Finnish Grid and Cloud Infrastructure (FGCI)

Data management plan (DMP)

Version 1.0
Approved by the board on 31.5.2017

This DMP is a living document that is updated and modified as the implementation progresses.

1. FCGI overview and division of responsibilities

The FGCI is a distributed infrastructure. The FGCI university and research institute partners handle production computations and data production relatively independently. As an infrastructure, FGCI has three primary strategies to promote good data management: training, support, and infrastructure design to make it easier to manage data well than poorly. It is our philosophy that the infrastructure should support this. **FGCI as a consortium not only strongly encourages, but also gives guidance and requirements for publishing datasets and open access publishing.**

The data produced by the FGCI can be divided in two parts:

1. **Data from scientific computations.** These large data sets **are owned and handled by the university partners,** and they are responsible for its storage and open access sharing, according to the guidelines and policies of each university. **This data is quantitative simulation data, and is highly valuable from a scientific point of view, as described in the main FGCI application**.

2. **FGCI internal data** on cluster usage, including technical information and statistics. This data is relatively small in scope, and is handled by CSC – IT Center for Science Ltd. following the processes and requirements set by EGI (European Grid Infrastructure). **This quantitative data is valuable for further development of grid technologies.**

The researchers and data owners using it have ultimate responsibility for the type 1 data management. However, the FGCI has a goal to share technical expertise and good scientific practices. Our size and resources provide a unique opportunity to make good data management as natural as using infrastructure, as opposed to a formality before and after research. A good plan will allow us to reach a significant number of scientific computing users within Finland. **To this end, the FGCI will make information on data management a key part of the FGCI activities.**

To facilitate the data management, the FGCI will require partners to follow open access policies, and will formulate a FGCI Data Management plan for the general principles within the infrastructure, and detailed Data Management guidelines that takes into account the specific policies and environment of each partner university, and gives university- and discipline-specific advice.

2. Management of FGCI data from computations

In this section, we describe our support for managing the data from computations contained in our infrastructure.

2.1. Existing data management policies and activities of partners

The FGCI partners already are individually rapidly developing resources for data management.

Through the collaboration of the FGCI, the lessons from these programs can be spread among institutions similarly, to how FGCI supports the transfer of technical knowledge.

FGCI partners have their own data management support and policies, which will be followed. These are published on the web and listed in Table 1. All take into account data management throughout the data lifecycle, from planning to archival and reuse.

Furthermore, each partner will name a **data management contact person**, who is in charge of enforcing this plan at his or her location. They will provide data-centric user support and training.

2.2. FGCI Data management principles and guidelines

Table 1. Data management policies and guidelines of FGCI user sites

University of Helsinki	http://www.helsinki.fi/kirjasto/en/get-help/management-research-data/research-data-policy/
Aalto University	http://www.aalto.fi/en/research/research_data_management/
University of Jyväskylä	http://openaccess.jyu.fi/en
Tampere University of Technology	http://scienceport.tut.fi/researchdata/datapolicy; http://scienceport.tut.fi/researchdata/opendata
University of Eastern Finland	http://www.uef.fi/web/open-uef/avoin-julkaiseminen
University of Oulu	http://www.oulu.fi/university/node/44598
University of Turku	http://www.utu.fi/en/research/open-science/Pages/home.aspx
Åbo Akademi University	http://www.abo.fi/forskning/oppen_vetenskap
Lappeenranta University of Technology	https://www.lut.fi/documents/10633/36609/LUT-Research-Data-Policy.pdf

The FGCI is an infrastructure, and thus day-to-day data management must ultimately be performed by end users. We have identified some ways by which we can ensure that users properly handle their data. Below, we state our concrete data management recommendations. In order to ensure that users follow these, we **recommend initial training in data management** before users may create accounts. For **projects applying for large storage capacity, we will require an initial data management plan** before storage is allocated. This will ensure that users consider data management as a part of research, instead of as an afterthought.

Primarily, we recommend users follow the **existing partner data management guidelines** with respect to openness and dissemination. When possible, we extend and improve these guidelines with a focus on scientific computation and seek to resolve any potential conflicts between policies.

The consortium **will expect compliance with the data management guidelines, and all partner sites provide support to all users of the infrastructure in this**. We will continually update the guidelines with current best practices and the latest recommended services.

FGCI as an infrastructure covers the "middle" of the data lifecycle, i.e. the actual storage and computation. However, our support services will cover all stages of the data lifecycle. **The FGCI is committed to open science principles and open publishing**. The data guidelines further explain the recommendations of implementing this and how to **leverage existing services**

We will provide a **FGCI landing page** on data management. This will have information on data management strategy specific to high-performance computing and data science work. It contains both new information and links to other nationally and partner-specific information, including local contacts. This information is not just available for centralized use, but for adapting to partner-specific documentation.

The consortium will **publish this data management plan as well as practical data guidelines and collate advice on data management practises** on its wiki pages:

https://confluence.csc.fi/display/FGCIOD/Instructions+for+FGCI+grid+users

These wiki-pages also provides a **joint platform for sharing best practices** by the data management contact persons and site administrators to support data management on their local infrastructure.

The FGCI recommends following the open science principles and open access publishing. In implementing this we recommend using suitable existing services such as DMPTuuli for project data management planning, the upcoming national digital preservation service portfolio for research, as well as international services such as Zenodo and EUDAT (see Table 2).

All relevant **publications** must be reported according to each organization's guidelines, ensuring they are sent to the national VIRTA publication reporting system. In general, this is done through the university reporting systems, which are also used for our internal reporting and **crediting to infrastructure**. All publications produced using FGCI must include a **reference to the infrastructure**. The reference to be used is the persistent identifier given to the infrastructure (urn:nbn:fi:research-infras-2016072533) through the Research Infrastructures service.

