



<b>Project acronym:</b>	U_CODE
<b>Project title:</b>	Urban Collective Design Environment: A new tool for enabling expert planners to co-create and communicate with citizens in urban design
<b>Call:</b>	H2020-ICT-2015
<b>Grant Agreement Number:</b>	688873

<b>Deliverable number:</b>	D1.3 (D4)
<b>Deliverable title:</b>	Data Management Plan
<b>Deliverable type:</b>	Report
<b>WP number and title:</b>	WP1: Project Management
<b>Dissemination level:</b>	Public
<b>Due date:</b>	Month 6 – 31 July 2016
<b>Lead beneficiary:</b>	TUDr – Coordinator
<b>Lead author(s):</b>	Sander Münster, Claudia Hawke, Anja Jannack, TUDr
<b>Contributing partners:</b>	all

## Document history

Version	Date	Author/Editor	Description
0.1	17.06.2016	Claudia Hawke, Sander Münster	Initial version sent to partners
0.2	19.07.2016	Anja Jannack	Incorporated partner input
0.3	21.07.2016	Sander Münster	adding paragraphs, editing
0.4	25.07.2016	Ulrich Hartmann, conject	conjectPM description + suggestions & comments
0.5	27.07.2016	Anja Jannack, Jörg Rainer Noennig	adding paragraphs, editing
1.0	29.07.2016	Claudia Hawke	Published version



## Table of contents

1. Executive Summary .....	4
2. Applied Methodology .....	4
2.1 Data set description .....	5
2.2 Data set reference and names .....	6
2.3 Data sharing .....	6
2.4 Standards and metadata .....	7
2.5 Archiving and preservation .....	11
2.5.1 Storage, backup, replication and versioning in U_Code .....	11
2.5.2 Long term data sharing platform .....	12
3. Budget .....	13
4. Attachment 1: Initial Datasets in U_CODE .....	13

## List of figures

Figure 1: Data Management Template .....	5
Figure 2: Upload Dialog with mandatory categories .....	7
Figure 3: Upload dialog with company category selections .....	8
Figure 4: Upload dialog with document type category selection .....	8
Figure 5: Upload dialog with Topic category selections .....	9
Figure 6: Upload dialog with Work Package category selections .....	9
Figure 7: Upload dialog with selected category choices (example) .....	10
Figure 8: Search options (example) .....	10



## 1. Executive Summary

In this report the initial Data Management Plan (DMP) for the U\_CODE project is presented. The report outlines how research data will be handled during and after the project duration. It describes what data will be collected, processed or generated with which methodologies and standards, whether and how this data will be shared or made open, and how it will be curated and preserved.

The Data Management Plan (DMP) describes the data management life cycle for all data sets. The purpose of the DMP is to provide an analysis of the main elements of the data management policy that will be used in U\_CODE with regard to all data sets that will be generated by the project. The data collected and generated by the different U\_CODE partners will have multiple formats. In general four different types are generated and processed 1.) text based data, 2.) visual based data sets, 3.) models, and 4.) software / source code data sets.

The Data Management Plan provides information on the following points:

- Data set description
- Data set reference and name
- Data sharing
- Standards and metadata
- Archiving and preservation (including storage and backup)

The DMP gives a first overview on the diversity, scale and amount of data which will be handled during the U\_CODE project. While the project is ongoing, conjectPM is used as the collaboration platform for the management of U\_CODE data.

The DMP is not a fixed document, but evolves during the lifespan of the project.

## 2. Applied Methodology

The methodology applied for drafting this initial DMP of U\_CODE is based on guidelines of the European Commission<sup>1</sup>. According to these guidelines all U\_CODE partners were asked to list and describe their datasets. The compiled list is presented in attachment 1<sup>2</sup> at the end of this document. The tables give details about the datasets generated in the project. These various datasets are stored at conjectPM for (internal) use during the project duration. Which data sets will be stored for open access will be decided later in the project.

This list addresses the main points on a dataset by dataset basis and reflects the current status of discussion and reflection within the consortium about the data that is going to be produced within the U\_CODE project. This list will evolve and develop over the lifetime of the project and will be kept up to date on the U\_CODE collaborative platform conjectPM.

---

<sup>1</sup>[https://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)

<sup>2</sup> Not intended for print on A4 paper.

