CoeGSS

Centre of excellence

D1.6 – DATA MANAGEMENT PLAN

Grant Agreement	676547
Project Acronym	CoeGSS
Project Title	Centre of Excellence for Global Systems Science
Topic	EINFRA-5-2015
Project website	http://www.coegss-project.eu
Start Date of project	October 1, 2015
Duration	36 months
Deliverable due date	31.03.2016
Actual date of submission	31.03.2016
Dissemination level	Public
Nature	Report
Version	1.0
Work Package	WP 1
Lead beneficiary	USTUTT
Responsible scientist/administrator	Gienger, Michael (USTUTT)
Contributor(s)	Koller, Bastian (USTUTT)
Internal reviewers	Camiciotti, Leonardo (Top-IX), Edwards, Margaret (CoSMo), Pollone, Michela (CSP), Broglio, Luca (CSP)
Keywords	WP 1, Data Management
Total number of pages	25

The CoeGSS ("Centre of Excellence for Global Systems Science") project is funded by the European Union. For more information on the project please see the website http:// http://coegss-project.eu/

The information contained in this document represents the views of the CoeGSS as of the date they are published. The CoeGSS does not guarantee that any information contained herein is error-free, or up to date.

THE CoeGSS MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, BY PUBLISHING THIS DOCUMENT.

Version History

	Name	Partner	Date
From	**Michael Gienger**	**USTUTT-HLRS**	**08.03.2016**
First Version	**Bastian Koller**	**USTUTT-HLRS**	**11.03.2016**
Second Version	**Michael Gienger**	**USTUTT-HLRS**	**21.03.2016**
Final Version	**Michael Gienger**	**USTUTT-HLRS**	**30.03.2016**
Reviewed by	**Leonardo Camiciotti, Margaret Edwards, Michela Pollone, Luca Broglio**	**Top-IX, CoSMo, CSP, CSP**	**28.03.2016**
Approved by	**ECM Board**		**31.03.2016**

Abstract

This deliverable is an initial version of the Data Management Plan (DMP) of the Centre of Excellence for Global Systems Science. As CoeGSS takes part in the Open Research Data Pilot, it volunteered to provide an initial version of the DMP within the first six months of the project's lifetime.

This document reports about the initial findings and can be seen as a baseline, which will be regularly updated during the evolution of the project. Thus, it is a living document, whose revisions will integrate new findings and also potential changes of the Open (Research) Data Pilot.

The document consists of the verbatim information found in the Grant Agreement and the Guideline documents, provided by the European Commission (EC). This information is augmented with the actual data about the CoeGSS pilots and the data sets that they will produce and use.

Table of Contents

1 Introduction....4
2 The guidelines for Data Management....5
3 The CoeGSS Data Management Process....7
4 So far identified data....10
5 Future plans....22
6 Conclusions....23
7 References....24

Table of Contents

Figure 1: High Level Process of the DMP updates....9

1 Introduction

This deliverable provides insights into the initial version of the CoeGSS Data Management Plan. Whilst a simple overview of data sets and their attributes normally would be expected (according to the official EC Data Management Templates), this document will go beyond, as proper Data Management is of utmost importance for the envisaged Centre of Excellence for Global Systems Science. Therefore, this deliverable contains the data sets and their descriptions, but also provides background on the mechanisms and view of CoeGSS on Data Management in general.

In Chapter 2, the official guidelines for Data Management are presented whereas Chapter 3 sketches the Data Management Process as implemented in CoeGSS. To support a clear understanding of the contents, Chapter 3 also gives a background on contractual obligations and decisions on how to manage data (and the different kinds of data sets and their access potentials). Chapter 4 presents then the so far identified data sets, originating from the three project pilots. Within Chapter 5, the future plans for Data Management are outlined, in particular focus has been put on upcoming, currently unknown data sets and finally, Chapter 6 provides the document conclusions.

It is important to notice that this deliverable is just a snapshot at Month 6 of the project and is treated as a living document, being regularly updated according to new findings and refinements of data used and generated by the project.

2 The guidelines for Data Management

2.1 Introduction

With the introduction of the Open Data Pilot and the offering to H2020 project proposals to participate in that pilot and to volunteer to provide Data Management Plans, the European Commission has officially released a set of guidelines for Data Management, which are listed below.

