REPORT ON OpenDreamKit DELIVERABLE D1.6

Data Management Plan V2

BENOÎT PILORGET

Due on	31/08/2018 (M36)
Delivered on	30/08/2018
Lead	Université Paris-Sud (UPSud)
Progress on and finalization of this deliverable has been tracked publicly at: `https://github.com/OpenDreamKit/OpenDreamKit/issues/22`	

ABSTRACT. This document aims at gathering general information concerning data produced within the OpenDreamKit project, and steps taken by the project to foster its dissemination and long term preservation. Such data includes but is not limited too:

(1) Code, including its history and possibly social metadata (issue tracking, ...)
(2) Publications
(3) Databases
(4) Reports, web site content, ...

Each site has filled this document with the information they could provide as of August 31st of 2018. There is little Intellectual Property discussion since, following the project's Open Source, Open Data, Open Publication, Open Management spirit, most of the data is available under some appropriate Free (as in Free Speech) license.

CONTENTS

1. PROJECT INFORMATION

- **Action to be implemented:**
  Open Digital Research Environment Toolkit for the Advancement of Mathematics - OpenDreamKit, project number 676541
- **Funding Programme:**
  OpenDreamKit is a Horizon 2020 European Research Infrastructure project
- **Objective:**
  It will provide substantial funding to the open source computational mathematics ecosystem, and in particular popular tools such as LINBOX, MPIR, SageMath, GAP, PARI/GP, LMFDB, SINGULAR, MathHub, and the IPython/Jupyter interactive computing environment. From this ecosystem, OpenDreamKit will deliver a flexible toolkit enabling research groups to set up Virtual Research Environments, customised to meet the varied needs of research projects in pure mathematics and applications, and supporting the full research life-cycle from exploration, through proof and publication, to archival and sharing of data and code.
- **Partners:**
  1- Université Paris-Sud (UPSud)
  2- Centre National de la Recherche Scientifique (CNRS)
  3- Jacobs University Bremen GGMBH (JacobsUni - Terminated)
  4- Université Grenoble-Alpes (UGA)
  5- Technische Universitaet Kaiserslautern (UNIKL)
  6- The Chancellor, Masters and Scholars of the University of Oxford (UOXF)
  7- Uniwersytet Slaski (USlaski)
  8- The University of Sheffield (USFD - Terminated)
  9- University of Southampton (Southampton - Terminated)
  10- The University Court of the University of St Andrews (USTAN)
  11- Université de Versailles-Saint-Quentin-en-Yvelines (UVSQ)
  12- The University of Warwick (UWarwick)
  13- Universitaet Zuerich (UZH - Terminated)
  14- Logilab (Logilab)
  15- Simula Research Laboratory AS (Simula)
  16- Universiteit Ghent (UGent)
  17- European XFEL (XFEL)
  18- Universität Erlangen-Nürnberg (FAU)
  19- University of Leeds (ULeeds)

- **Timeframe:**
  The project lasts 48 months, from the 1st of September 2015 until the 31st of August 2019.

2. Data responsibility

- Every partner will be technically and legally responsible for keeping, disseminating and preserving data created within their labs, even in cases where works are led by researchers coming from various partners.
- Results are owned by the beneficiary that generates them. Two or more beneficiaries own results jointly generated under the terms stated in article 26.2 of the Grant agreement.
- The Research Executive Agency (referred as 'the Agency') may -with the consent of the beneficiary concerned- assume ownership of results to protect them up to four years after the 31st August 2019 under the terms stated in article 26.4 of the Grant agreement.
- Each beneficiary must examine the possibility of protecting its results and must adequately protect them under the terms stated in article 27 of the Grant agreement.
- Each beneficiary must up to four years after the 31st of August 2018 take measures aiming to ensure exploitation of its results under the terms stated in article 28 of the Grant agreement.
- Unless it goes against their legitimate interests, each beneficiary must take measures as soon as possible to disseminate its results under the terms stated in article 29 of the Grant agreement.

