7th Meeting of the VT Project on

Life Science Reference Data Integration

Integrating ELIXIR reference datasets within the European Grid Infrastructure

**Date**: Tue 13 May 2015 (13/05/2015) 12:00 CET – Webex Teleconference

Contents

[Participants 1](#_Toc419294026)

[Agenda 1](#_Toc419294027)

[Minutes 2](#_Toc419294028)

[Discussion / Progress report 2](#_Toc419294029)

[Planned actions 4](#_Toc419294030)

[Useful Links 4](#_Toc419294031)

# Participants

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** |  | **Affiliation** | **Country** |
| Fotis E. Psomopoulos | FP | AUTH | GR |
| Afonso Duarte | AD | ITQB | PT |
| Marios Chatziangelou | MC | IASA | GR |
| William Karageorgos | WK | IASA | GR |
| Diego Scardaci | DS | INFN/EGI.eu | NL |
| Giacinto Donvito | GD | INFN Bari | IT |

# Agenda

* Review of the agreed actions (selected subset from ensembl / elixir services, OneData approach, etc - see also minutes from last teleconference)
* Decision for next steps / actions towards a first implementation

# Minutes

## Discussion / Progress report

**Actions from previous meeting**

*1*: Specific subsets of ensembl by following up with the users from the survey

* ***FP***: Follow-up email to be sent to all (see detailed list [here](https://docs.google.com/document/d/1iqgCFKLYr12EfS3QQk1UX3_aeBC15lgnB8iSDJG-0Y4/edit#heading=h.x34i4bq2duqo)). Additional details to be requested from users include: specific subset(s), corresponding tools/applications to be used in conjunction.

*2*: Identify in [ELIXIR registry](https://elixir-registry.cbs.dtu.dk) whether (and which) subsets are already listed

* ***FP***: Registry currently contains subsets, but not in the form of direct access to the respective datasets but as tools that are directly applied to those. As an example, the [plants subset of ensembl](http://plants.ensembl.org/) is not provided as a dataset but only through targeted applications such as [BLAST](https://elixir-registry.cbs.dtu.dk/#/tool/EBI/Ensembl%20Genomes%20Plants%20Blast/1) and [Sequence search](https://elixir-registry.cbs.dtu.dk/#/tool/EBI/Ensembl%20Genomes%20Plants%20sequence%20search/1).
* ***FP***: However we can try and provided the corresponding datasets as [direct flat files](http://plants.ensembl.org/info/website/ftp/index.html) (according also to the specification from the users derived by the survey).

*3*: Identify in ELIXIR registry whether (and which) index files are listed. Alternatively, we can use the Homo Sapiens index data as a test case (Test Case 2 for the VT)

* ***FP***: The ELIXIR registry doesn't currently provide any entries with such [index files](http://ccb.jhu.edu/software/tophat/igenomes.shtml). However, they do provide links for the corresponding tools that employ them, such as [Bowtie](https://elixir-registry.cbs.dtu.dk/#/tool/DebianMed/bowtie%20(full%20annotation)/1.1.1) and [TopHat](https://elixir-registry.cbs.dtu.dk/#/tool/IFB%20ELIXIR-FR/TopHat%20-%20cloud%20IFB/2.0.13) (same situation as with subsets previously #2).

*4*: Follow-up with OneData as an option for replicating data across Grid and Cloud resources.

* ***GD***: Currently working on a Task Force for the Human Brain project that has complementary approaches. They have established a test-bed, applicable to both Grid and Cloud solutions, that can be used also for the work of this VT.

**Discussion**

***FP***: Pending final confirmation from the discussions in the EGI Conference in Lisbon, the Ensembl Plant subset can be used as Test Case 1. What are the next steps towards an implementation?

***GD***: We need a list of applications that can be used with the specific dataset, in order to test and evaluate the replication process. This means that we require at least a high-level workflow of applications that can use the replicated data.

***FP***: For the Ensembl dataset (Test Case 1), we will start a second round of communication asking also about the applications to be used with the proposed datasets (new Action 1). Regarding the applications, the most basic one will be BLAST with the rest to be defined by the users.

***FP***: We can establish two test cases for this VT.

*Test Case 1*: Ensembl data (plant database, to be confirmed). Widely used scenario, general applicability, however it will be difficult to define genera workflows (other than the first steps)

*Test Case 2*: Index data files (NGS analysis). Necessary for most NGS analyses, and a well defined set of tools that use these data files. However, relatively smaller target group.

***DS***: Can we re-use the data of the Chipster test case?

***FP***: It depends on the particular implementation. Usually there is no change to the index files, as they are produced from reference data. Format is also a standard, as it used by specific tools. However, Chipster may have included custom libraries and / or specific data access permissions, so we have to consult with them (new Action 2).

***MC***: The important data source for us is the replica and not the master.

***FP***: Correct. The master source usually will be outside of EGI resources. Therefore we can expect that the first replica will be with EGI, as well as any consequent replicas. We focus on the techniques and methods that combine the replication with the respective tools.

***DS***: It is also important to know the exact interface of the involved tools with the specific dataset. Depending on the interface, we may decide that a replication will take place for a Cloud infrastructure (such as in the case of using the dataset with a full path), or for a Grid infrastructure. In any case, the list of applications should also include the data interfaces necessary.

## Planned actions

*Action 1*: Round of communication with users regarding particular subset as well as usual tools/applications used for Test Case 1. (***FP***)

*Action 2*: Communication with the Chipster team to establish availability and/or connectivity to the index files Test Case 2 (***FP***)

*Action 3*: For each Test Case, provide a list of the applications (with characteristics, URL to tool, interface to data etc), as well as sample workflows. (***FP***)

*Face-to-face meeting in EGI Conference*: Wednesday, May 20th, 15:00 - 15:30 Lisbon time

*Next teleconference*: Wednesday, June 3rd, 12:00 Amsterdam time

# Useful Links

**Pilot Wiki Page**: <https://wiki.egi.eu/wiki/VT_Life_Science_Data_Integration>

**Dedicated DocDB**: <https://documents.egi.eu/public/ShowDocument?docid=2361>

**Pilot Mailing List**: <https://mailman.egi.eu/mailman/listinfo/lsrdi-vt>

**Resources Google Sheet**: [Google Sheet](https://docs.google.com/spreadsheets/d/1gcSNCNBdwIgNN8URrj0b2w4WT18VA04BHOi5csT0HGM/edit#gid=1080851537)