EGI-Engage

Data Management Plan

D2.4

Date	31 August 2015
Activity	NA2
Lead Partner	EGI.eu
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2556

Abstract

This document describes the initial data management plan for the research data that will be generated within EGI-Engage. For each dataset, it describes the type of data and their origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	*Partner/Activity*	*Date*
From:	Sergio Andreozzi	NA2/EGI.eu	21/07/2015
Moderated by:	Matthew Dovey (JISC)	PMB	24/08/2015
Reviewed by	Yannick Legre (EGI.eu) Francisco Hernandez (VLIZ)	NA1 SA2	24/08/2015
Approved by:	AMB		31/08/2015

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author/Partner*
v.1	21/07/2015	Table of Content	Sergio Andreozzi, EGI.eu
v.2	11/08/2015	Initial draft	Jesus Marco de Lucas (IFCA) Eric Yen (TWGrid) Ingemar Häggström (EISCAT) Alexandre Bonvin (Univ. Utrecht) Davor Davidović (IRB) Sergio Andreozzi (EGI.eu)
v.3	14/08/2015	Complete draft ready for external review	Kimmo Mattila (CSC) Sergio Andreozzi (EGI.eu)
v.4	28/08/2015	Updated document based on feedback from the external review	Sergio Andreozzi (EGI.eu) Sy Holsinger (EGI.eu)
FINAL	28/08/2015		

TERMINOLOGY

A complete project glossary is provided at the following page: http://www.egi.eu/about/glossary/

Contents

1 Introduction

The EGI-Engage project participates in the pilot action on open access to research data. Research data is defined as information, in particular, facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, and images. The focus of the open research data pilot in Horizon 2020 is on research data that is available in digital form [R1].

The Open Research Data Pilot applies to two types of data: 1) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible; 2) other data (e.g. curated data not directly attributable to a publication, or raw data), including associated metadata.

The obligations arising from the Grant Agreement of the projects are (see article 29.3): Regarding the digital research data generated in the action (‘data’), the beneficiaries must: 1) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following: the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible; other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan'; 2) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex 1, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.

This document describes the initial data management plan[1] for the research data that will be generated within EGI-Engage. For each dataset, it describes the type of data and their origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

As recommended in [R1], this document will be further developed before the mid-term and final project reviews with more detailed information related to the discoverability, accessibility and exploitation of the data.

[1] Data management plan: document detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved.

2 Datasets

Within the EGI-Engage project, research data will be mainly generated or collected by the Work Package 6 through the various Competence Centres (CC). The following sections provide details of each relating to type, origin and scale of data, standards and metadata, data sharing (target groups, impact and approach) and archive and preservation, according to the suggested template (see Annex 1 of the guideline document provided by the EC [R1]). All CCs except for EPOS have provided an initial data management plan. EPOS will provide inputs in a future update of this document.

2.1 ELIXIR Competence Centre

Data management plan contact: kimmo.mattila@csc.fi

No scientific data will be generated within the EGI ELIXIR competence centre, however ELIXIR, as an infrastructure, does manage life science data produced by life scientists. Thus this section will focus on the data managed by ELIXIR instead of the data produced by ELIXIR.

2.1.1 Data description

2.1.1.1 Types of data

The ELIXIR CC will focus on services working with life science data. More specifically, it will provide technical solutions to use cases proposed in the EXCELERATE grant on the management of genomics data: Marine metagenomics, Plant genomics and phenotype and Human sensitive data.

2.1.1.2 Origin of data

The data managed by ELIXIR is produced and submitted by scientists. ELIXIR repositories collect, integrate and provide access to the data.

2.1.1.3 Scale of data

The biggest data collections in life sciences are in the order of petabytes (PB), however, it is likely that the ELIXIR CC will work with smaller data sets. A single whole human genome raw data is roughly 200 GB. However, there are also lots of fairly small files. More information can be found in [R2].

2.1.2 Standards and metadata

Some standards like the standard formats in the marine or the plain domain are still under development. Some of the standards for capturing and exchanging genomic data that might be used in the use cases are described in BioSharing [R3]. Part of the data may be stored to public data repositories (e.g. ENA) that have clearly defines metadata models. More details will be provided in a future update.

2.1.3 Data sharing

2.1.3.1 Target groups

The target audience would be users interested to submit or use Metagenomics, Plant and Human data.

2.1.3.2 Scientific Impact

Sharing data is essential to get data for scientific discoveries such as comparative environmental metagenomic analyses or finding genes related to a disease.

2.1.3.3 Approach to sharing

ELIXIR promotes open data access [R4], but naturally human data might be sensitive therefore requires authorised access. On the web page referenced, there is also a statement from the BioMedBridges project on "commonly agreed principles of data management and sharing".

2.1.4 Archiving and preservation

Services for archiving and preservation within ELIXIR are listed in https://www.elixir-europe.org/services.

