4th Meeting of the EGI-ELIXIR Pilot

Integrating ELIXIR reference datasets within the European Grid Infrastructure

**Date**: Wed 25 Feb 2015 (25/02/2015) 12:00 CET – Webex Teleconference

Contents

[Participants 2](#_Toc409094196)

[Agenda 2](#_Toc409094197)

[Minutes 3](#_Toc409094198)

[Part A: Overview of the general actions for the pilot 3](#_Toc409094199)

[Part B: Progress report in Pilot Tasks 4](#_Toc409094200)

[Task 1: Identify existing life science datasets in EGI 4](#_Toc409094201)

[Task 2: Identify reference datasets for replication 4](#_Toc409094202)

[Task 3: EGI AppDB extension to a dataset registry 5](#_Toc409094203)

[Task 4: Tools for data replication 5](#_Toc409094204)

[Task 5: Analysis tools to work with data replicas 6](#_Toc409094205)

[Task 6: Integration with ELIXIR Registry 6](#_Toc409094206)

[Part C: Discussion and further Actions 7](#_Toc409094207)

[Useful Links 7](#_Toc409094208)

# Participants

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** |  | **Affiliation** | **Country** |
| Fotis E. Psomopoulos | FP | AUTH | GR |
| Afonso Duarte | AD | ITQB | PT |
| Diego Scardaci | DS | EGI.eu | NL |
| Gergely Sipos[[1]](#footnote-1) | GS | EGI.eu | NL |
| Giacinto Donvito | GD | INFN Bari | IT |
| Brane Leskosek | BL | NGI\_SI | SI |
| Lukasz Dutka | LD | CYFRONET | PL |
| Marios Chatziangelou | MC | IASA | GR |
| William Karageorgos | WK | IASA | GR |

# Agenda

* Review of the agreed actions
* Progress report on the 6 tasks
* Discussion for next steps / actions

# Minutes

## Part A: Overview of the general actions for the pilot

**1. ELIXIR Publicity**

* EGI has published a newsletter article regarding the Pilot (together with FP and Rafael Jimenez).

The article can be found here:

<http://www.egi.eu/news-and-media/newsletters/Inspired_Issue_18/egi-elixir.html>

The article will be cross-posted and promoted within the ELIXIR network.

* *FP*: to follow-up on the article visibility

**2. ELIXIR Letters of Endorsement**

*FP* is in communication with Nicola Kay (ELIXIR). The EGI-ELIXIR pilot is not yet an officially acknowledge pilot in ELIXIR, but it remains within the ELIXIR action portfolio as a potential activity for beginning 2015. Letters of endorsement from the involved ELIXIR Nodes will be helpful in further promoting the pilot.

The 2015 ELIXIR all-hands meeting will take place on Monday 30th and Tuesday 31st March 2015 at the Wellcome Trust Genome Campus, Hinxton, UK

* *GD* and *BL* will be attending the event (INFN and Head of SI Node respectively)

Progress

* *FP* has acquired a Letter of Endorsement from the ELIXIR-GR Node (GR Node status is Observer)
* *AD* with Joao Pina contacted the ELIXIR-PT Node, however no response so far
	+ There is an upcoming National Event on ELIXIR → follow-up communication then.
* *DS* has requested a Letter of Endorsement from the Swedish ELIXIR Node (BILS)
	+ No response as of yet → follow-up communication
* *DS* has requested a Letter of Endorsement from the Finnish ELIXIR Node (CSC)

*FP*: There has been some communication between *FP*, *GS*, Tommi Nyrönen (Head of Finnish Node) and Niklas Blomberg (Director of ELIXIR). Though there are potential synergies for the ELIXIR-endorsed EGI pilot, it should evolve a bit before the Finnish node endorses it. Ideally, we should wait until the Competence Center starts and then make improvements on what is needed for the scientific use case. In particular, the strategic actions for the EGI-ELIXIR pilot (also endorsed by Niklas Blomberg) are the following:

*SA1*: we coordinate ELIXIR-EGI service pilot interactions via the Engage Competence Center

*SA2*: CC project leading to e.g. ELIXIR pilot starts in 1st March and CC in M6 with kick-off in Lisbon in May

Additionally, there were the following recommendations regarding the actual work in the pilot:

*REC1*: focus on enabling virtual machines that encapsulate a collection of useful scientific tools.