Intellectual property bindings are to be found in the OpenDreamKit Consortium Agreement. However, almost all the data creation process is directly linked to the addition of codebase to opensource softwares. All the data creation is given open access by nature. Data for opensource softwares is furthermore preserved anc accessible on the softwares' platforms and/or repositories. At the project level, the Coordinator will be a technical help to partners putting their data and publications on open access platforms if it is necessary.

3. Necessary resources for implementation

3.1. Publications : Green and Gold access (financial resources)

- Green access: no financial resources are required if partners choose this option for their publications
- Gold access: partners who have chosen this option are free to acquire the right from publishers to edit their publications in open access. UPSud, CNRS, JacobsUni, UJF, USFD, Southampton, USTAN, UZH, and Simula have planned gold open access publication charges in their estimated budget (CF Grant Agreement).

3.2. Open access for data and publications (human resources)

- Partners will use their own process with their administration to publish in open access the decided document
- The Coordinator will support partners that are standing without a clear process of their own
- The Coordinator will supervise ,and intervene if help needed, publication at the consortium of data and publications on European platforms.

3.3. Data without open access

- No specified resources were planned at this stage.
- All data created and managed by OpenDreamKit participants are open access.

4. DATASETS

All the following datasets are public and accessible via open access platforms and/or softwares.

4.1. Website

Data storage and security: The website is located on Github: https://github.com/OpenDreamKit/OpenDr

Dissemination: This dataset is the main dissemination tool of OpenDreamKit, and therefore disseminates itself

Preservation: By using the distributed system git to manage most of our data, we assure a local copy of the data within each participant machine. We rely on the Github external platform for public access. If it should happen that this platform is not available any more, the data can easily be moved away to another platform.

- *Name of data* The OpenDreamKit website
- *Nature of data* Text and metadata concerning OpenDreamKit participants and activities
- *Reuse of existing data* Data such as status reports are accessible to the all opensource software communities so they can follow the evolution of the project.
- *Mean of production* Written by OpenDreamKit participants
- *Data standard* Source code is written in Markdown language and converted into html.
- *Link* opendreamkit.org; https://github.com/OpenDreamKit/OpenDreamKit.github.io

4.2. Additions to SAGE codebase

Data storage and security: All addition to the SAGE codebase will be stored within the distributed SAGE repository on the trac server trac.sagemath.org. For smaller datasets, we might use other distributed Git repositories and store a central clone on platforms such as github. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: The SAGE codebase is publicly accessible through the trac server trac.sagemath.org and distributed within the SAGE software. For other data, we have an open access and open source policy and will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of our data, we assure a local copy of the data within each participant machine. We rely on external platforms (Trac and Github) for public access. If it should happen that those platforms are not available any more, the data can easily be moved away to another platform.

Partners involved: UPSud, CNRS, UGA, OXF, UVSQ

(1) Software code

- *Licence:* GPL
- *Nature of data:* Code
- *Reuse of existing data:* The data is added to the already large existing SAGE codebase.
- *Mean of production:* Code implementation by OpenDreamKit participants.
- *Data standard:* The code is mostly written in Python and Cython, also using the Rest syntax for documentation and SAGE coding conventions.
- *Usage for further experiments:* The code is merged in the software and can be distributed and reused through the Software. Through the git history, one can trace back older versions of the code and re-enable a former state of the software.
- *Link:* trac.sagemath.org

(2) Database of strongly regular graphs
- *Licence:* GPL
- *Nature of data:* A mix of data generators implemented in Python and data stored as json objects. A description can be found in http://arxiv.org/abs/1601.00181.
- *Reuse of existing data:* A part of this data was created, independently of OpenDreamKit, by a number of SAGE contributors. D. Pasechnik acknowledges OpenDreamKit support for his work on this topic.
- *Means of production:* Computers, mostly using SAGE and GAP, and some purpose-written Python code.
- *Usage for further experiments:* All the data in this item is available from SAGE.
- *Link:* `http://doc.sagemath.org/html/en/reference/graphs/sage/graphs/strongly_regular_db.html`