2.2 LifeWatch Competence Centre

Data management plan contact: Jesus Marco de Lucas (marco@ifca.unican.es)

2.2.1 Data description

2.2.1.1 Types of data

The LifeWatch competence centre will generate/collect mainly test datasets as part of larger datasets, to analyse the LW-EGI CC framework. For example, one month of data collected at a water reservoir, or six different simulation outcomes related to it.

2.2.1.2 Origin of data

Instruments in the water reservoir.

2.2.1.3 Scale of data

Gigabytes of data in a database that can be exported in the CSV file format.

2.2.2 Standards and metadata

Under investigation.

2.2.3 Data sharing

2.2.3.1 Target groups

The data can be interesting for other research teams that make similar analysis at other water reservoirs.

2.2.3.2 Scientific Impact

The data can potentially underpin scientific publications.

2.2.3.3 Approach to sharing

The embargo period is usually two years as the data is exploited by an SME. The datasets released are usually limited in scope (e.g. 1/10th of total data). The repository is located at the IFCA data centre and freely accessible via web [R5], but registration is needed.

2.2.4 Archiving and preservation

Copies are kept in WORM tapes, and in a separate server (400 km away) of the company. Main repository uses RAID technology and has not lost any data in the last 10 years. The data are automatically synchronised across the servers.

2.3 Disaster Mitigation Competence Centre

Data management plan contact: Eric Yen (Eric.Yen@twgrid.org)

2.3.1 Data description

2.3.1.1 Types of data

There are two main types of data:

- Observation data from tidal gauge, weather stations, rainfall, radar data, satellite data and images, bathymetry, historical records of earthquake and tsunami, etc.
- Waveform at any target site, potential source of a historical tsunami event, changes of rainfall, wind field and path of typhoon or any special weather event, dispersion path of aerosol or volcano ashes, are the primary simulation results.

2.3.1.2 Origin of data

Government agency of weather, earthquake, tsunami, and volcano; or research institutes that own the data needed by the CC.

2.3.1.3 Scale of data

Data scale of the whole collection and generated data would be few TB to 10s of TB. Variation is possible due to the resolution of the generated output.

2.3.2 Standards and metadata

The ISO 19156 standard for Observation and Measurement data model was selected. For weather and climate data, the centre will also comply with the Climate and Forecast convention (CF) (e.g. NetCDF). Both of these specifications are included in the new metadata model called ADAGUC Data format standard.

2.3.3 Data sharing

2.3.3.1 Target groups

The data can potentially underpin scientific publications. Scientists of tsunami, earthquake, volcano, weather, and climate changes; scientists, policy makers of disaster mitigation strategy and studies.

2.3.3.2 Scientific Impact

The data can support new discoveries such as the sources and characteristics of potential tsunami sources or new ways of hazards simulation and analysis. The data can also support new modelling schemes and the change processes of climate and disaster events.

2.3.3.3 Approach to sharing

Almost every government has strict regulation for announcement of weather and natural hazards, so the centre is focusing on research instead of releasing results to the public. Moreover, sharing is still up to the clearance of right for dissemination from the original agency. At least during the project years, the data collected or generated would be shared in a restricted way and for academic purposes only.

2.3.4 Archiving and preservation

The data will be organised and managed in a repository over the distributed infrastructure. The CC plans to have no less than three copies of the data set at different sites. Academia Sinica (Taiwan) is in charge of the long-term data preservation.

2.4 EISCAT_3D Competence Centre

Data management plan contact: Ingemar Häggström (ingemar.haggstrom@eiscat.se)

2.4.1 Data description

2.4.1.1 Types of data

Development of value-added products (e.g. processes, combined data, plots).

2.4.1.2 Origin of data

EISCAT Incoherent Scatter radar low-level data.

2.4.1.3 Scale of data

A few TB/year will be produced within EGI-Engage. EISCAT data are of a larger order of magnitude.

2.4.2 Standards and metadata

A mixture of standards depending on type. For long-term preservation, the format hdf5 will be used.

2.4.3 Data sharing

2.4.3.1 Target groups

Mainly, space and environmental researchers.

2.4.3.2 Scientific Impact

This research data can underpin scientific publications.

2.4.3.3 Approach to sharing

Current value-added products are open to all from day zero, but low-level data is not. Discussions on the new products are still on going.

2.4.4 Archiving and preservation

Data are stored on a few e-Infrastructures, mirrored and synchronised. There are two levels of storage: a large short-term, and a reduced long-term.

2.5 MoBrain Competence Centre

Data management plan contact: Alexandre Bonvin (a.m.j.j.bonvin@uu.nl)

2.5.1 Data description

2.5.1.1 Types of data

There is research data involved in the activity, but this is not produced with EGI-Engage resources, but from other EU projects (e.g. the I3 iNext infrastructure project [R6] for which a data management plan has been drafted). The types of data produced by those other projects are experimental NMR, Xray, SAXS and cryo-EM data.

2.5.1.2 Origin of data

Biological samples (owned by the end users of the facilities).