Please note: for the convenience of reading this document as a standalone part, the following two subchapters are 1:1 cited from the "Guidelines to Data Management in Horizon 2020" document[1], as released by the EC.

2.2 The Data Management Plan Template

The Data Management Plan (DMP) template as provided by the European Commission prescribes the following attributes for each of the datasets used or generated within the project:

- ***Data set reference and name:***
  Identifier for the data set to be produced
- ***Data set description:***
  Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.
- ***Standards and metadata:***
  Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.
- ***Data sharing:***
  Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).
- ***Archiving and preservation (including storage and backup):***
  Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

[1] Guidelines on Data Management in Horizon 2020 - https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

2.3 Additional guidance for Data Management Plans

Further data management principles are also provided in the Guidelines. These are of utmost importance for the Centre and its services. This is further clarified in Chapter 3.1.

Scientific research data should be easily:

1. ***Discoverable***
   DMP question: are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?
2. ***Accessible***
   DMP question: are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)?
3. ***Assessable and intelligible***
   DMP question: are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are datasets provided in a way that judgments can be made about their reliability and the competence of those who created them)?
4. ***Usable beyond the original purpose for which it was collected***
   DMP question: are the data and associated software produced and/or used in the project usable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?
5. ***Interoperable to specific quality standards***
   DMP question: are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing recombination with different datasets from different origins)?

3 The CoeGSS Data Management Process

3.1 Different kinds of data sets in CoeGSS

As CoeGSS seeks to get best possible benefits out of the combination of High Performance Computing (HPC) and High Performance Data Analytics (HPDA), data set use is of utmost importance and core of the project work. Whilst this is one aspect, especially also the results will need special attention. Results in the term of CoeGSS are defined as different kind of data sets, e.g.:

- Results of Pre and Post-Processing
- Intermediate Results
- Performance Data
- Experiment Setup Data
- Input and Output Results
- …

Note: The Data Management Process will heavily depend on the Intellectual Property Rights (IPR) management within the project. Thus, a close alignment between both activities is needed and will be ensured.

3.2 Accessing and storage of data

As described in the Grant Agreement, CoeGSS shall list the relevant data also at some of the suggested places, such as:

- https://www.datacite.org
- https://www.openaire.eu

For the sake of completeness, but not pursued further within this deliverable, the Open Access approach of CoeGSS is defined as follows:

CoeGSS will follow the green Open Access approach whenever possible and will upload its deliverables to the institutional repositories set-up at several partners' sites and will also exploit the contractual right to a post-print online publication after a grace period (e.g. 6 or 12 months). Open access journals and conferences (Gold Open Access) will be targeted for dissemination as long as appropriate publication options, with high levels of impact, are available.

As Open Access publication rights are complex in particular if the green road is followed, the partners will rely on their internal Open Access consultant offices. CoeGSS partners will use their institution internal Open Access funds and in addition, the project has set aside 10.000 € for Open Access publications following the Gold model.

3.2.1 Access to existing data sets

An important service that CoeGSS will make available for its end users will be the provisioning of a data repository. One of the biggest challenges in these kinds of simulations (e.g. Synthetic Information Systems) is to get access to real-life data sets. The search for these available data sets (and meta-information about their availability and licenses amongst others) is often cumbersome and a hurdle, especially for non- or just moderate-experienced users.

Thus, CoeGSS aims to provide access to available data sets in the best possible way:

- By identifying and linking to existing data sets
- By providing specific information regarding the respective data sets.

3.2.2 Access to and or hosting of community data sets

In addition to the activities in 3.2.1, CoeGSS will also support the access to and the hosting of community data sets. That means, that the centre stakeholders may provide the infrastructure to host data sets for communities, however it has to be clear that this will finally depend on the business model (as data hosting is expensive).

3.2.3 Access to and management of CoeGSS results

The CoeGSS Consortium intends to publish data about the pilots, the conditions under which these activities are done and the potential related results. In addition to the data itself, the CoeGSS team will also use the data as basis to create success stories to show the impact of the centre and also to attract new customers.

3.2.4 Open access to research data – contractual obligations

The Grant Agreement Article 29.3 defines the handling of open access to research data as follows:

"R*egarding the digital research data generated in the action ('data'), the beneficiaries must:*

(a) 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:

> *(i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;*
>
> *(ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';*

(b) 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).

This does not change the obligation to protect results in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39, all of which still apply.

