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 with them. There is an increasing demand to compute the data where the reference datasets are located.

## Pilot Scope

Recognizing the need for this kind of support the EGI and ELIXIR communities defined and launched this pilot project in December 2014. The project aims to facilitate the discovery of existing reference datasets in EGI and will develop and deploy services that allow the replication of life science reference datasets by data providers, resource providers and researchers, and also ease the use of these datasets by life science researchers within analysis applications. The project is expected to run for 9 months and includes tasks such as Life Science datasets identification and replication, Data Analysis Tools, EGI AppDB extension to dataset registry and Integration with the ELIXIR Registry.

## Target audience and expected outcomes

This survey is aimed towards the identification of datasets more commonly used by Life Sciences researchers. Going beyond a simple listing of frequently used datasets, the survey focuses on gathering information on dataset use within HPC infrastructures, as well as information on software tools and platforms commonly employed for data analysis.

Feedback from the survey will affect the focus and prioritization on particular datasets at the initial stages of the Pilot, while ensuring maximum impact for the researchers’ community.

## Additional Resources

* Pilot Wiki Page: https://wiki.egi.eu/wiki/Integrating\_Reference\_Datasets
* Dedicated DocDB: https://documents.egi.eu/public/ShowDocument?docid=2361
* Pilot Mailing List: https://mailman.egi.eu/mailman/listinfo/elixir-pilot

# 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. Most commonly, a dataset is defined as the content of a single database table, or a single statistical data matrix.

Within the context of the EGI-ELIXIR pilot, examples of Datasets are the following (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

**A1**. How often do you use publicly available data for your work?

1. Many times every day
2. About once a day
3. A few times a week
4. Once a week
5. A couple times a month
6. Rarely / Never

**A2**. In your work, how much do you depend on publicly available data (as opposed to in-house produced data):

1. very dependent
2. dependent
3. neutral
4. independent
5. very independent

**A3**. On average, how significant the computational bottleneck is when using public data access and utilization in your work?

1. very significant
2. significant
3. neither significant nor insignificant
4. insignificant
5. very insignificant

## Section B: Data Repositories

**B1.** Please rank the following Data Repositories in terms of access frequency for your work?

1. NCBI (<http://www.ncbi.nlm.nih.gov/>)
2. Ensembl (<http://www.ensembl.org/index.html>)
3. ΕΒΙ (<http://www.ebi.ac.uk/>)
4. UniProt (<http://www.uniprot.org/>)
5. PDB (<http://www.rcsb.org/>)
6. Other

**B2**. What is your preferred mode of access for those Data 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)

**B3**. Do you maintain locally an instance of a public Data Repository?

* Yes
* No

If “***Yes***”:

* Do you employ an HPC infrastructure at your institution? (Yes/No)
* How useful would be the integration of selected Data Repositories within the European Grid Infrastructure?

1. very useful
2. useful
3. neither useful nor insignificant
4. insignificant
5. very insignificant

## Section C: Databases and Datasets

**C1**. Please list your 3 most frequently used publicly available Databases:

**C2**. What is the average size of a data set used in your work:

1. Less than 10 MB
2. 10-100 MB
3. 1-10 GB
4. 10-100 GB
5. Over 1 TB

**C3**. Which is your preferred means of using this data:

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

**C4**. Please list some of your most commonly used tools for using this data: