



# 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**



Percentage of 33 complete survey responses

#### Data set catalogue

About v Contact v	Contact ¥			© ▼ Search		
Applications Database	Home	Software Marketplace	Cloud Marketplace	Datasets	People	

#### **Platforms and capacity**

- Federated EGI Cloud integrates • The 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.
- integrates OpenStack, The system OpenNebula and Synnefo clouds from 30 institutes, offering 300 TB storage and 4000 CPU cores.

#### **Tools identified for data replication**

CVMFS B2SAFE





Home > Datasets					
Datasets	Search o	latasets	ି ୧	Order By Name	<ul> <li>▼ Ascending</li> <li>▼ ■ ■</li> </ul>
Top Rated Newest Most Visited Recently Updated All Datasets Natural Sciences Engineering & Technology Medical & Health Sciences		2 locations 1 version Alternative splice events and the available: AltSplice, AltExtron, an events, and isoform splice patter	d AEdb. AltSplice and AltExtron pr rns. AEdb contains: AEdb-Sequenc	of genes from human and other model ovide information on alternative intror e (sequence and properties of alternativ AEdb-motif (data and sequence of know	n/exons, alternative splice vely splice exons),
Agricultural Sciences Social Sciences Humanities Support Activities Other			en EMBL - EBI and the Sanger Insti	tute to develop a software system whic tions also references outgroup organisr	

- 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