</tr>
<tr>
<td>&lt;40% - 80%</td>
<td>56%</td>
</tr>
</tbody>
</table>

Percentage of the total budget allocated for RDM

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - Don’t know</td>
<td>33%</td>
</tr>
<tr>
<td>&lt;10%</td>
<td>11%</td>
</tr>
<tr>
<td>&lt;20%</td>
<td>11%</td>
</tr>
<tr>
<td>&lt;30%</td>
<td>9%</td>
</tr>
<tr>
<td>&lt;40% - 80%</td>
<td>35%</td>
</tr>
</tbody>
</table>
Life cycle – ELIXIR Services

- Indicators guide and inform the managers of Emerging Services
- Capacity building support from ELIXIR community
- Monitoring of usage trends and the scientific impact provides information to support life cycle management
Open Data has a large value

The economic benefit of Landsat data for the year 2011 was estimated to be $1.70 billion for US users and $400 million for international users resulting in a total annual value of $2.19 billion.