Life Sciences Datasets Questionnaire: responses – v1.0 *(28/02/2015)*

Integrating ELIXIR reference datasets within the European Grid Infrastructure

Contents

[Introduction 2](#_Toc412891515)

[Response Summary 2](#_Toc412891516)

[Detailed Survey Responses 3](#_Toc412891517)

[Section A: Background 3](#_Toc412891518)

[A1a. Affiliation – Organization Type 3](#_Toc412891519)

[A1b. Affiliation – Country 3](#_Toc412891520)

[A1c. Providing answers on behalf 4](#_Toc412891521)

[Section B: Work with Reference Datasets 5](#_Toc412891522)

[B1. How often do you use reference dataset for your work? 5](#_Toc412891523)

[B2. How do you access the reference dataset? 5](#_Toc412891524)

[B3. Where do you perform your analysis run work with reference datasets? 6](#_Toc412891525)

[B4. On average, how significant is the data transfer bottleneck in your analysis when you work with reference datasets? 7](#_Toc412891526)

[B5. On average, how significant in the computational bottleneck in your analysis when you work with reference datasets? 8](#_Toc412891527)

[Section C: Datasets, Databases, Data Repositories 9](#_Toc412891528)

[C1. Please list your most frequently used reference datasets (or databases, data repositories that provide such datasets). 9](#_Toc412891529)

[C2. What is your preferred mode of access for those datasets/databases/repositories? 10](#_Toc412891530)

[C3. Which is your preferred means of interacting with the reference data: 11](#_Toc412891531)

[C4: How useful would be the integration of selected Data Repositories within the European Grid Infrastructure? 11](#_Toc412891532)

[Section D: Comments and Follow-up 13](#_Toc412891533)

[D1. Other Comments 13](#_Toc412891534)

[D2. Contact information (confidential) 13](#_Toc412891535)

# Introduction

This documents provides the details results of the survey regarding the reference datasets in Life Sciences.

The document of the survey can be found [here](https://documents.egi.eu/secure/RetrieveFile?docid=2361&version=2&filename=LS%20DB%20Questionnaire%20EGI-ELIXIR%20Pilot_v1.0_final.docx). The URL of the survey is [here](https://survey.egi.eu/index.php/652565/lang-en).

The current version (1.0) of this document was produced with the results retrieved up until (and including) February 28th 2015.

## Response Summary

|  |  |
| --- | --- |
| Full responses | 27 |
| Incomplete responses | 37 |
| **Total responses** | **64** |

The results listed below are based on the Full Responses (27).

# Detailed Survey Responses

## Section A: Background

### A1a. Affiliation – Organization Type

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| Research Center | 11 | 40.74% |
| University | 13 | 48.15% |
| Foundation | 1 | 3.70% |
| Government | 1 | 3.70% |
| Company | 0 | 0.00% |
| Other | 1 | 3.70% |



*Other*

* Non profit organization

### A1b. Affiliation – Country

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| Canada | 1 | 3.70% |
| Denmark | 1 | 3.70% |
| Finland | 1 | 3.70% |
| France | 2 | 7.40% |
| Greece | 9 | 33.33% |
| Israel | 1 | 3.70% |
| MAC | 1 | 3.70% |
| Netherlands | 5 | 18.51% |
| Portugal | 1 | 3.70% |
| Spain | 1 | 3.70% |
| Switzerland | 2 | 7.40% |
| UK | 1 | 3.70% |
| USA | 1 | 3.70% |

### A1c. Providing answers on behalf

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| Myself | 20 | 74.07% |
| Research group | 4 | 14.81% |
| Project (please name it) | 1 | 3.70% |
| Community/collaboration/experiment (please name it) | 0 | 0.00% |
| Other (please name) | 2 | 7.41% |

*Comments/Other*

* CHAIN-REDS
* I try to answer on behalf of the science services of CSC.
* ELIXIR-DK



## Section B: Work with Reference Datasets

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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| At least a few times a week | 18 | 66.67% |
| Approximately once a week | 2 | 7.41% |
| A few times a month | 3 | 11.11% |
| Less than once a month | 4 | 14.81% |
| Never | 0 | 0.00% |



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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| Publicly available | 25 | 92.59% |
| Access is granted through my team/project/collaboration | 0 | 0.00% |
| I arranged access directly with data provider | 1 | 3.70% |
| Other | 1 | 3.70% |

*Other*

* I use the CHAIN-REDS Knowledge Base and Semantic Search Engine



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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| On my laptop/desktop | 8 | 29.63% |
| On a server/mainframe/cluster in my institute | 12 | 44.44% |
| On a server/mainframe/cluster in a partner institute | 3 | 11.11% |
| On a commercial infrastructure/cloud (e.g. Amazon, Microsoft Azure) | 0 | 0.00% |
| On an e-infrastructure (e.g. EGI, PRACE, Embassy Cloud, Nordugrid, etc.) | 2 | 7.41% |
| Other | 2 | 7.41% |

*Other*

* Laptop/server/EGI
* On a Surfsara cloud



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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| very significant - my analysis is very slow because of slow data download/upload | 3 | 11.11% |
| significant - my analysis is often impacted by slow data download/upload rate | 13 | 48.15% |
| insignificant - data transfer rate has little, or no impact at all on my analysis | 11 | 40.74% |



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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| very significant - my analysis is very slow because of complex computational algorithms | 6 | 22.22% |
| significant - my analysis is often slow because of complex computational algorithms | 16 | 59.26% |
| insignificant – my analysis on reference data does not require complex computational algorithms | 5 | 18.52% |



