PROJECT WORKPLAN

Table of Contents

Project title

Integrating ELIXIR reference datasets within the European Grid Infrastructure

Proposers

Fotis E. Psomopoulos, Giacinto Donvito

Coordinator

Fotis E. Psomopoulos (AUTH, <fpsom@issel.ee.auth.gr>)

Contributors

The project is contributed by experts from:

- User Community Support Team and Research Champions from EGI
  - Main contact: Gergely.Sipos@egi.eu
- Resource providers from EGI
  - Main contact: Tiziana Ferrari Tiziana.Ferrari@egi.eu
- EGI technology developers and technology providers
  - Main contact: Diego Scardaci Diego.Scardaci@egi.eu
- ELIXIR Storage task force
  - Main contact: Giacinto Donvito giacinto.donvito@ba.infn.it
- ELIXIR Service registry task force
  - Main contact: Jon Ison jison@ebi.ac.uk

Contributing individuals are:

- Marios Chatziangelou (IASA), Developer and team leader for EGI the Applications Database (AppDB)
- William Karageorgos (IASA/GR), Developer of the EGI Applications Database
- Alexander Nakos (IASA/GR), Developer of the EGI Applications Database
- Sotiris Sotiropoulos (IASA/GR), Developer of the EGI Applications Database
- Petros Eskioglou (IASA/GR), Developer of the EGI Applications Database
- Afonso Duarte (ITQB), EGI Research Champion, Life sciences
- Fotis E. Psomopoulos (AUTH), EGI Research Champion, Life sciences
- Giacinto Donvito (INFN Bari), Italian Grid Initiative
- Lukasz Dutka (CYFRONET), PLGrid
- Jona Javorsek (ELIXIR.SI/JSI), ELIXIR.SI / NGI Slovenia
- Gergely Sipos (EGI.eu), leader of User Community Support Team
- Diego Scardaci (EGI.eu), User Community Support team and Tool development manager
- Rafael C Jimenez (ELIXIR), Technical coordinator
- Nuno Ferreira (EGI.eu), User Community Support team
- <ADDITIONAL NAMES TO BE ADDED>

Motivation

There has been significant work done in the EGI in the past to help the deployment and discovery of services, where “services” can be either computationally oriented (such as batch queues) or application oriented (such as web-services, ready-to-use applications embedded in portal gateways or encapsulated in Virtual Machine Images). However in bioinformatics many services used for analysis purposes rely on public reference datasets. Reference dataset are getting big and users struggle to discover, download and compute with them. There is an increasing demand to compute the data where the reference datasets are located. EGI members already host some biological reference datasets across the infrastructure, however currently EGI neither provides discovery capabilities for available datasets, nor provides guidelines for those who wish to use these datasets or would like to replicate additional datasets onto EGI sites. The project will facilitate the discovery of existing reference datasets in EGI and will develop and deploy services that allows the replication of life science reference datasets by data providers, resource providers and researchers, and the use of these datasets by life science researchers in analysis applications.

There are several life science domains where computing with reference datasets is necessary. For example, the areas of Epigenetics and Metagenomics are prominent among these. In both of these areas major bottlenecks in the analysis workflow are the sequence comparisons against reference datasets, computing with multiple reference datasets, and performing various data enhancement procedures (i.e. the interconnection of the data involved within a specific study to external data). Given that a crucial point in such studies is the massive amount of data involved as input (query data) [1], the researchers usually opt for a local replication of the reference dataset in order to best allocate the available computational resources, often within the respective institution. Consequently, researchers usually go for more targeted, focused and therefore limited studies, missing out in the process significant data patterns and motifs that could emerge from the inclusion of "Big Data" techniques and methods [2].

Through the provision of a data registry within EGI, such studies could potentially scale up to include most, if not all, relevant reference datasets. Although the amount of data currently available as reference dataset is vast, the existing methods and widely used techniques can only hint at the knowledge that can be potentially extracted and consequently applied for addressing a plethora of key issues, ranging from personalized healthcare and drug design to sustainable agriculture and nutrition. Appendix A describes further life science application examples that would benefit from the results of this project.

