6th Meeting of the VT Project on

Life Science Reference Data Integration

Integrating ELIXIR reference datasets within the European Grid Infrastructure

**Date**: ~~Tue 21 Apr 2015 (21/04/2015) 12:00 CET – Webex Teleconference~~ (*rescheduled*)

 : Mon 27 Apr 2015 (27/04/2015) 12:00 CET – Webex Teleconference

Contents

[Participants 2](#_Toc417909515)

[Agenda 2](#_Toc417909516)

[Minutes 3](#_Toc417909517)

[Part A: Short Discussion on Tuesday, April 21st 3](#_Toc417909518)

[Part B: Overview of the general actions for the pilot (April 27th) 4](#_Toc417909519)

[Part C: Discussion and further Actions 4](#_Toc417909520)

[Useful Links 7](#_Toc417909521)

# Participants

|  |
| --- |
| ***Tue 21 Apr 2015*** |
| **Name** |  | **Affiliation** | **Country** |
| Fotis E. Psomopoulos | FP | AUTH | GR |
| Afonso Duarte | AD | ITQB | PT |
| Diego Scardaci | DS | INFN/EGI.eu | NL |
| Giacinto Donvito | GD | INFN Bari | IT |
| ***Mon 27 Apr 2015*** |
| **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 |

# Agenda

* Progress report on the 6 tasks
* Discussion for next steps / actions (incl. a face-to-face meeting at the EGI Conference in Lisbon)

# Minutes

## Part A: Short Discussion on Tuesday, April 21st

Points raised by Gergely Sipos through email:

* Progress of the change in project name (wiki page different, no change in mailing list)
* Identifying who we need to interview from those who answered the survey (***FP***), arrange interviews (?)
* Talk to EBI (because many datasets are there) - ***Gergely***. Update: I wrote to Steven last week to ask for a Skype meeting about this. I hope to get this call later this week.
* Deploy suitable tools to the EGI-Engage testbed (JRA2.1 ***Lukasz Dutka***, ***GD***)

***FP***: The change in wiki page was successful. Change in mailing list requires a change in the SSO group; leads to issues with access permissions to the existing resources. Still investigating with the EGI IT support, should be resolved by 27/04/2015.

***FP***: Prepared a list of the current contact points from the circulated survey

[https://docs.google.com/document/d/1iqgCFKLYr12EfS3QQk1UX3\_aeBC15lgnB8iSDJG-0Y4/edit#](https://docs.google.com/document/d/1iqgCFKLYr12EfS3QQk1UX3_aeBC15lgnB8iSDJG-0Y4/edit%23)

Highlighted people that requested Ensembl data for replication. Still need to investigate whether Refseq and/or NR/NT are available through EBI.

***AD***: "all" dbs from EBI at <http://www.ebi.ac.uk/services/proteins> and some use provide access to NR

Actions:

1. (***FP*** & ***GS***?) discuss with EBI if RefSeq and NR/NT are also provided by them
2. (***FP*** & ***GS***?) setup a list of questions for the contact points regarding the datasets

***GD***: The table on the technologies existing for data replication is ready

[https://docs.google.com/spreadsheets/d/1fftRZxBd1NSeSK8y8YIhts28TrTpTcJIQ52JMjtPPBI/edit?pli=1#gid=0](https://docs.google.com/spreadsheets/d/1fftRZxBd1NSeSK8y8YIhts28TrTpTcJIQ52JMjtPPBI/edit?pli=1%23gid=0)

***GD***: as soon as the reference datasets are selected, we can proceed with testing them with these technologies

***FP***: setup a face-to-face meeting for the participants/interested parties in the VT at the EGI conference in Lisbon. Should take place in the beginning of the week, ideally before the ELIXIR CC session.

***FP***: due to limited time should take place during one of the coffee breaks, or before/after a lunch break.

Action:

1. ***FP*** Setup a doodle poll for setting up the meeting (<http://doodle.com/ugvqi2cpibvwsuzt>).