(3) Database of special Hadamard matrices
- *Licence:* GPL
- *Nature of data:* A mix of data generators implemented in Python and hard-coded as text data.
- *Reuse of existing data:* A part of this data was created, independently of OpenDreamKit, by a number of SAGE contributors. D. Pasechnik acknowledges OpenDreamKit support for his work on this topic.
- Means of production: Computers, mostly using SAGE and GAP, and some purpose-written Python code.
- *Usage for further experiments:* All the data in this item is available from SAGE.
- *Link:* `http://doc.sagemath.org/html/en/reference/combinat/sage/combinat/matrices/hadamard_matrix.html`

4.3. Additions to GAP codebase

Data storage and security: All addition to the the GAP codebase will be stored within the distributed GAP repository on github or a related repository. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: The GAP codebase is publicly accessible through github and mostly distributed as part of GAP. For other data, we have an open access and open source policy and will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of our data, we assure a local copy of the data within each participant machine. We rely on external platform for public access. If it should happen that those platforms are not available any more, the data can easily be moved away to another platform. Datasets which are no longer being actively worked on will also be archived in the University's research data (USTAN) repository and on zenodo.

Partners involved: USTAN, OXF

- Name of data: Additions to the GAP codebase
- *Nature of data:* Software code
- *Licence:* GPL
- *Reuse of existing data:* The data is added to the already large existing GAP codebase.
- *Means of production:* Code implementation by USTAN and other participants, building on existing elements of GAP, which have been implemented by many people.
- *Data standard:* Follows the conventions of GAP.

- *Usage for further experiments:* The code is merged in the software and can be distributed and reused through the Software. Through the git history,ne can trace back older versions of the code and re-enable a former state of the software.
- *Link:* http://www.gap-system.org

4.4. Additions to PARI/GP

Data storage and security: All additions to the PARI/GP software are stored on their server which is hosted at the University of Bordeaux, France.

Dissemination: The PARI/GP code, documentation and various binaries are publicly available on the server pari.math.u-bordeaux.fr. These are also packaged for major linux distributions.

Preservation and future access: The PARI/GP software has existed since 1979 and its website since 2003. It gets supports from various French and European institutions and has been hosted at the University of Bordeaux since its creation.

Partners involved: CNRS

4.5. Additions to MathHub portal

Data storage and security: MathHub data is stored, versioned, and protected by the state-of-the art GIT system.

Dissemination: All data is to be ingested into the MMT system developed by Jacobs University. It is the main topic of WP6 to design a sustainable, flexible and distributed system that will allow for such efforts.

Preservation and future access: All data will be hosted publicly on the MathHub portal (`http://mathhub.info`), a dedicated information portal for active documents and data (flexiformal knowledge with integrated semantic services). We can also expect that all the data produced will be protected under GIT as well, and hosted on GitHub or similar services.

Partners involved: JacobsUni, UZH, FAU

- Name of data:
- *Nature of data:*
- *Licence:* Original data will be licensed under an open knowledge license (see `http://opendefinition.org`), transformed data will be licensed as open as the original license allows it.
- *Reuse of existing data and Means of production:* Most data will be generated by transforming and semantic preloading of existing data sources (the mathematical data bases from WP6.)
- *Data standard:* The data created will be in the form of OMDoc/MMT flexiformalizations (representations of mathematical knowledge and data at flexible levels of formality). Since this data will aim to provide specifications for software, flexiformalisations for mathematical knowledge, and specifications for the implementation of mathematical knowledge into mathematical software, this process might require many different formats.
- *Usage for further experiments:* In order to enable computer algebra software systems such as SAGE and GAP to benefit from this system, some of the data might be packaged as part of those distributions (through snapshotting of the federation of Git repositories hosting data).
- *Link:* http://www.gap-system.org

4.6. Additions to nbdime

Nbdime is a new subproject of Jupyter for diff and merge of Jupyter notebooks.

Data storage and security: All source code is stored in public repositories on `www.github.com` using the distributed version control system git. This means full copies of all files

including their edit history are located on both external cloud infrastructure with professional backup routines and the personal computers of developers, and in the event of failure of either system restoring is trivial.