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.

3.3 The process

Figure 1 illustrates the approach taken towards Data Management and the update of the Data Management Plan within CoeGSS at a high level. It shows the different phases and breakpoints, which will be defined within the next months after finalization of this document. Initially updates of the DMP were requested in accordance with the project reviews, but this needs clarification in which format this shall happen. Nonetheless, DMP maintenance will be a continuous activity within the project, with regular updates.

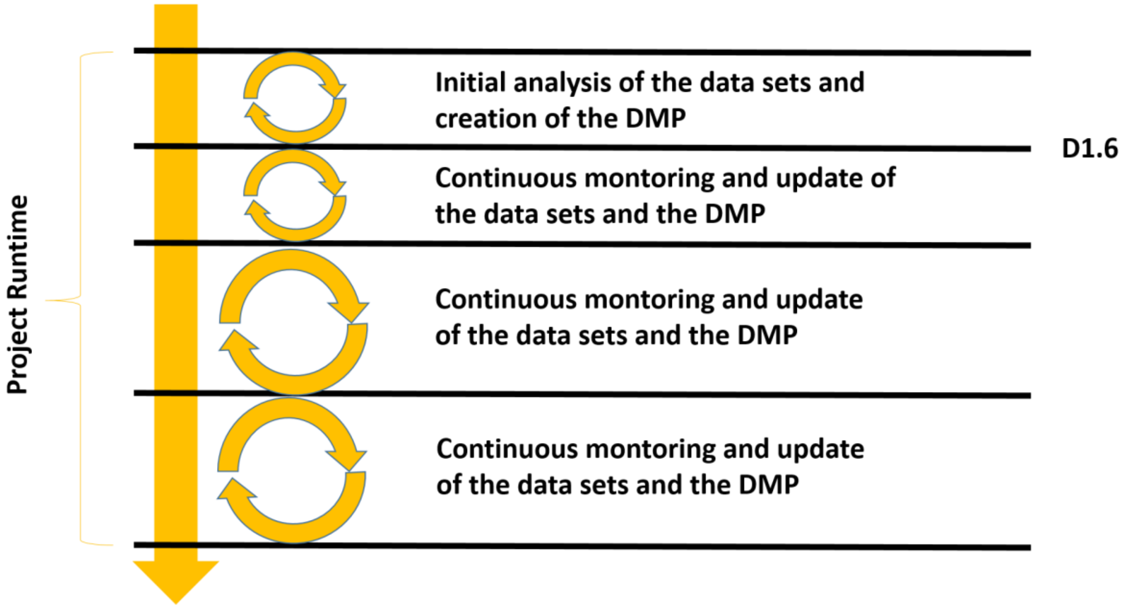
Project Runtime
Initial analysis of the data sets and creation of the DMP
D1.6
Continuous montoring and update of the data sets and the DMP
Continuous montoring and update of the data sets and the DMP
Continuous montoring and update of the data sets and the DMP

Figure 1: High Level Process of the DMP updates

3.4 Repositories

The repositories, guaranteeing the data accessibility, as chosen by CoeGSS, are:

- Final data and paper storage accessibility: Zenodo – https://www.zenodo.org/
- Staging data storage: the CoeGSS Portal provides a data storage instance based on the Comprehensive Knowledge Archive Network (CKAN) application package, where specific folders and data are made available for all kinds of CoeGSS users. Data can be shared publicly, but have not been appropriately validated so that the data will be marked as not final. Furthermore, for ease of use, all kinds of data will be annotated with meaningful metadata in order to simplify the search activity.

CoeGSS will offer the data through the Zenodo archive and the CoeGSS CKAN instance. The latter will be used as staging data repository and clearly marked as such. The data submitted to Zenodo will be appropriately referenced as final. Papers archived at Zenodo might have data associated with them.

4 So far identified data

This chapter presents the so far identified data sets for each of the CoeGSS pilots. The description will follow the template as introduced in Chapter 2.

4.1 Health Habits

4.1.1 Data Sets

Data set reference and name

CoeGSS-HealthHabits-Input: Gridded Population of the World (GPW)

Data set description

This data set describes various relevant aspects with regards to the population distribution across the world.

Standards and metadata

Data is stored in text and XML format.

Data sharing

Data can be used for research purposes with acknowledgment to the project. Data is stored locally but can be downloaded from http://sedac.ciesin.columbia.edu/data/collection/gpw-v3.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Input: Health data from European Open Data Portal

Data set description

These are health statistics about the European population.

Standards and metadata

Data is stored in CSV, JSON and XML format.

Data sharing

Open Access; http://www.europeandataportal.eu/data/en/group/health?page=4; Data can be shared with the partners of the consortium.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Input: Population data from the European Open Data Portal

Data set description

These are demographic statistics about the European population.

Standards and metadata

Data is stored in CSV, JSON and XML format.

Data sharing

Open Access; http://www.europeandataportal.eu/data/en/group/population-and-society; Data can be shared with the partners of the consortium.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Output: CoeGSS Health Habits synthetic populations

Data set description

This data set describes the synthetic population of a country.

Standards and metadata

Data is stored in CSV or JSON format.

Data sharing

Not defined yet.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Output: CoeGSS Health Habits simulation output

Data set description

This data set describes the output of a simulation of the agent based model.

Standards and metadata

Data is stored in CSV or JSON format.

Data sharing

Not defined yet.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Performance: CoeGSS Health Habits use case performance analysis data

Data set description

This data set is describing the performance data of the various runs in order to get information about the application performance on different systems.

Standards and metadata

Not defined yet.

Data sharing

Not defined yet.

Archiving and preservation (including storage and backup)

Not defined yet.

4.1.2 Application Data

Data set reference and name

CoeGSS-HealthHabits-Simulation: CoeGSS Health Habits use case agent based model

Data set description

This is the agent-based model for the spread of tobacco use.

Standards and metadata

Not defined yet.

Data sharing

Not defined yet.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Simulation: CoeGSS Health Habits use case synthetic population

Data set description

This is the code that generates synthetic population based on input data.

Standards and metadata

Not defined yet.

Data sharing

Not defined yet.

Archiving and preservation (including storage and backup)

Not defined yet.

Data set reference and name

CoeGSS-HealthHabits-Experiment: CoeGSS Health Habits use case experiment setup data

Data set description

Special scripts, required for interoperability, allowing seamless integration at HLRS and PSNC (and others). In particular, modify.sh, job-submit.sh and post-process.sh have been used.

Standards and metadata

Bash and Perl scripts.

Data sharing

job-submit.sh and post-process.sh are under the license GPL v3, whereas modify.sh is under Apache v2 license. The application can be shared under the CoeGSS Open Access policy.

Archiving and preservation (including storage and backup)

All kinds of Open Access platforms are supported, 50 MB of data are expected.

4.2 Green Growth

4.2.1 Data Sets

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, gross domestic product

Data set description

This data set is describing the gross domestic product per country in between 1990 and 2014.

Standards and metadata

Data is stored in CSV format.

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, population data

Data set description

This data set is describing the population data per country in between 1990 and 2014.

Standards and metadata

Data is stored in CSV format.

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, spatial population data

Data set description

This data set is describing a gridded data world map in 1990, 1995, 2000, 2005, 2010 and 2015.

Standards and metadata

Data is stored in CSV format.

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data set size is small, 600 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, working population

Data set description

This data set is describing the working population at an age of 14 to 65 years in between 1960 and 2014.

Standards and metadata

Data is stored in CSV format.

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, motor vehicles in total numbers

Data set description

This data set is describing the amount of motor vehicles in use in between 2005 and 2014.

Standards and metadata

Data is stored in Microsoft Excel format.

Data sharing

The data set is published under proprietary license so that sharing of data is not permitted.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, motor cars per habitant

Data set description

This data set is describing the amount of cars per 1.000 habitants in between 2003 and 2014.

Standards and metadata

Data is stored in CSV format.

Data sharing

The data set is published under proprietary license so that sharing of data is not permitted.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, electric car sales statistics

Data set description

This data set is describing the statistics of electric car sales in between 2010 and 2014.

Standards and metadata

Data is stored in CSV format.

Data sharing

The data set is published under proprietary license so that sharing of data is not permitted.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, open street map data

Data set description

This data set is describing the street map data for all countries in 2015.

Standards and metadata

Data is stored in PBF format.

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data set size is of medium size, 30 GB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, total street kilometres

Data set description

This data set is describing the overall amount of kilometres on the streets in gridded cells in 2015.

Standards and metadata

Data is stored in text format (based on the open street map data).

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data set size is very small, 10 MB in total.

Data set reference and name

CoeGSS-GreenGrowth-Input: CoeGSS Green Growth input data, Twitter data stream

Data set description

This data set is describing the online sources for Twitter to obtain social contagion.

Standards and metadata

The data is available using the Twitter Application Programming Interface (API).

Data sharing

The data set is publicly available and can be shared under the Open Access policy.

Archiving and preservation (including storage and backup)

The data size not known at the moment.

Data set reference and name

CoeGSS-GreenGrowth-Output: CoeGSS Green Growth output data, electric car sales projection

Data set description

This data set defines the main outcomes of the Green Growth pilot, in particular estimations for the global distribution of electric cars reflecting different scales of complexity will be presented.

Standards and metadata

The data is stored in HDF5 format.

Data sharing

The data set can be shared under Open Access policy.

Archiving and preservation (including storage and backup)

Currently, one execution run produces 10 GB of data. Different scales are planned.

4.2.2 Application Data

Data set reference and name

CoeGSS-GreenGrowth-Simulation: CoeGSS Green Growth simulation data, SimPop

Data set description

This data set is concerned with the generation of synthetic populations based on surveys and auxiliary data.

Standards and metadata

The data set is produced in R and therefore relies on the R proprietary formats.

Data sharing

The data set can be shared under Open Access policy.

Archiving and preservation (including storage and backup)

No information yet.

Data set reference and name

CoeGSS-GreenGrowth-Simulation: CoeGSS Green Growth simulation data, PANDORA

Data set description

This data set is concerned with the generation of synthetic populations for clusters using the agent-based simulation framework PANDORA.

Standards and metadata

All kinds of inputs are generated using C++.

Data sharing

The data set uses the GNU Public License.

Archiving and preservation (including storage and backup)

No information yet.

Data set reference and name

CoeGSS-GreenGrowth-Simulation: CoeGSS Green Growth simulation data, Pandas library

Data set description

This data set represents a high-performance data analytics tool.

Standards and metadata

This data set relies on the programming language Python.

Data sharing

For this data set, the BSD license is used.

Archiving and preservation (including storage and backup)

No information yet.

Data set reference and name

CoeGSS-GreenGrowth-Experiment: CoeGSS Green Growth experiment setup data

Data set description

This data set defines the relevant data such as scripts and configurations to execute the simulations.

Standards and metadata

To be defined.

Data sharing

To be defined.

Archiving and preservation (including storage and backup)

No information yet.

4.3 Global Urbanisation

4.3.1 Data Sets

Data set reference and name

CoeGSS-Urbanization-Input: CoeGSS Global Urbanization use case input data

Data set description

This data set defines the input data for the Global Urbanization pilot. Therefore, specific data sets are required to define the simulations, which rely on population and socio-economic data, pollution indicators as well as transportation data.

Standards and metadata

Data is stored in traditional text and CSV formats, however SQLite databases are considered as well.

Data sharing

The data sets are publicly available so that sharing of data is in general permitted.

Archiving and preservation (including storage and backup)

Data need to be stored or linked, however the amount of data cannot be estimated yet.

Data set reference and name

CoeGSS-Urbanization-Output: CoeGSS global Urbanization use case output data

Data set description

This data sets describes all kinds of output of the Global Urbanization use case.

Standards and metadata

The data format relies on text base and CSV, however SQLite databases are considered as well.

Data sharing

For simulation, the CoSMo platform is applied, which relies on proprietary licenses. This statement also holds for data that is created using the platform.

Archiving and preservation (including storage and backup)

Data need to be stored, however the amount of data cannot be estimated yet.

Data set reference and name

CoeGSS-Urbanization-Performance: CoeGSS Global Urbanization use case performance analysis data

Data set description

This data set is describing the performance data of the various application runs in order to get information about the performance on different systems.

Standards and metadata

Data is stored in traditional as well as CRAY proprietary formats.

Data sharing

The data is used to optimize the production runs on the clusters. Data can be shared under Open Access policy.

Archiving and preservation (including storage and backup)

Data need to be stored, however the amount of data cannot be estimated yet.

4.3.2 Application Data

Data set reference and name

CoeGSS-Urbanization-Simulation: CoeGSS Global Urbanization use case simulation application

Data set description

This data set represents the code to generate synthetic populations for the CoSMo platform.