Tasks:

The project between EGI and ELIXIR consists of the following tasks:

1. **Identify existing life science datasets in EGI:** Identify existing biological reference dataset replicas within the EGI infrastructure together with their key characteristics that make them usable for analysis (such as dataset version, source, access mode, related analysis tools, size, update frequency, tools used for replication, etc.). The task will survey resource providers and life science users of EGI and will look for datasets in the EGI information system and/or other EGI registries. The expected output of the task is an informative table about the datasets that are available on EGI and their key characteristics for users and resource providers → Milestone 1
    - Leader: Gergely Sipos (EGI.eu)
    - Contributors: Fotis E. Psomopoulos (AUTH), Afonso Duarte (ITQB)
    - Estimated length: 2 months
2. **Identify reference datasets for replication:** Identify key biological reference datasets from life sciences that would benefit from replication to EGI sites for example to increase their availability or scalability of access. The task will identify, engage with and survey life science data providers and data users including developers of the ELIXIR tools registry[1]. The expected output of this task is an informative table about life science reference datasets that should be made available on EGI, together with their key characteristics for resource providers and users to replicate them and to use them (ie. metadata describing for example the size, update frequency, preferred access mode, related tools, etc.) → Milestone 2
    - Leader: Fotis E. Psomopoulos(AUTH)
    - Contributors: Giacinto Donvito (INFN), Nuno Ferreira (EGI.eu)
    - Estimated length: 3 months

[1] With special attention to the domains of Epigenetics, Metagenomics, Crop Genomics because these are mentioned in the motivations section.

3. **EGI AppDB extension to a dataset registry:** Extend the EGI Applications Database[2] (AppDB) with new capabilities to expose information about biological reference datasets and their replicas across EGI. Key characteristics of these datasets should be made available by AppDB in the form of metadata for life science users. The initial dataset metadata schema should consist of basic attributes such as name, locations, size, and type; when input from tasks 1 & 2 becomes available, the schema should be revisited in order to identify any additional characteristics that may need to be included. A new access group should also be created, in order to allow particular individuals to input the actual initial metadata, once tasks 1 & 2 are complete. → Deliverable 1
    - Leader and partners: William Karageorgos (IASA), Marios Chatziangelou (IASA)
    - Estimated length: 7 months
4. **Tools for data replication:** Identify and propose suitable software tools, software configurations, operational practices and documentations to those who want to replicate key biological reference datasets to the EGI infrastructure. The tools can be relevant for resource providers to replicate complete datasets for groups of users, and can be relevant for life science researchers to replicate parts of reference datasets for custom analysis. The task will also setup a distributed testbed where the proposed tools and configurations can be tested and validated with real reference datasets and applications by life science communities. The expected outputs of the task are
    i. Recommended services to replicate reference biological datasets to EGI (software, software configurations, operational practices, documentation) → Deliverable 2.
    ii. A distributed testbed where the recommended service portfolio for replication is deployed and where reference life science datasets are replicated → Deliverable 3
    iii. An evaluation of the recommended services on the testbed by resource providers and by life science users. (e.g. online survey or face to face workshop) → Milestone 3
    - Leader: Giacinto Donvito (INFN/Bari)
    - Partners: Fotis E. Psomopoulos (AUT), Afonso Duarte (ITQB), Jona Javorsek (JSI), Marios Chatziangelou (IASA/GR), Lukasz Dutka (CYFRONET)
    - Estimated length: 6 months
5. **Analysis tools to work with data replicas:** Identify and provide guidance for the use of key life science software applications and tools that can be used to work with reference datasets on EGI. These tools can be used by life science researchers to define and execute custom analysis that work on reference datasets hosted on EGI. The task will review the identified tools on the distributed testbed of Task 4, and will provide information for the users about these tools at a central location, ideally as software profiles in EGI AppDB. → Deliverable 4

[2] https://appdb.egi.eu

- Leader: Afonso Duarte (ITQB)
- Partners: E. Psomopoulos (AUTH), Nuno Ferreira (EGI.eu), ELIXIR (delegates to be involved)
- Estimated length: 3 months

6. **Integration with ELIXIR Registry:** Collaboration work between the developers of the EGI AppDB and the ELIXIR service registry to federate information about 'biological reference datasets' from AppDB to the ELIXIR registry. The task will make content from the EGI AppDB visible for the broader life sciences community. Output of this task is technical integration between the ELIXIR Registry and the EGI AppDB, so content about reference datasets hosted on EGI can be federated from the EGI AppDB into the ELIXIR registry. → Deliverable 5.
   - Leader: Marios Chatziangelou (IASA)
   - Partner: Jon Ison (EBI)
   - Estimated length: 2 months