Dissemination: Data can be accessed through the public Github repositories, providing open access. All source code is published under the standard open source licence for source code related to the Jupyter project, namely the "Modified BSD License".

Preservation and future access: All data from Simula is published through public repositories under an open source licence, and copyright is assigned to the Jupyter project. This ensures the results can be kept alive and developed further alongside the Jupyter project. The Jupyter project has multiple international partners, both inside and outside Europe, both academic and commercial, ensuring continuation far beyond the end of OpenDreamKit.

Partners involved: Simula

- *Name of data:* nbdime project
- *Nature of data:* Software code
- *Licence:* Modified BSD Licence
- *Reuse of existing data:* The data reuses conventions from the existing Jupyter codebase, and reuses external open source software libraries where applicable.
- *Means of production:* Code implementation by SRL participants.
- *Data standard:* The code is written in Python and Javascript, using Jupyter coding conventions.
- *Usage for further experiments:* When completed, researchers and developers can use this tool to merge Jupyter notebooks when working with git repositories, an important improvement to a reproducible scientific workflow.
- *Link:* https://github.com/martinal/nbdime

4.7. **Additions to the LMFDB database**

The LMFDB database is currently hosted at UWarwick. As well as the LMFDB database itself, the different components of the data are created and stored in a variety of places, and we only list here those for which ODK researchers are responsible.

(1) The LMFDB database

Name of data: The LMFDB database

Licence: Under discussion by the LMFDB developers.

Nature of data: A mongo database. For a detailed description of its contents, see `https://github.com/LMFDB/lmfdb-inventory`. The website `http://www.lmfdb.org/` provides a user interface to the database and documents its contents, including its origin, extent and reliability.

Reuse of existing data: Some of the data in LMFDB existed for many years, for example the Cremona Elliptic Curve Database (see `ecdata` below), while others have been computed specifically for the project. Contributors to LMFDB are listed at `http://www.lmfdb.org/acknowledgment`.

Means of production: Computers, mostly using special purpose custom-written software, which in itself constitutes a significant research output by the contributors.

Data standard: Each section of the LMFDB has its data quality documented and accessible via the website. For example, see `http://www.lmfdb.org/EllipticCurve/Q/Source`.

Usage for further experiments: All the LMFDB data is accessible through its website. Future plans include provision of an API for accessing the data systematically.

Link: `http://www.lmfdb.org/`

Partners involved: UWarwirck

(2) `ecdata`

Name of data: The Cremona Elliptic Curve Database

Licence: Under discussion.

Nature of data: A collection of plain text files containing tables of elliptic curves defined over Q, together with their arithmetic invariants, contained in a `git` repository at `https://github.com/JohnCremona/ecdata`. For a detailed technical description of their content and format, see `https://github.com/JohnCremona/ecdata/blob/master/doc/file-format.txt`. The website `http://johncremona.github.io/ecdata/` provides a simple user interface to the database and documents its contents, including its origin, extent and reliability.

Reuse of existing data: All this data was computed by John Cremona, with additional contributions from Andrew Sutherland and Jeremy Rouse.

Means of production: Computers, mostly using special purpose custom-written software, which in itself constitutes a significant research output by the contributors.

Data standard: Documented at `http://johncremona.github.io/ecdata/`.

Usage for further experiments: All the data in `ecdata` is made available through the following channels: as an optional package in SAGE (with a small subset as standard); as an optional package in PARI/GP; as standard in Magma; and through the LMFDB. All of these, allow all researchers free access to use the data for their own investigations.

Link: `http://johncremona.github.io/ecdata/`

(3) `ecnf-data`

Name of data: Database of Elliptic Curve over number fields

Licence: Under discussion.

Nature of data: A collection of plain text files containing tables of elliptic curves defined over algebraic number fields other than Q, together with their arithmetic invariants, contained in a `git` repository at https://github.com/JohnCremona/ecnf-data. For a detailed technical description of their content and format, see https://github.com/JohnCremona/data/blob/master/ecnf-format.txt.

Reuse of existing data: This data was computed, by John Cremona and several others.

