9th Meeting of the VT Project on

Life Science Reference Data Integration

Integrating ELIXIR reference datasets within the European Grid Infrastructure

**Date**: Mon 22 June 2015 (22/06/2015) 12:00 CEST – Webex Teleconference

Contents

[Participants 1](#_Toc422757415)

[Agenda 1](#_Toc422757416)

[Minutes 2](#_Toc422757417)

[Discussion / Progress report 2](#_Toc422757418)

[Planned actions 4](#_Toc422757419)

[Useful Links 5](#_Toc422757420)

# Participants

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** |  | **Affiliation** | **Country** |
| Fotis E. Psomopoulos | FP | AUTH | GR |
| Afonso Duarte | AD | ITQB | PT |
| Kostas Koumantaros | KK | GRNET | GR |
| Marios Chatziangelou | MC | IASA | GR |
| Alex Nakos | AN | IASA | GR |
| William Karageorgos | WK | IASA | GR |
| Diego Scardaci | DS | INFN/EGI.eu | NL |
| Marica Antonacci | MA | INFN Bari | IT |

# Agenda

* Discussion on the 1st use-case for implementation (see also relevant [document](https://documents.egi.eu/secure/RetrieveFile?docid=2361&version=2&filename=RefGenomeIndexReplication_v1.0.docx))
* Next steps towards implementation

# Minutes

## Discussion / Progress report

**Discussion**

***FP***: Quick overview of the document describing the 1st Pilot Case: "*Replication of reference genome index files for NGS data analysis*" (see also relevant document in [DocDB](https://documents.egi.eu/secure/RetrieveFile?docid=2361&version=2&filename=RefGenomeIndexReplication_v1.0.docx)). Main points:

* Two options of reference genome index data; pre-built (Illumina) or manually constructed (Ensembl source files)
* Four major tools for using index data (Bowtie, TopHat, Cufflinks and R/BioConductor)
* Simple option for testing purposes would be to use pre-build files together with the first part of the workflow (TopHat and Cufflinks)

***FP***: Although the technical implementation of the proposed workflow is currently beyond the scope of this VT, it is evident that there is significant overlap between the two use-cases.

***WK***: The current implementation of AppDB regarding datasets doesn't allow for annotation files or index files.

***FP***: Not an issue. Index files are considered as a single dataset, and treated accordingly. Same thing is evident in the Elixir registry, where annotation files (gtf) are available as datasets.

***WK***: Not a problem if we proceed like this, but we should keep it in mind for future reference.

***AD***: Is Illumina providing this files open to all?

***FP***: Yes, the pre-build index files are publicly available.

***AD***: Should one always fetch the files from the Ensembl / UCSC ? In terms of data replication does it makes sense to keep a copy of it ?

***FP***: Both options for obtaining reference index files are equivalent. Most researchers usually go for the pre-built files, but there are several cases where index files are built from scratch (see also Chipster case), especially in cases of non major genomes.

***FP***: In any case, and regarding the first implementation, it makes sense to keep a simple setup, therefore I propose the use of the pre-built index files.

***AD***: Joao Pina from PT has setup a process of using R (and BioConductor?) across Grid resources. We could use this in this case (*Action #1*).

***FP***: What should be the next steps towards a first implementation?

***KK***: We should publish some guidelines what needs to be done on the site side, i.e. what is the list of products /apps/libs we need to be deployed, and then construct some scripts for Grid users to run.

***FP***: In terms of guidelines, we can improve the current document on the Pilot with the necessary information (*Action #2*).

***KK***: ok, but we need however deploy tophat2 and cufflinks on the Grid sites.

***MC***: We should also include the applications that will be used for the datasets within AppDB, in order to test connections between dataset entries and application entries (Action for future).

***MC***: What are the current methods/techniques that will be used to test data replication and dataset entry in AppDB?

***KK***: We can use CVMFS; definitely for the application deployment and probably also for the dataset deployment. In the datasets case however, we need to know and understand the specifics of the datasets.

***FP***: The pre-build index files may take up ~40GB of storage, but I will update the document accordingly (*Action #2*).

***KK***: I'll talk with my admins to get some feedback on this (*Action #3*).

***DS***: Take under consideration that CVMFS has some issues (especially with big numbers of small-size files).

***FP***: This will not be an issue in our case; we have few files of large size.

***FP***: In summary; first goal is to run the pilot using CVMFS both for the application and for the pre-build index files, and test it initially on the NGI-GR nodes. In the future, we should also consider testing on at least an additional site, in order to verify the data replication process.

## Planned actions

*Action 1*: Connect with NGI-PT (Joao Pina) for information on using R/BioConductor across Grid nodes (***AD***)

*Action 2*: Update the Pilot Case document with additional details on tools (list of products /apps/libs dependencies), as well as estimates on the dataset size for pre-built indexes (***FP***).

*Action 3*: Start a investigative process for implementation on the NGI-GR node (***KK***).

*Next teleconference*: Monday, July 20th, 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)