## 2.1 Data set description

The data collected and generated by the different U\_CODE partners will have multiple formats and vary in size from a few MB's to several GB's. The formats range from interview transcripts, survey results, protocols, pictures, visual recordings up to software prototypes, and test data. So far four types of general data sets are identified:

- **text based data:** interviews, surveys (scientific), publications, reports,
- **visual data:** logfiles graphs, visual protocols, pictures, UML diagrams
- **models:** models, digital models, conceptual framework
- **software data:** prototype, software prototypes, test data, source code

The Initial DMP template asked the U\_CODE partners to describe their different data sets according to the following items:

*DATA SET – name; DATA SET - nature of data; Lead; WP; Task/ Deliverable, time in which data is generated/collected, type of data, data format, publication date, source of data, how is the data generated/collected, how is the data processed; restriction on using the data; standards; metadata; data sharing; preservation and backup; duration of preservation (short-term, long-term, ...), related dataset; underpins scientific publication; License*

Nr	DATA SET - name	DATA SET Type - nature of data	Lead	WP	Task/ Deliver.	time in which data is generated/collected	Type of data	data format	Publication Date	Source of data
Explanation & filling examples		eg. interviews, survey results, software prototypes, software, publications, production, test data, conceptual framework, models,	TU Delft, TU Delft, SEEN, COIN, DPT, SITSa, GMP	1-3			audio, video, text, pictures, code, models...	xls, docx, jpeg, pdf, ppt, mp3 ...		
Data set 1										
Data set 2										

how is the data generated/collected	how is the data processed	Restriction on using the data, suggestions by now	audience, if yet known	standards	metadata	data sharing
		open access, open to qualified researchers, confidential: only for U_CODE members	e.g. other research groups, users of ...	reference to existing suitable standards of the discipline.	If standards do not exist, an outline on how and what metadata will be created.	Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and access software and other tools for enabling re-use, and definition of whether access be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identifying, 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 (ethical rules of personal data, intellectual property, commercial, privacy-related security-related).

preservation and backup	duration of preservation (short-term, long-term, ...)	related dataset	underpins scientific publication	License
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			

Fig. 1: Data Management Template

Due to the fact that data collection and creation is an ongoing process, questions such as the detailed description of data nature, exact scale, to whom those data may be useful or if these data underpin a scientific publication will be answered in the updated versions of the DMP. Moreover the question on the existence or non-existence of similar data and the possibilities for integration and reuse are not finally agreed between the U\_CODE partners and will be reported later.

## 2.2 Data set reference and names

A first collection of datasets has been compiled in Attachment 1 at the end of this document. A comprehensive pattern for naming the produced datasets of the project to be published open access is going to be developed. As an example one approach could be the following:

UCODE\_Data\_"WPNo"."DatasetNo."\_"DatasetTitle"UCODE\_Data\_WP1.1\_UserGeneratedContent). This depends also on the long term data sharing platform to be chosen.

conjectPM is used to organize, manage and monitor the collected and generated data sets of the U\_CODE project. Due to the structure of the collaboration platform conjectPM (for a detailed explanation see Section 2.4) a unified name structure is not necessary to handle the various data sets during the project duration.

## 2.3 Data sharing

ConjectPM is used to share and manage the collected and generated data sets within the U\_CODE project. It provides a well-organized structure to make it easy for research teams to find, better understand and reuse the various data by creating a consistent and well structured research data pool (see also 2.4).

**Open access policy:** By default all of the created data in U\_CODE shall be made available open access. Reasons for not making the data open will derive from

- **legal properties** (e.g. missing copyrights, participant confidentiality, consent agreements, or intellectual property rights)
- **scientific** and/or **business** reasons (e.g. pending publications, exploitation aspects)
- **technical issues** (e.g. incomplete data sets).

The collected and generated data can be classified into two categories 1) **short term intermediate data** (stored at conjectPM), and 2.) **long term data** (stored in repositories, such as ZENODO or OpARA). The long term data have different levels of open accessibility:

- data with restricted access to the U\_CODE partner creating this data set;
- data with restricted access to U\_CODE project partners;
- data that is to be published and shared as open source to researchers only;
- data that is to be published and shared as open source to everyone.

The decisions on data publication and the level of accessibility will be taken per dataset and by the responsible U\_CODE partner who created the dataset. This will be documented



in this (or future) versions of the data management plan. The updated version of the DMP shall detail the information on data sharing, including access procedures, embargo periods, and outlines of technical mechanisms for dissemination for open accessible data sets.

Strategies to limit restrictions may include: anonymising or aggregating data. Questions to be considered when further developing the open access policy in U\_CODE are:

- How do we make the data available to others?
- With whom are we sharing the data, and under what conditions?
- What kind of restrictions are needed and why?
- What actions are we planning to minimise these restrictions?

## **2.4 Standards and metadata**

The U\_CODE project will create diverse data to detail project content and moreover create data needed to enable other researchers to use and regenerate output data in a systematic way. The documentation can take the form of publications, manuals and README files on how to use the software, in addition to scripts for running the software.