Standards and metadata

The CoSMo simulation suite relies on proprietary formats.

Data sharing

The CoSMo simulation suite applies a proprietary license, which prevents data sharing in general.

Archiving and preservation (including storage and backup)

Data need to be stored, however the amount of data cannot be estimated yet.

Data set reference and name

CoeGSS-Urbanization-Experiment: CoeGSS Global Urbanization use case experiment setup data

Data set description

All additional scripts and application customizations refer to this data set.

Standards and metadata

Standardized Bash and Python scripts.

Data sharing

As those scripts rely to the CoSMo simulation suite, the same licenses hold so that sharing of data is not possible.

Archiving and preservation (including storage and backup)

Data need to be stored, however the amount of data cannot be estimated yet.

5 Future plans

As already highlighted in the sections above, this document is based on available data and acts as a baseline for the CoeGSS Data Management Plan. However, at the current point of time, neither can all data sets be identified at this time nor are all of them completely analysed. As a consequence, solid plans and detailed actions are required in order to deal with upcoming data sets and their possible issues.

The CoeGSS DMP for the following months foresees several plans and actions:

- **Additional data sets**
  CoeGSS is dealing intensively with data analytics, which has been highlighted in the various project deliverables. Nevertheless, there may be data gathered that are not directly involved in the analytics or computation phase, such as website data or questionnaires as well as communication with portal users.
  During the lifetime of the project, mechanisms to share added value information with the outside world will be defined and instantiated, as long as the information does not correspond to any intellectual property right of a user or the project itself.
- **Incomplete data sets**
  The described data sets in Chapter 4 show the amount of data that CoeGSS is dealing with. As highlighted, the presented data is not complete, the growing applications demand fine-grained information and additional data sources, so that further information is expected. In addition, the described data sets are not complete, especially data sharing regularities are not defined for all mentioned data sets when writing this document.
  Therefore, CoeGSS will overcome that situation by regular questionnaires and legal consultation to compile the relevant information and to provide a full set of available data to the project and third party users.
- **Licensed data sets**
  In order to make licensed data sets available to other users and stakeholders, there are several investigations possible. For example, for data that are owned by CoeGSS stakeholders, a non-disclosure agreement with the interested party is conceivable to share the data. However, this procedure does not apply for all kinds of licensed data so that CoeGSS will define the appropriate methods and procedures to make as much data as possible available to its users.
- **Data controller**
  For data, which are generated by the portal, the website or other possible access points, CoeGSS follows the process of instantiating a so called "data controller" whenever data property is not directly clear.
  The data controller is represented as an institution or body that determines the purpose and means of processing the personal data. In particular, the controller is responsible for ensuring the quality of data as well as ensuring the safety measures protecting the data. In general, the data controller is reflected as the hosting operator of the service whose storage responsibility is granted officially by the Executive Centre Management.

6 Conclusions

The importance of data for the CoeGSS activity requires focused and well elaborated processes to ensure a proper data management, which needs to take into account contractual obligations, impact generation and finally the IPR of data as used and produced.

This document provides an initial Data Management Plan, based on the actual research of the available facts within the project (data, generated data, IPR, Grant Agreement and Consortium Agreement), according to the EC Guidelines and finally, based on the research of the available Open Access Repositories.

We have currently opted to gather all these facts under this deliverable, to aim for completeness of the data within the project and to further refine the possible formats, either internally, to improve understanding (i.e., decrease ambiguity) and also externally, to cater for mining of data (the inclusion into the Zenodo repository).

7 References

7.1 List of Abbreviations

API	Application Programming Interface
BSD	Berkeley Software Distribution
CA	Consortium Agreement
CKAN	Comprehensive Knowledge Archive Network
CoeGSS	Centre of Excellence for Global System Science
CSV	Comma-separated values
DMP	Data Management Plan
DoA	Description of Action
EC	European Commission
EGI	European Grid Infrastructure
ESFRI	European Strategy Forum on Research Infrastructures5
GA	Grant Agreement
GB	Gigabyte
HPC	High Performance Computing
HPDA	High Performance Data Analysis
JSON	JavaScript Object Notation
IPR	Intellectual Property Rights
MB	Megabyte
TB	Terabyte
ToR	Terms of Reference
WP	Work Package
XML	Extensible Mark-up Language