*REC2*: create ways to either include relevant datasets or their selected parts as part of the virtual machine image, or/and provide a (read-only) computable storage interface to those virtual machines that hold the relevant ELIXIR datasets.

*REC3*: focus at first on non-sensitive datasets, e.g. plants, fish

Actions

* *FP* will provide *GD* and *BL* the Request to the ELIXIR-GR and the Letter of Endorsement, to be used as templates.
* *BL* will procure a Letter of Endorsement from the SI ELIXIR Node
* *GD* will procure a Letter of Endorsement from the Italian ELIXIR Node.
* *BL* will communicate with the 2015 ELIXIR all-hands meeting organization committee, in order to include a flash presentation of the EGI-ELIXIR pilot in the agenda (by Tuesday 03/03/2015)
* *FP* will prepare a short presentation for the meeting (description of the pilot, information from the survey)

## Part B: Progress report in Pilot Tasks

### Task 1: Identify existing life science datasets in EGI (GS)

No update

*GS*: Difficult to receive feedback from the NGIs.

*GS*: It might be problematic to identify existing setups in EGI, with the current setup.

### Task 2: Identify reference datasets for replication (FP)

Progress

* Circulate Questionnaire on Reference Datasets (target audience: end users)
URL: https://survey.egi.eu/index.php/652565/lang-en
* Get initial feedback regarding relevant datasets

The results of the survey will be detailed on a separate document, which will be updated periodically to include new responses (DocDB repository)

Actions:

* Construct a second survey for dataset providers, after receiving some feedback from the Dataset Users survey. The second survey can be in the form of a document to directly engage Dataset providers
	1. who is using it,
	2. how is it provided,
	3. how often is it accessed, etc
* Send a second round of emails for additional responses to the first survey
* Include an analysis of the responses for the following events
	1. 2015 ELIXIR all-hands meeting in Hinxton (30-31/03/2015)
	2. EGI Conference in Lisbon (18-22/05/2015)

### Task 3: EGI AppDB extension to a dataset registry (WK)

Progress

* First iteration of the back-end/front-end implementation for the Datasets in AppDB
URL: <https://appdb-dev.marie.hellasgrid.gr/store/dataset>

General Comments

*WK*: Default view of a dataset contains one or more locations. The primary location of the dataset is denoted as “MASTER”, all replicates are denoted as “REPLICA”

*WK*: The classification scheme currently used is adopted from the one used in AppDB (only higher level for the time being)

*WK*: We aimed for compatibility with the ELIXIR Dataset Registry (<https://elixir-registry.cbs.dtu.dk/>). However, the Registry view is mostly ontological, so additional investigation is required.

*MC*: There are two options in showing different versions of datasets in AppDB.

1. Separately (current implementation)
2. Within each entry – multiple versions in each entry

*FP*: I support option #2. The higher level listing should have only the main dataset (latest version) and within each entry to have an extra tab for older versions. Additional point, keep also date of last release along with the actual version.

*AD*: The current implementation though (option #1) is more user friendly. Especially as there are many researchers that only use one particular version of a dataset (for various reasons)

*GD*: However, option #2 is more contained. And there are no restrictions for having multiple versions.

*FP*: Keeping all versions could be a problem on the long term – too many datasets with no guaranteed utilization by the end users.

*LD*: It is not a problem. It is possible to replicate everything – we have the necessary technologies in place.

*GD*: True. The actual replication can be driven by the end users’ requests.

*WK*: We can definitely do that. If the data is there, we can let the user decide in the end which entries / versions to be visible at the end.

*WK*: There has been a design decision that all Dataset Entries will be entered manually by the user. Which means that the data entry mechanism is not yet ready/adapted for communication with other Dataset Registries (such as the ELIXIR registry).