Milestones/Deliverables

1. **M1:** Task 1 will provide an informative table about the datasets that are available on EGI together with their key characteristics for users and resource providers (ie. metadata describing the datasets such as size, update frequency, preferred access mode, etc.). This milestone will be used by Task 3 and Task 6 to implement metadata structures in AppDB and the ELIXIR registry to provide useful information about datasets. The milestone will be used also by Task 5 to identify analysis tools that can work with the existing reference dataset replicas.
2. **M2:** Task 2 will provide an informative table about life science reference datasets that should be made available on EGI, together with their key characteristics for resource providers and users to replicate them and to use them (ie. metadata describing for example the size, update frequency, preferred access mode, related tools, etc.). This milestone will be used by Task 3 and Task 6 as to implement the data structure that should be used by AppDB and the ELIXIR registry to provide information about datasets, and to populate these registries with content.
3. **D1:** Task 3 will deliver an extend version of the EGI Applications Database to expose information about biological reference datasets and their replicas across EGI.
4. **D2:** Task 4 will deliver recommended services to those who want to replicate key biological reference datasets to the EGI infrastructure. The tools can be relevant for resource providers to replicate complete datasets for groups of users, and can be relevant for life science researchers to replicate parts of reference datasets for custom analysis.The services are expected be software, software configurations, operational practices, documentation.
5. **D3:** Task 4 will deliver a distributed testbed where the recommended services (D2) are deployed and where reference life science datasets are replicated.
6. **M3:** Task 4 will provide an evaluation of the recommended services for dataset replication (D2). The evaluation will be performed by resource providers and life

science users on the distributed testbed (D3) in the most suitable way, e.g. online survey, face-to-face workshop.

7. **D4:** Task 5 will provide information about key life science software applications and tools that can be used by life science researchers to define and execute custom analysis that work on reference datasets hosted on EGI. The information will be published at some central location, ideally as software profiles in EGI AppDB.
8. **D5:** Task 6 will deliver technical integration between the ELIXIR Registry and the EGI AppDB, so content about reference datasets hosted on EGI can be federated from the EGI AppDB into the ELIXIR registry.

Benefits to ELIXIR:

The project will benefit ELIXIR by establishing

1. A set of tools and recommendations that would help ELIXIR members and partners
    - Achieve more balanced load on storage resources across their sites
    - Unload user analysis jobs from large centres to partner sites (with data replicas)
    - Perform data processing at national or home institute resources
2. A pilot infrastructure that includes
    - Key datasets for life science analysis workflows
    - Information about applications and tools that researchers can choose from to work with reference datasets
    - A registry that provides information for users about the reference datasets and about the tools that are available to interact with these data
3. A group of experts who can
    - Guide the setup of production infrastructures based on the pilot infrastructure
    - Themselves become providers in production systems.

Timeline

9 months in total, with start as soon as possible.

Resources

EGI and ELIXIR will share and contribute equality to cover the cost of this pilot. Contributions will initially be covered from already running projects (such as EGI-InSPIRE), but opportunities for additional funding will be explored during the work. The partners may organise a joint workshop during the project to help the project achieve certain goals.

Appendix A: Example applications that would benefit from the project

Considering a real-world example for the use of the proposed data registry, the case of crop genomics may be also a relevant one, especially when one considers the problem on the pan-genome level which has been delivering interesting result in other studies [3]. The sequencing of large and complex genomes of crop species, facilitated by new sequencing technologies and bioinformatic approaches, has provided new opportunities for crop improvement [4]. There are several tools that address important issues such as plant-specific molecular mechanics and genome engineering. However, with the advance of new sequencing technologies, many questions and challenges arise that address new and exciting areas such as the crop microbiome [5], plant epigenetics [6] etc. A common layer in all those questions is both the management of large sets of data (requiring the use of compute infrastructures) and the combination of multiple reference datasets for pattern extraction (such as UniProt, Reactome and Ensembl among others). Therefore, an integration of these reference sets into the existing compute infrastructures may well prove to be a tipping point in the definition of the next generation of crop genomics research and Life Sciences research in general.

References

1. John D McPherson, "*A defining decade in DNA sequencing*", Nature Methods 11, 1003–1005 (2014) doi:10.1038/nmeth.3106
2. Francesca Finotello and Barbara Di Camillo, "*Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis*", Briefings in Functional Genomics (2014) doi: 10.1093/bfgp/elu035
3. Fotis E. Psomopoulos, Victoria I. Siarkou, Nikolas Papanikolaou, Ioannis Iliopoulos, Athanasios S. Tsaftaris, Vasilis J. Promponas and Christos A. Ouzounis, "*The Chlamydiales Pangenome Revisited: Structural Stability and Functional Coherence*", Genes 2012, 3(2), 291-319; doi:10.3390/genes3020291
4. Michael W Bevan and Cristobal Uauy, "*Genomics reveals new landscapes for crop improvement*", Genome Biology 2013, 14:206, doi:10.1186/gb-2013-14-6-206
5. Gabriele Berg, Martin Grube, Michael Schloter and Kornelia Smalla, "*Unraveling the plant microbiome: looking back and future perspectives*", Front. Microbiol., 4 June 2014, doi: 10.3389/fmicb.2014.00148
6. McKeown PC and Spillane C., "*Landscaping plant epigenetics*", Methods Mol Biol. 2014;1112:1-24, doi: 10.1007/978-1-62703-773-0_1.