Means of production: Computers, mostly using special purpose custom-written software, which in itself constitutes a significant research output by the contributors.

Data standard: Not yet documented.

Usage for further experiments: All the data in `ecnf-data` is made available through the LMFDB, allowing all researchers free access to use the data for their own investigations.

Link: `http://johncremona.github.io/ecnf-data/`

(4) `bianchi-data`

Name of data: Database of Bianchi modular forms

Licence: Under discussion.

Nature of data: A collection of plain text files containing tables of Bianchi modular newforms of degree 1 over imaginary quadratic fields of class number 1, contained in a `git` repository at `https://github.com/JohnCremona/bianchi-data`.

Reuse of existing data: This data was computed by John Cremona.

Means of production: Computers, using special purpose custom-written software, which in itself constitutes a significant research output by the contributor.

Data standard: Not yet documented.

Usage for further experiments: All the data in ttbianchi-data will be made available through the LMFDB, allowing all researchers free access to use the data for their own investigations.

Link: `http://johncremona.github.io/bianchi-data/`

4.8. **Data of experimental results published in scientific publications**

Data storage and security: All addition to the software codebases will be stored within the distributed repository of each software (the github central clone). For smaller datasets, including results of experiments, and publication sources, other distributed git repositories will be used and a central clone on platforms such as github will be stored. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: For other data,there is an open access and open source policy which will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of the data, a local copy of the data within each participant machine is assured. External platforms (trac and github) for public access are relied on. If it should happen that those platforms are not available any more, the data can easily be moved away to another platform.

Partners involved: UGA

- *Name of data:* Experimental results of parallel computation benchmarks,
- *Nature of data:* any data related to an experiment, including timings, memory usage, input data, scripts used to run the experiments and description of the software stack used for their production
- *Licence:* Creative commons BY-ND for experiments' data, GPL for scripts
- *Reuse of existing data:* The data is open for reuse without modification.
- *Mean of production:* Experiments run by UGA participants.
- *Data standard:* Not yet documented
- *Usage for further experiments:* All information provided should adhere to the standards of reproducible research, in order to allow reproduction of this data.
- *Link:* None yet

4.9. **Additions to LinBox**

Data storage and security: All addition to the software codebases will be stored within the distributed repository of each software (the github central clone). For smaller datasets, including results of experiments, and publication sources, other distributed git repositories will be used and a central clone on platforms such as github will be stored. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: For other data,there is an open access and open source policy which will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of the data, a local copy of the data within each participant machine is assured. External platforms (trac and github) for public access are relied on. If it should happen that those platforms are not available any more, the data can easily be moved away to another platform.

Partners involved: UGA

- *Name of data:* Additions to the codebase of the `LinBox` software
- *Nature of data:* Software code

- *Licence:* LGPL
- *Reuse of existing data:* The data is added to the already large existing codebase.
- *Mean of production:* Code implementation by GA participants.
- *Data standard:* The code is mostly written in C++, also using the Doxygen syntax for documentation.
- *Usage for further experiments:* The code is merged in the software and can be distributed and reused through the Software. Through the git history, one can trace back older versions of the code and re-enable a former state of the software.
- *Link:* github.com/linbox-team

4.10. **Additions to Givaro**

Data storage and security: All addition to the software codebases will be stored within the distributed repository of each software (the github central clone). For smaller datasets, including results of experiments, and publication sources, other distributed git repositories will be used and a central clone on platforms such as github will be stored. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: For other data,there is an open access and open source policy which will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of the data, a local copy of the data within each participant machine is assured. External platforms (trac and github) for public access are relied on. If it should happen that those platforms are not available any more, the data can easily be moved away to another platform.