2.5.2 Standards and metadata

The end results are typically deposited into public databases like the PDB [R7] or EMDB for cryo-EM data.

2.5.3 Data sharing

2.5.3.1 Target groups

The raw data are usually so complex that they are only of use to expert users in structural biology that have been trained in a specific technique. The processed and derived data typically deposited in public databases (see 2.5.2) are of use to researchers in life sciences in general and for biotech and pharmaceutical companies.

2.5.3.2 Scientific Impact

This research data can underpin scientific publications.

2.5.3.3 Approach to sharing

Data are shared via databases (e.g. again PDB, EMDB), with possibly an embargo period until publication. Other datasets (e.g. the results of computations) can be shared via EUDAT or other repositories like SBGRID for structural biology. For such an example see: https://data.sbgrid.org/dataset/131/

2.5.4 Archiving and preservation

From a university perspective, data are to be kept for 10 years. Currently, there is no proper archiving mechanism in place at the particular site (Utrecht University). At the moment, policies and services rely on what is provided by the database service providers where data are deposited.

2.6 DARIAH Competence Centre

Data management plan contact: Davor Davidović (davor.davidovic@irb.hr)

2.6.1 Data description

2.6.1.1 Types of data

During the project, the centre will generate/collect data that come from the research activities in the fields of Arts and Humanities. Common types of research data generated and collected in A&H are books, letters, emails, paintings, photographs, manuscripts, various digital collections, audio/video materials, etc. However, in the research activities related to EGI-Engage, the focus is on digitised data, i.e. the information/data stored in different digital formats, such as plain files (text, photo, audio and video), metadata, collections, and annotations. The collected data are very heterogeneous, both in source (origin) and the format and metadata used for their digital preservation.

2.6.1.2 Origin of data

The digitised data used in these research activities originates from the physical objects/artefacts used in the research activities connected to A&H, for example, books, audio and video materials, paintings, archaeological artefacts, etc. that can be found in museums, libraries, etc. However, the focus is on existing digitised collections of these physical artefacts that are generated, operated and managed by the members of the DARIAH community. Thus, the main sources of data for this related research are those digitised data provided by various DARIAH members. Some of DARIAH members already operate their own digital repositories, which will be used as a data source.

2.6.1.3 Scale of data

It is hard to estimate the scale of the research data because of a large number of different sources and the amount of information that is stored. More detailed information of the size of data generated/collected (in terms of GB/TB) will be available when the survey on e-Infrastructure needs of DARIAH community is finished and inputs analysed (end September). For example, a pilot-project that creates a database of Bavarian dialects in Austria using gLibrary collects data from an existing database of Bavarian dialects that contains about 50,000 headwords and approximately 70,000 records plus 3,000 multimedia files.

2.6.2 Standards and metadata

Currently, the DARIAH community does not promote any specific metadata standard. The adopted metadata formats vary from case to case. Also, there is no recommendation about any long-term preservation format and thus no domain-specific data format is used or recommended. Thus, an individual approach for each use case is required.

2.6.3 Data sharing

2.6.3.1 Target groups

The data collected within the DARIAH Competence Centre will be useful primarily to the members of the DARIAH community. In addition, it is believed that the wider audience having strong interests in exploiting A&H data can benefit in using these data. For example, the data can be used in education (e.g. digital books, newspapers, other educational materials), museums (e.g. presenting their exhibits in digital format or long-term archiving of digital copies of their entire collections), libraries, or archives.

2.6.3.2 Scientific Impact

This research data can underpin scientific publications.

2.6.3.3 Approach to sharing

For now, no further information on how data will be shared and accessed is known. A concrete answer to that question will be possible upon the completion of an e-Infrastructure survey. As far as known, the majority of data are stored and shared via various data repositories that can be widely accessed. The repositories are mostly institutional (i.e. DARIAH member institutions such as computing/storage centres or research organisations).

2.6.4 Archiving and preservation

Since the data are highly diverse and heterogeneous with no recommended standard, it is impossible to answer this question. The implementation of the repositories, safe guarantee, number of copies, etc. is on individual data/repository providers. The plan is to implement several digital repositories for a specific DARIAH use cases (e.g. Bavarian dialects) using gLibrary framework that allows storing the data on different storage elements (local, grid and cloud storage elements). The partners in the Task 6.6 plan to provide a part of their EGI storage resources for the VO that will be established for the research purposes of DARIAH community.

3 References

No	***Description/Link***
R1	Guidelines on Data Management in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
R2	BioMedBridges workshop on e-Infrastructure support for the life sciences – Preparing for the data deluge http://zenodo.org/record/13942#.Vcy8kBNVhHw
R3	BioSharing https://www.biosharing.org/search/?q=genomics&content=standards
R4	https://www.elixir-europe.org/open-access
R5	Repository for LifeWatch – Water Reservoir data: http://doriiie.ifca.es
R6	Infrastructure for NMR, EM and X-ray crystallography for translational research http://inext-eu.org/
R7	Protein Data Bank in Europe www.pdbe.org