## 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Answer** | **Count** | **Percentage** | **Notes** |
| RefSeq, RefSeq Bacteria | 8 | 29.62% | Sequence Database |
| NR/NT | 7 | 25.92% | Sequence Database |
| Ensembl  | 7 | 25.92% | Sequence Database |
| UniProtKB Reference Proteomes | 6 | 22.22% | Sequence Database |
| PDB | 4 | 14.81% | Structural Database |
| UniRef100 | 3 | 11.11% | Sequence Database |
| SRA | 3 | 11.11% | Sequence Database |
| 1000genomes | 2 | 7.40% | Sequence Database |
| ucsc genome browser tables, refseq, exons, conservation tracks  | 2 | 7.40% | Sequence Database |
| KEGG | 2 | 7.40% | Metabolic Database |
| GEO, GEO DATASETS | 2 | 7.40% | Gene Expression Database |
| OMIM | 2 | 7.40% | Sequence Database |
| chEMBL | 2 | 7.40% | Reaction Database |
| GENCODE, ENCODE Mouse and Human  | 2 | 7.40% | Sequence Database |
| JGI Fungi Portal | 1 | 3.70% | Species Database |
| COSMIC | 1 | 3.70% | Cancer Database |
| NCBI Pubmed/Pubmed central | 1 | 3.70% | Text Database |
| MouseDB | 1 | 3.70% | Species Database |
| HGNC | 1 | 3.70% | Sequence Database |
| HAMAP | 1 | 3.70% | Sequence Database |
| PROSITE | 1 | 3.70% | Sequence Database |
| RegulonDB | 1 | 3.70% |  |
| RSAT | 1 | 3.70% |  |
| TRANSFAC | 1 | 3.70% |  |
| MMDB | 1 | 3.70% |  |
| GENbank | 1 | 3.70% | Sequence Database |
| CHAIN-REDS | 1 | 3.70% |  |
| Enclyclopedia of Life (Text content)  | 1 | 3.70% | Text Database |
| Medline abstracts | 1 | 3.70% | Text Database |
| ENA | 1 | 3.70% |  |
| EGA | 1 | 3.70% |  |
| GoNL | 1 | 3.70% |  |
| RNAseq | 1 | 3.70% | Sequence Database |
| nextprot  | 1 | 3.70% |  |
| Drugbank | 1 | 3.70% |  |

The percentage has been calculated based on the number of different responses (i.e. 27), and not with regard to the number of different entries (total 73, multiple entries per response).

Finally, there were more generic responses that are not included in the above tables, such as:

* NCBI (contains many different repositories)
* sequence databases, BLAST databases
* Ontologies (OWL format)
* Bird genomes, various sources
* Plant genomes, various sources

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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| Flat file (FASTA, FASTQ, PDB, etc) | 23 | 85.19% |
| NoSQL approaches | 0 | 0.00% |
| Through provided API | 2 | 7.41% |
| Direct database connection | 2 | 7.41% |
| Software Platform (e.g. Galaxy) | 0 | 0.00% |



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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| Custom (in-house) developed tools | 15 | 55.56% |
| Existing platforms and services (e.g. Galaxy) | 7 | 25.93% |
| Programming frameworks (e.g. Bioconductor) | 3 | 11.11% |
| Other | 2 | 7.41% |

*Other*

* Custom + various software
* and Galaxy sometimes



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

|  |  |  |
| --- | --- | --- |
| **Answer** | **Count** | **Percentage** |
| very useful | 14 | 51.85% |
| useful | 6 | 22.22% |
| insignificant | 1 | 3.70% |
| I don’t know enough about the topic to judge this | 6 | 22.22% |



## Section D: Comments and Follow-up

### D1. Other Comments

e.g. further explanation to some of the answers you provided

|  |  |
| --- | --- |
|  | **Other Comments** |
| **1** | tools to download specific slices of the database (as tabix for 1000G) are welcome. |
| **2** | We should be able to select more than one answers in a few of your questions. |
| **3** | I was originally a researcher in Europe and recently moved to USA. I still use European tools, mainly through ExPASy. As a mere user, I rely these databases and tools (though my knowledge is very limited in programming and how these work) |
| **4** | The biggest problem in accessing data is the dispersed nature of meta data. For example: bacterial genome assemblies can be obtained from NCBI (although not without scripting efforts), but the corresponding meta data about the bacterial strains is available via the GOLD database. The GOLD database is not accessible via an API or flatfiles so analysis over a large number of genomes is virtually impossible without a lot of hand work. |
| **5** | The datasets need annotating in consistent terms, at least "EDAM format" and possible "EDAM data" for bio datasets. |
| **6** | We use the nt/nr database at the moment via the cloud where we stored the database locally. The ideal situation would be that we could access the database from a shared resource though which we could work on either a cluster/cloud/galaxy because the storage place of the database is one of the biggest limitations |

### D2. Contact information (confidential)

|  |  |
| --- | --- |
|  | **Contact Emails** |
| **1** | panagiotis.ioannidis@unige.ch |
| **2** | anastasia.gioti@biology.uoc.gr |
| **3** | katnastou@biol.uoa.gr |
| **4** | lilachfr@gmail.com |
| **5** | cnikolaou@biology.uoc.gr |
| **6** | apolyzos@bioacademy.gr |
| **7** | epk34@drexel.edu |
| **8** | b.b.oudemunnink@amc.uva.nl |
| **9** | v.dejager@nioo.knaw.nl |
| **10** | mntertilis@biol.uoa.gr |
| **11** | mariakamilari@gmail.com |
| **12** | rafael.mayo@ciemat.es |
| **13** | aduarte@itqb.unl.pt |
| **14** | kimmo.mattila@csc.fi |
| **15** | jison@ebi.ac.uk |
| **16** | b.b.oudemunnink@amc.uva.nl |