Partners involved: UGA

- *Name of data:* Additions to the codebase of the `Givaro` software
- *Nature of data:* Software code
- *Licence:* LGPL
- *Reuse of existing data:* The data is added to the already large existing codebase.
- *Mean of production:* Code implementation by UGA participants.
- *Data standard:* The code is mostly written in C++, also using the Doxygen syntax for documentation.
- *Usage for further experiments:* The code is merged in the software and can be distributed and reused through the Software. Through the git history, one can trace back older versions of the code and re-enable a former state of the software.
- *Link:* github.com/linbox-team

4.11. **Additions to FFLAS-FFPACK**

Data storage and security: All addition to the software codebases will be stored within the distributed repository of each software (the github central clone). For smaller datasets, including results of experiments, and publication sources, other distributed git repositories will be used and a central clone on platforms such as github will be stored. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: For other data,there is an open access and open source policy which will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of the data, a local copy of the data within each participant machine is assured. External platforms (trac and github) for public access are relied on. If it should

happen that those platforms are not available any more, the data can easily be moved away to another platform.

Partners involved: UGA

Data storage and security: All addition to the software codebases will be stored within the distributed repository of each software (the github central clone). For smaller datasets, including results of experiments, and publication sources, other distributed git repositories will be used and a central clone on platforms such as github will be stored. All the present data is public, and there is no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

Dissemination: For other data,there is an open access and open source policy which will advertise the data sets accordingly.

Preservation and future access: By using the distributed system git to manage most of the data, a local copy of the data within each participant machine is assured. External platforms (trac and github) for public access are relied on. If it should happen that those platforms are not available any more, the data can easily be moved away to another platform.

Partners involved: UGA

5. OPENACCESS AND OPENDATA POLICY

5.1. **Publications**

Partners will be following the process explained below. The lead partner of the publication takes responsibility for the open access process for peer-reviewed publications.

(1) Publication in the journal of their choice: Partners must be keeping in mind that the Agency asks projects not to accept more than 6 month embargo (to check the publishers' policies, use http://www.sherpa.ac.uk/romeo/

(2) Give open access to all peer-reviewed publications: Maximum of 6 months embargo
- Partners upload publications on http://arxiv.org/, or on another open access platform of their choice
- Warn the Coordinator so that they keep list of OpenDreamKit publications updated on the project website

5.2. **Data related to open access publications**

The lead partner of a publication will send to the Project Manager all the data related to the peer-reviewed publication once it is given full open access. Data related to a publication are all the data needed to reexamine the research leading to the publication. The Project Manager will publish the data linked to publications on http://zenodo.org/. Thanks to the publications' DOIs, the data will be linked to publications.

5.3. **Openaire**

Once publications are published on an open access platform and their data published on Zenodo, the OpenAire website will do the linkage between them. All OpenDreamKit published work will be present on the OpenDreamKit page available on https://www.openaire.eu/. In order for publications and data to appear on this website, one must state when completing forms on open access platform for publications and Zenodo websites that the concerned work is being financed by Horizon 2020 project number 676541.

5.4. Other data in Open Access

As far as it can be foreseen, all data produced and managed within the frame of the OpenDreamKit project will be available in Open Access. The project management itself occurs publicly on the project github repository `https://github.com/OpenDreamKit/OpenDreamKit`. Most of the work is moreover targeted at improving Open Source software.

5.5. Data storage, access, and security

(1) **OpenData access**

Data produced within theOpenDreamKit frame must be accessible up to four years after the beginning of the project. Therefore partners must ensure access to their open access results. The latter will all be available on EU funded platforms (Zenodo and Openaire) built and managed specifically to allow European research to give long-term open access to their results. The OpenDreamKit consortium therefore expresses its faith in the long-term availability of their work, considering the quality of the concerned platforms. Furthermore the data produced within the project will stay accessible on the opensource software platforms and on their repositories (github etc.).

