Draft Data Management Plan for OpenDreamKit

Benoît Pilorget (Editor)

January 12, 2016

Contents

1 Introduction

[1] EdN:1

2 Datasets

2.1 UPSud

Most of the data created by UPsud 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.

Data storage and security All addition to the 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 he 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 platform (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.

1. Example

   Name of data Photos from scottish landscapes

   Licence ??

   Nature of data Image, photo

[1] EDNOTE: MK@Benoît, please write somem intro here what this document wants to do

Reuse of existing data We used old photos from 50 decades ago as well as paintings

Mean of production Camera

Data standard Photos are in black and white and printed with particular paper and ink

Usage for further experiments Photos will be put on this website so that they can be viewed by the max people. For this they will need to use that software

Link ??

2. Dataset 1

Name of data Addition to the SAGE codebase

Nature of data Software code

Licence GPL

Reuse of existing data The data is added to the already large existing SAGE codebase.

Mean of production Code implementation by UPsud participants.

Data standard The code is mostly written in Python, 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

3. Dataset 2

Name of data The OpenDreamKit website

Nature of data Text and metadata concerning OpenDreamKit participants and activities

Reuse of existing data

Mean of production Written by OpenDreamKit participants

Data standard Source code is written in Markdown language and converted into html.

Usage for further experiments

Link opendreamkit.org, https://github.com/OpenDreamKit/OpenDreamKit.github.io

2.2 CNRS

This subsection will contain all datasets the CNRS is currently able to describe

Data storage and security Quickly explain how data are stored and protected within yout institution

Dissemination How data can be disseminated -> openaccess etc

Preservation and future access How data can be preserved and available in the next years

2.3 Jacobs University

The data created by the Jacobs University team will be in the form of OMDoc/MMT flexiformalizations (representations of mathematical knowledge and data at flexible levels of formality). Most data will be generated by transforming and semantic preloading of existing data sources (the mathematical data bases from WP6.)

All data will be hosted publically on the MathHub portal (`http://mathhub.info`), a decicated information portal for active documents and data (flexiformal knowledge with integrated semantic services). MathHub data is stored, versioned, and protected by the state-of-the art GIT system.

Original data and 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.

2.4 University of 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 Data Storage: 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 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 We 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.