*FP*: Indeed. For the time being we focus on manual entry of the Datasets, by the Data Providers. However, we must keep into consideration that integration / communication with the ELIXIR Dataset Registry may be required in the future.

Actions

* Further refine the front-end, for entries to include references to older version entries
* Include reference/link to the ELIXIR Dataset Registry, when applicable
* Identify potential communication protocols with the ELIXIR Dataset Registry

### Task 4: Tools for data replication (GD)

Progress

* Created a table with all identified technologies for data replication.
URL: <https://docs.google.com/spreadsheets/d/1fftRZxBd1NSeSK8y8YIhts28TrTpTcJIQ52JMjtPPBI/edit#gid=0>

General Discussion

*GD*: There are several use-cases that have identified applicable to datasets (such as Transfer, Access, Sharing, etc).

*GD*: In progress of identifying the respective contacts for each technology.

*LD*: There is also the concept of getting mounted datasets as VMs, accessible either through Cloud or Grid applications

*GD*: There are two issues with that:

1. Size of the VM is largely dependent on the dataset, does not scale well, especially when discussing PB of data
2. There is always the issue of local availability when requested

*LD*: Agreed. The Dataset access mechanisms should be located on the data layer of the infrastructure.

*DS*: Same issue was with the Chipster use case (GAPF VT).With focused on Block Storage and CVFMS

Actions

* continue updating the table, and connect technologies with Reference Datasets (Task 1) and corresponding applications (Task 5).
* identify the relevant people for each technology
	+ we can invite these people invited as a speaker within a workshop to facilitate discussion.

### Task 5: Analysis tools to work with data replicas (AD)

Progress

* Created a table with all tools / applications corresponding to the Reference Datasets identified in Task 1. Available also as a short report in DocDB.
URL: <https://docs.google.com/spreadsheets/d/1gcSNCNBdwIgNN8URrj0b2w4WT18VA04BHOi5csT0HGM/edit?usp=sharing>

General Discussion

*AD*: Identified two general categories of applications related to Reference Datasets

1. Web-based applications: user-friendly but limited by the platform deployed (usually Java applications)
2. Command-line applications: more powerful and can be easily deployed on EGI resources. However usually require some expertise from the end user.

*AD*: Two options to proceed:

1. Identify additional tools, not necessarily applicable to the key Reference Datasets (wide)
2. Identify additional tools specific to the key Reference Datasets (deep)

*FP*: It makes more sense to go for option #2, i.e. more tools that can be employed with the top Reference Datasets. In that case, the replication process will be easily and efficiently tested using the most relevant tools.

Actions

* Continue updating the table, with additional focus on application already listed on AppDB

### Task 6: Integration with ELIXIR Registry (MC)

In progress

*MC*: There is a planned meeting with Jon Ison from the ELIXIR registry (Wed 25/02/2015).

Actions

* identify the list of the metadata that would be useful for the ELIXIR registry in order to secure (in a level) a smooth integration
* circulate the outcome of the discussion in the Pilot mailing list.

## Part C: Discussion and further Actions

*Actions for next period*

* Follow-up on the ELIXIR Nodes’ Endorsement Letter (**General**)
* ??? (**T1**)
* Construct a second survey for dataset providers (**T2**)
* Send a second round of emails for additional responses to the first survey (**T2**)
* Prepare a document with the survey’s results (**T2**)
* Further refine the front-end, for entries to include references to older version entries (**T3**)
* Include reference/link to the ELIXIR Dataset Registry, when applicable (**T3**)
* Identify potential communication protocols with the ELIXIR Dataset Registry (**T3**)
* connect technologies with Reference Datasets (T1) and corresponding applications (T5) (**T4**)
* identify the relevant people for each technology (**T4**)
* update tool table, with additional focus on application already listed on AppDB (**T5**)
* identify the list of the metadata that would be useful for the ELIXIR registry in order to secure (in a level) a smooth integration (**T6**)
* circulate the outcome of the discussion in the Pilot mailing list (**T6**)

*Next teleconference*: Monday, March 23rd, 12:00 CET

# Useful Links

**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>

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

1. After a follow-up call [↑](#footnote-ref-1)