(2) **Data management policy per partner (when relevant)**

- UPSud
  Most of the data created by UPSud is related to the software SAGE and is progressively incorporated into the SAGE codebase. There might also be smaller data sets of tutorials, documentation and teaching content independent of the SAGE codebase which will be stored accordingly to the size and needs.
- USTAN
  Most of the data created by UStan is related to the software GAP and will be incorporated into the GAP codebase. There might also be smaller data sets of tutorials, documentation and teaching content independent of the GAP codebase which will be stored accordingly to the size and needs.
- Simula
  All data sets produced by the work at SRL so far is in the form of source code and associated documentation and example files. The size of the produced data is small, on the order of tens of MB.
- USFD
  The outputs of the University of Sheffield will be mainly in the form of code. Where other data is needed it will be as a prerequisite for running a particular demonstration.
  - **:** *Data storage and security*: The code will be made available by github, or other suitable software version control and management systems, under BSD licenses. The code is public so there are no concerns about unauthorised access. For backups we will be relying on the distributed nature of git storage and back up facilities managed by the repository.
  - **:** *Dissemination*: Code will be available for dissemination by public accessibility and through the Open Data Science website (http://opendsi.cc) which is also github hosted.
  - **:** *Preservation and future access*: The git model ensures we will have local back ups of repositories across multiple machines, but the main data provision moving forward will be github. The University of Sheffield also has deals with figshare for making data available. We will exploit this mechanism of sharing as appropriate.
- Southampton
  There are no significant data sets associated with the work at Southampton. The most important data is resulting code and associated documentation and tutorials.

The details below refer to this data set, and we expect the data set to be fairly small (order of 1 GB).

: *Data storage and security*: The code is stored in a distributed repository (git at the moment), and a central clone of this repository is stored with Github.com in the cloud. We may use multiple repositories, and store a central copy of each on Github.com.

: *Security*: All the code is public, and there no concern about unauthorised access. Through cloud hosting and local clones of repositories, there are backups and redundancy.

: *Dissemination*: Data can be accessed through the public repositories, and the public website (probably this URL: http://joommf.github.io, tbc), providing open access.

: *Preservation and future access*: They rely on provision of the data through github.com but maintain local copies of the repository in case github.com ceases to exist or suffers from catastrophic technology failure. It is likely that other online repository hosting providers would be able to fill the gap (bitbucket.org is an existing alternative). The University of Southampton offers long term storage of small data sets for 10 years – the repositories would fall into this categories. While the data wouldn't be conveniently accessible, this provides an extra layer of backups, from which accessible repositories and websites could be created easily.

- UVSQ
  Most of the data created by UVSQ is related to the software SAGE and will be incorporated into the SAGE codebase. There might also be smaller data sets of tutorials, documentation and teaching content independent of the SAGE codebase which will be stored accordingly to the size and needs.

5.6. **Long term data preservation**

As seen earlier, most of the data (datasets, code, documentation, website content, etc) produced within OpenDreamKit is version controlled and available from public GitHub repositories. It is therefore accessible to anyone on the web and, by the distributed nature of git, backed-up in many places. GitHub's weight in today's IT landscape – well illustrated by its takeover by Microsoft – gives an additional insurance against its disappearance; it has become "too big to fail" any time soon.

Nevertheless a serious back-up solution is necessary to guarantee the long term preservation of our data. An effortless solution was found with the new universal archive for software code called www.softwareheritage.org/.

Software Heritage is originally a French initiative from the INRIA (National Institute for Research in Computing Science) which is now sponsored by companies such as Microsoft, intel, Société Générale, Google, GitHub or DANS. The goal of the archive is to "collect, preserve, and make accessible source code for the benefits of present and future generations." Software is to be taken in a broad sense, and encompasses all kinds information mentioned earlier.

Though the archive is still at its beginning, it already includes all the public repositories from GitHub, which means that all the software components from OpenDreamKit are preserved. Each software component is assigned a unique identifier that is intrinsically bound to it. It does not rely on third parties, so it is truly persistent.

We have checked on a case-by-case basis that the software components not hosted on GitHub are still covered by Software Heritage, e.g. through other channels like Debian Source Package archives.

6. Definitions

- Data: a set of factual information saved on a medium, produced and collected according to various processes in the research process.

Disclaimer: this report, together with its annexes and the reports for the earlier deliverables, is self contained for auditing and reviewing purposes. Hyperlinks to external resources are meant as a convenience for casual readers wishing to follow our progress; such links have been checked for correctness at the time of submission of the deliverable, but there is no guarantee implied that they will remain valid.