3rd Meeting of the EGI-ELIXIR Pilot

Integrating ELIXIR reference datasets within the European Grid Infrastructure

**Date**: Wed 14 Jan 2015 (14/01/2015) 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 | AUTH | GR |
| Afonso Duarte | ITQB | PT |
| Diego Scardaci | EGI.eu | NL |
| Gergely Sipos | EGI.eu | NL |
| Giacinto Donvito | INFN Bari | IT |
| Jan Jona Javoršek | NGI\_SI | SI |
| Joao Pina | LIP | PT |
| Lukasz Dutka | CYFRONET | PL |
| Marios Chatziangelou | IASA | GR |
| William Karageorgos | 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**

Sara Coelho is preparing an EGI newsletter article for the Pilot (together with FP and Rafael Jimenez). The article will be cross-posted and promoted within the ELIXIR network.

* As soon as the article is online, engage the ELIXIR PT node (AD)

Gergely Sipos has promoted the EGI-ELIXIR Pilot in two official reports:

1. Final Report of the EGI-Inspire project
2. BioMedBridges deliverable

**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.

Actions

* Jan will procure a Letter of Endorsement from the SI ELIXIR Node
* Diego will inquire the Swedish ELIXIR Node (BILS) for such a Letter
* Giancinto will check with the Italian Head of Node for such a Letter
* FP will check with the Greek ELIXIR Node for such a Letter (GR Node status is Observer)
* FP: follow-up with Kay regarding deadlines for Letter procurement. Inform Jan (NGI\_SI) for facilitating Letter from the SI node

## Part B: Progress report in Pilot Tasks

### Task 1: Identify existing life science datasets in EGI

In progress

Gergely will talk to the EUDAT people regarding the LS datasets they support

Task 1 Milestone (*M3*)

* Provide a document describing examples of LS datasets from past experiences in NGIs

### Task 2: Identify reference datasets for replication

Initial questionnaire has been drafted as a document. Version 0.5 is available in DocDB

Two distinct target audiences for the questionnaire:

1. Dataset providers
   1. who is using it,
   2. how is it provided,
   3. how often is it accessed, etc
2. Reference Dataset users (i.e. Life Science researchers)
   1. which datasets do you use,
   2. which location would you prefer,
   3. how often do you use it, etc

Current version of the questionnaire is focused on the 2nd category, i.e. Reference Dataset Users.

General comments:

1. Include a short summary, introducing the EGI-ELIXIR pilot (1-2 sentences, with references to the online resources)
2. Include a small paragraph explaining the aim of the questionnaire, target audience and expected outcomes
3. Amend the title to address “Datasets” rather than “Databases”
4. Clarify that the focus of the pilot is on Flat File replication, and not (e.g.) MySQL databases.
5. Start with the definition of the “Dataset”
6. Address the issue of NoSQL databases (are they being considered for use, etc) → Section C
7. Include a question regarding T5, i.e. the use of applications using this data (standard app or own apps) → Section D

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
* Produce a final version of the current questionnaire and circulate within the Pilot participants for feedback (deadline Fri 23/01/2015)
* Migrate final questionnaire to LimeSurvey and circulate it on potential emailing lists.

Task 2 Milestones (*M3*)

* Circulate Questionnaire
* Get initial feedback regarding relevant datasets
* First version of the second questionnaire

### Task 3: EGI AppDB extension to a dataset registry

In progress

Start of implementation of the first building blocks of the schema in AppDB

Further meta-data can be added on the fly

Integrate the results of the T2 questionnaire into the current schema (emphasis on the second survey)

Task 3 Milestone (*M3*)

* Provide initial visibly functionality on AppDB, with more emphasis on the back-end

### Task 4: Tools for data replication

In progress

Starting an action together with some collaborators from CNR in Italy ELIXIR

Connection with EBI; follow the ELIXIR Task Force. However, that group is not really involved yet in tracking such technologies, still in the process of writing white papers, user requirements and use-cases.

Contact people and highlight what we know on storage technologies, in order to produce a first draft document of relevant ideas and technologies

Will provide a clear draft of technologies that we should evaluate, focusing on interaction with people about these technologies (estimated time of document delivery end of January)

Task 4 Milestone (*M3*)

* produce a table listing relevant people and technologies (solutions from use-cases). As a follow-up point, we can invite these people invited as a speaker within a workshop to facilitate discussion.

### Task 5: Analysis tools to work with data replicas

In progress

Clearer ideas now based on T1 and T2.

AD with JP tried to engage the ELIXIR PT node, however no feedback so far.

Will use feedback from the questionnaire (T2) regarding the datasets.

Task 5 Milestone (*M3*)

* prepare a document with the tools associated with the databases/datasets in the respective Google Sheet file

### Task 6: Integration with ELIXIR Registry

In progress

contact with Jon Ison in order to discuss about meta-data they want to share after integration with AppDB → circulate the outcome of the discussion in the Pilot mailing list.

Task 6 Milestone (*M3*)

* identify the list of the metadata that would be useful for the ELIXIR registry in order to secure (in a level) a smooth integration

## Part C: Discussion and further Actions

*Milestones / Actions for* ***M3***

* Provide a document describing examples of LS datasets from past experiences in NGIs (**T1**)
* Circulate Questionnaire (**T2**)
* Get initial feedback regarding relevant datasets (**T2**)
* First version of the second questionnaire (**T2**)
* Provide initial visibly functionality on AppDB, with more emphasis on the back-end (**T3**)
* produce a table listing relevant people and technologies (solutions from use-cases). As a follow-up point, we can invite these people invited as a speaker within a workshop to facilitate discussion. (**T4**)
* prepare a document with the tools associated with the databases/datasets in the respective Google Sheet file (**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**)

*Next teleconference*: Wednesday, February 25th, 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)