Research **data that is prepared for sharing** has to be stored either in the IDA service, or in a similar organizational/national/international archive. When choosing a storage and sharing service, the user must consider legal and ethical issues and ensure that the stability and availability of the chosen service are suitable for long-term storage. The service must also give a persistent identifier for the datasets so that there is a way to refer to the data. We recommend that all datasets are stored in a service from which they can be shared and that they are licensed so that others can use them (e.g. Creative Commons licenses).

Table 2. Recommended and supported data management solutions

Open science and research services, including IDA and Etsin	http://openscience.fi/services
EUDAT	https://eudat.eu/
Zenodo	https://zenodo.org/
DMPTuuli	https://www.dmptuuli.fi/

The choice of licenses has to be done taking into account legal and ethical aspects of the data. Software and databases have their own license recommendations. For open datasets we recommend Creative Commons BY 4.0 and for open metadata CC0. The former requires attribution to the original creator and the latter waives all rights ensuring maximum visibility for metadata.

All public datasets must be described in Etsin. If a dataset is described in another service, there must be a reference to this description in Etsin (for instance with persistent identifier). The description of a dataset has to include administrative, technical and descriptive metadata according to current standards. The goal is to ensure good findability of the data and adequate levels of information for others to easily evaluate the possibility for utilization of the data. On top of the metadata description, a link to the data and possible license information has to be included. All datasets produced using FGCI must include a reference to the infrastructure just like the publications.

2.3. FGCI-provided data management tools

FGCI provides a variety of tools and integrations into the data life cycle, although our core service is computation and storage during computation only.

Each partner provides core day to day data storage for research activities. In general, this storage space is large and fast, but not backed up. A smaller home directory space is provided for backed up codes and critical configuration. For backed up data storage, other partner university storage locations must be used. Each local cluster has integration into the local resources. Data can be stored in both per-user folders, or group folders for collaboration. In all cases, data is protected by filesystem permissions.

On all HPC environments provided by FGCI, we have installed the iRODS commands, **which allow direct access to the IDA storage service from clusters**. This allows direct staging to and from long-term storage. There is also a grid storage node from FGI phase of the infrastructure available based on dCache technology.

The cPouta cloud service provides several tiers of data storage: default, non-backed up disks, high-performance IO storage, and normal backed up storage.

2.4. Implementing data management training and instruction

By far, the hardest data management problems are on the human side, and our consortium's training processes assist here. Support is provided both nationally and locally, with the consortium serving as a conduit for best practices to be shared. Our aim is that local support staff, who are already involved in FGCI support, are able to "know their customers" and provide the most useful support. Our goal is to nationalize local best practices, and allow them to be re-adopted and specialized again by local groups.

FGCI, as a part its' data management activities, will **organize events for data management planning and a roadshow to all the partners.** The timetable for the FGCI data management activities is given in Table 3. The idea is to **provide targeted training and offer support for support staff**, so that they can better support researchers from the bottom up. The consortium will make use of the Open science training materials. CSC's existing training framework including trainings on data science and data management will be offered to users.

Table 3. FGCI data management activity timetable.

Start	End	Activity	Status
Jun 2016	Oct 2016	Formulation and acceptance of operational level agreement	Completed
Nov 2016	May 2017	Formulation of data management plan and detailed guidelines	First version completed
Aug 2017	Aug 2017	Internal data management workshop	In preparation
Aug 2017	Oct 2017	**FGCI roadshow:** information on grid usage and data management to all partners	In preparation
Jun 2017	Dec 2023	Followup of data management, including updating the data management plan and its implementation, e.g. taking into use new data management technologies	In planning

3. Management of FGCI internal data

In this section, we outline the data produced by FGCI in the daily operation of its cluster.

3.1. Types of data

Primarily, FGCI data is of cluster status, usage, and job statistics. This is primarily useful for reporting and development of FGCI resources. Data is collected automatically on the FGCI clusters as a normal part of the open source platforms we run. For example, the batch system contains records of all jobs run. This provides data automatically in a standard structured and interoperable form. All cluster software and automated configuration is considered data.

3.2. Documentation and quality

Since the FGCI setup is automatic, the software stack collecting the data is known and reproducible. Because all data comes from standard open source systems, documentation and structuring is automatic. We will prefer the standard forms from these systems when releasing the data, and defer most documentation to the authoritative upstream sources by linking. Data quality matches that of the FGCI infrastructure: the data documents the actual performance of the system.

3.3. Storage and backup

FGCI operational data is backed up our partner's backup systems as a part of normal operations of systems of this scale. The total size of data is small relative to the capacity of FGCI and partner systems.

3.4. Ethics and legal compliance

The relevant cluster operational data is non-personal and the FGCI can release it independently.

Usage data may be released only in a sufficiently anonymous form. Partners, in conjunction with guidelines produced by the FGCI, will conduct anonymization.

3.5. Data sharing and long-term preservation

The cluster data is reported on a yearly basis as part of annual reporting. Summaries are also included in FGCI and partner reports.

Software and other code is made available under the MIT license on Github from the CSC organization account.

4. References

- Creative Commons tool for choosing licenses: https://creativecommons.org/choose
- Persistent identifier for FGCI-infrastructure is urn:nbn:fi:research-infras-2016072533, which is resolved in National Library's URN.FI -service: http://urn.fi/urn:nbn:fi:research-infras-2016072533
- Open science training materials http://avointiede.fi/koulutus