Life Sciences Databases Questionnaire

Integrating ELIXIR reference datasets within the European Grid Infrastructure

# Definitions

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

**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). Within the context of the EGI-ELIXIR pilot, examples of Data Repositories are the following:

**Database**: an organized collection of data, usually implemented via a traditional DBMS (such as MySQL, PostgreSQL, Oracle etc). Within the context of the EGI-ELIXIR pilot, examples of Databases are the following:

**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:

# 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. Through provided API
3. Direct database connection
4. 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: