Life Sciences Datasets Questionnaire

Integrating ELIXIR reference datasets within the European Grid Infrastructure

# Introduction

In bioinformatics many services used for analysis purposes rely on public reference datasets. Reference dataset are getting big and users struggle to discover, download and compute them. This increase is accompanied with an increasing demand to compute the data where the reference datasets are located.

Recognizing the need to tackle the data processing issue the EGI e-infrastructure and ELIXIR research infrastructure communities launched a pilot project in December 2014. The project aims to: a) facilitate the discovery of existing reference datasets that are already hosted on EGI resources; b) develop and deploy services that allow the replication of life science reference datasets on the EGI e-infrastructure; c) ease the use of these replicated datasets for life science researchers and their analysis applications.

## About EGI

The European Grid Infrastructure (EGI) delivers integrated computing and storage services to European researchers, driving innovation and enabling new solutions to answer the big questions of tomorrow. EGI is a publicly funded e-infrastructure that includes resources provided by about 350 resource centres who are distributed across 56 countries in Europe, the Asia-Pacific region, Canada and Latin America.

## About ELIXIR

ELIXIR unites Europe’s leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information. ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians – to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built.

## Purpose of this survey

This survey was setup to collect information from life science communities about reference datasets that are commonly used by them and that would benefit from being replicated onto EGI resources to enable scalable access and processing for research.. Going beyond a simple listing of frequently used datasets, the survey also focuses on capturing technical information on these datasets, as well as information on software tools and platforms commonly applied for their analysis.

Feedback from the survey will help the project define focus areas priorities ensuring maximum impact for life science research communities.

## Further information

* Pilot Project Page: https://wiki.egi.eu/wiki/Integrating\_Reference\_Datasets
* EGI: <http://egi.eu>
* ELIXIR: http://www.elixir-europe.org/

# Definitions

In order to better understand and answer the following questions, the following definitions are being used:

**Dataset**: a collection of data that pertains to a single entity. A dataset can be defined in the form of a database table, as a database, as a file or set of files, a single statistical data matrix, etc.

A few examples for life science datasets (focusing mostly on flat-file datasets): BLAST Datasets (NT/NR, UniRef), PDB Datasets, 1000 Genomes, etc.

**Database**: an organized collection of data, usually implemented via a traditional DBMS (such as MySQL, PostgreSQL, Oracle etc).

**Data Repository**: a logical collection of data from related but different databases. Most commonly, data repositories relate also to the physical grouping of databases (co-location).

# Questions

## Section A: Background

**A1a**. Affiliation – Organization Type

1. | | Research Center
2. | | University
3. | | Foundation
4. | | Government
5. | | Company
6. | | Other

**A1b**. Affiliation – Country

**A1c**. Providing answers on behalf

1. | | Myself
2. | | Research group
3. | | Project + Name it: …
4. | | Community/collaboration/experiment + Name it: …
5. | | Other + Name it: …

## Section B: Work with Reference Datasets

**B1**. How often do you use reference dataset for your work?

1. | |At least a few times a week
2. | |Approximately once a week
3. | |A few times a month
4. | |Less than once a month
5. | | Never

**B2**. How do you access the reference dataset?

1. | | Publicly available
2. | | Access is granted through my team/project/collaboration
3. | | I arranged access directly with data provider

**B3**. Where do you perform your analysis run work with reference datasets?

1. | | On my laptop/desktop
2. | | On a server/mainframe/cluster in my institute
3. | | On a server/mainframe/cluster in a partner institute
4. | | On a commercial infrastructure/cloud (e.g. Amazon, Microsoft Azure)
5. | | On an e-infrastructure (e.g. EGI, PRACE, Embassy Cloud, Nordugrid, etc.)

**B4**. On average, how significant is the data transfer bottleneck in your analysis when you work with reference datasets?

1. | |very significant - my analysis is very slow because of slow data download/upload
2. | |significant - my analysis is often impacted by slow data download/upload rate
3. | |insignificant - data transfer rate has little, or no impact at all on my analysis

**B5**. On average, how significant in the computational bottleneck in your analysis when you work with reference datasets?

1. | |very significant - my analysis is very slow because of complex computational algorithms
2. | |significant - my analysis is often slow because of complex computational algorithms
3. | |insignificant – my analysis on reference data does not require complex computational algorithms

## Section C: Datasets, Databases, Data Repositories

**C1**. Please list your most frequently used reference datasets (or databases, data repositories that provide such datasets). Examples of such datasets are NR/NT, RefSeq, UniRef100, Ensembl Plants, Plaza, 1000genomes, MouseDB etc.

**<**freeform text answer >

**C2**. What is your preferred mode of access for those datasets/databases/repositories?

1. | |Flat file (FASTA, FASTQ, PDB, etc)
2. | |NoSQL approaches
3. | |Through provided API
4. | |Direct database connection
5. | |Software Platform (e.g. Galaxy)

**C3.** Which is your preferred means of interacting with the reference data:

1. | |Custom (in-house) developed tools
2. | |Existing platforms and services (e.g. Galaxy)
3. | |Programming frameworks (e.g. Bioconductor)
4. | |Other:

**C4:** How useful would be the integration of selected Data Repositories within the European Grid Infrastructure?

1. | |very useful
2. | |useful
3. | |insignificant
4. | |I don’t know enough about the topic to judge this

## Section D: Comments and Follow-up

**D1**. Other Comments (e.g. further explanation to some of the answers you provided)

**D2**. Contact information