## Part B: Overview of the general actions for the pilot (April 27th)

**1. Change of Project Name**

The project has been renamed, as well as the corresponding resources

1. *Mailing list*: lsrdi-vt@mailman.egi.eu (life science reference data integration VT)
2. *Wiki Page*: https://wiki.egi.eu/wiki/VT\_Life\_Science\_Data\_Integration
3. *SSO Group*: lsrdi-vt

## Part C: Discussion and further Actions

**Overview of the current status**

1. Identification of reference dataset: we have a draft list of potential datasets, with the key dataset emerging to be ensembl (<http://www.ensembl.org/>).
2. AppDB interface: we have a first implementation of the AppDB front- and back-end (<https://appdb-dev.marie.hellasgrid.gr>), with connection to the Elixir Dataset registry (<https://elixir-registry.cbs.dtu.dk>) as well as the respective data sources.
3. We have a list of the technical tools required for implementation of the data replication ([https://docs.google.com/spreadsheets/d/1fftRZxBd1NSeSK8y8YIhts28TrTpTcJIQ52JMjtPPBI/edit?pli=1#gid=0](https://docs.google.com/spreadsheets/d/1fftRZxBd1NSeSK8y8YIhts28TrTpTcJIQ52JMjtPPBI/edit?pli=1%23gid=0))
4. We have a list of tools that can be used in conjunction with LS Reference Datasets.

***FP***: Given that we are in M5 of the VT project, what are the next steps necessary for a first implementation?

***MC***: If Ensembl is an identified dataset, are we using flat files are the entirety of the database? It should be noted that the ELIXIR data registry identifies "datasets" as complete databases, which might be an issue for us.

***FP***: Although Ensembl is a complete database system (also accessible as a MySQL dump), we will focus only on flat files which can be produced through the Ensembl interface. This decision is both from the requirements of the VT project as well as supported by the users' survey feedback.

***MC***: Ultimately AppDB and the ELIXIR registry will be connected bidirectionally; datasets added in AppDB will be listed in the ELIXIR registry and vice versa.

***FP***: Is there a technical problem in providing through AppDB a subset of an already listed "complete" dataset?

***WK***: There is already support for such dependencies in AppDB; a subset is defined as "derived" from a different entry

***FP***: We should identify:

* *Action 1*: Specific subsets of ensembl by following up with the users from the survey
* *Action 2*: Identify in ELIXIR registry whether (and which) subsets are already listed

***DS***: There was some talk with ELIXIR people (including CSC) and there is some interest in specially formatted data (such as indexes for NGS analysis)

***FP***: It is indeed an interesting option

* *Action 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)

***DS***: It is critical to find which users are going to actually use the replicated data. How can we exposed this data to the user?

***FP***: What are the current technical options of providing access to the data through the EGI infrastructure?

***DS***: There are three options:

1. create a new VO that corresponds to the VT to be used as a testbed
2. use an existing VO / Community (a good example is our case is biomed / LSGC)
3. use the Cloud platform approach and employ the catch-all VO; select a couple of sites and setup a block storage.

***FP***: Option c seems to be the easiest and most convenient one

***MC***: Such an option has several problems; there is an issue of available storage (as we are talking about large files), data will not be visible to other VOs or to Grid resources, etc. We should aim for more general options.

***DS***: We can share data within multiple VOs. In the Cloud infrastructure we have 2 kind of storages: block storage and object storage. The latter exposes a CDMI interface. We already tested integration between cloud and grid (processing in cloud and data stored in grid) but it is quite complex to be implemented by people without grid knowledge and only interested to cloud.

***MC***: We are discussing about having 1 replica for all potential technologies. If we use 1 replica for Grid resources and 1 replica for Cloud computing, we duplicate the effort in technologies, we duplicate the effort in maintaining the updates and we duplicate the necessary storage. It's not an efficient approach

***DS***: We should look for high-level approaches such as OneData

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

*Actions for next period*

* *Action 1*: Specific subsets of ensembl by following up with the users from the survey
* *Action 2*: Identify in ELIXIR registry whether (and which) subsets are already listed
* *Action 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)
* *Action 4*: Follow-up with OneData as an option for replicating data across Grid and Cloud resources.

*Next teleconference*: Wednesday, May 13th, 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)