EGI-Engage
http://www.egi.eu/about/egi-engage

Replicating Reference Data Sets within EGI

Project motivations

- Many services in Biomedical Sciences rely on public reference datasets. These datasets are getting bigger and users struggle to discover, download and compute with them. There is an increasing demand to make data available close to computing facilities for user applications.
- The European Grid Infrastructure (EGI) federates national and regional IT providers into a pan-European infrastructure. EGI members offers compute and storage capacity in cloud and grid access modes for research communities and industry.
- Let's explore possibilities to form a baseline infrastructure from EGI, ELIXIR and other complementary resources to offer reference data sets on national compute facilities!

Goals

1. Identify and replicate significant life science reference data sets into EGI
2. Extend EGI Applications Database and ELIXIR-BioMedBridges Registry to dataset replica catalogues
3. Propose analysis tools and prepare best practices for researchers on how to work with data set replicas

Members

- EGI and ELIXIR members and partners from CH, DK, FI, FR, IT, NL, PL, PT, SI, UK
- The project received endorsement letters from ELIXIR Greece and ELIXIR Slovenia.

First results

Top #10 identified reference data sets

1. Ensembl
2. RefSeq
3. NR/NT
4. UniProtKB
5. PDB
6. GEO
7. SRA
8. 1000genomes
9. UniRef100
10. KEGG

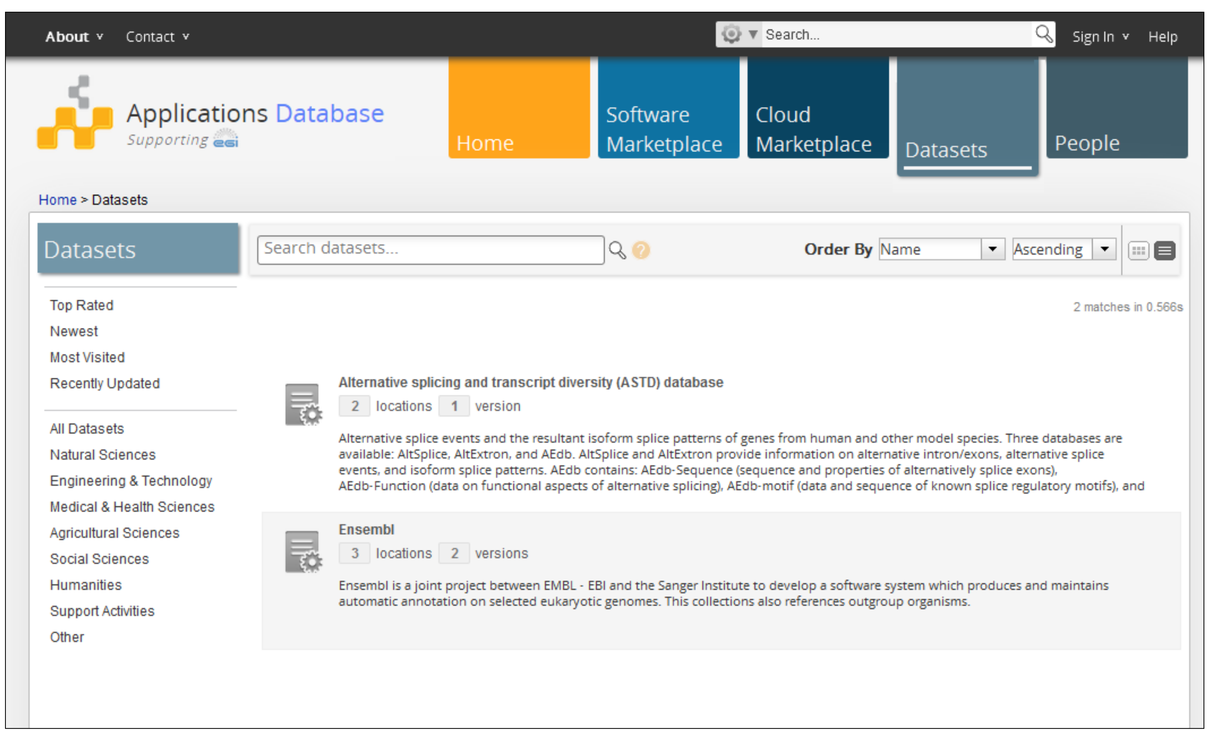
- Sequence DBs **64%***
- Structural / Reaction DBs **16%***
- Specialized DBs **14%***
- Text DBs **6%***

* Percentage of 33 complete survey responses

Platforms and capacity

- The EGI Federated Cloud integrates institutional clouds since May 2014.
- It is agnostic to the underlying cloud management frameworks and provides a gluing layer based on open standards and open source implementations.
- The system integrates OpenStack, OpenNebula and Synnefo clouds from 30 institutes, offering 300 TB storage and 4000 CPU cores.

Data set catalogue



Tools identified for data replication

- CVMFS
- B2SAFE
- Data Avenue
- DynaFeds
- File Transfer Service
- Globus Transfer
- OneData
- OwnCloud

Further information:
http://go.egi.eu/datasetreplication
Suggestions and requests to join:
support@egi.eu