To enable a consistent description of all datasets provided by the project, a template table is used to describe metadata of each dataset including title, author, description, formats, etc. (see attachment 1). U\_CODE partners collect and create data sets on their own or by co-creating these data sets together. Due to the diversity of the project partners involved there were no community data standards identified yet.

The collaboration platform – conjectPM – used for the management of U\_CODE enforces the categorization of any document uploaded in order to impose a common structure in the metadata of the document repository. On project initialization, participants agreed on the following mandatory categories to be assigned to a document: Company, Document Type, Topic and Work Package. The values assignable to the respective categories are shown in the following screens.

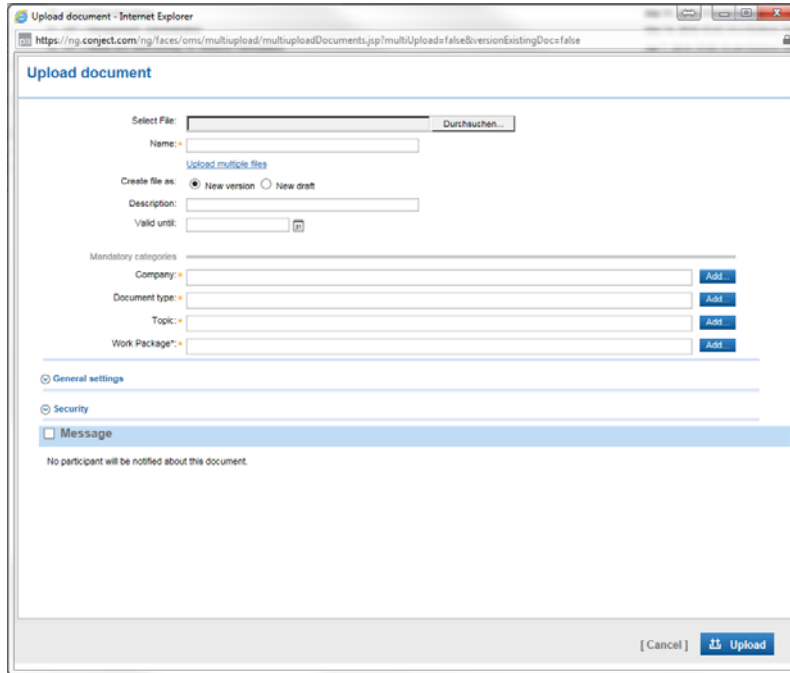


Fig. 2: Upload Dialog with mandatory categories

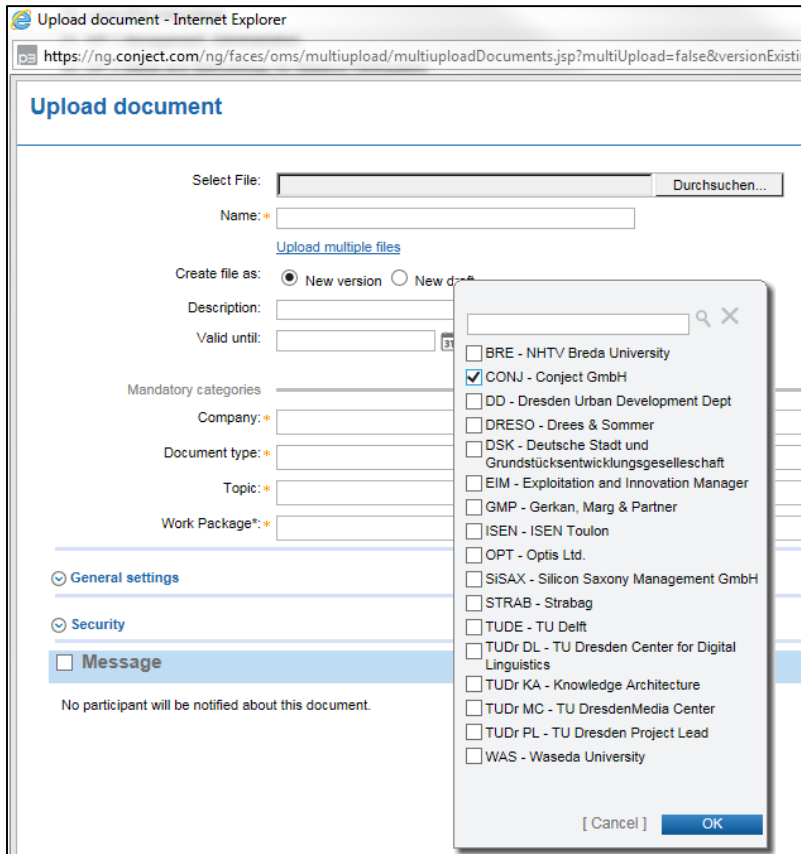


Fig. 3: Upload dialog with company category selections



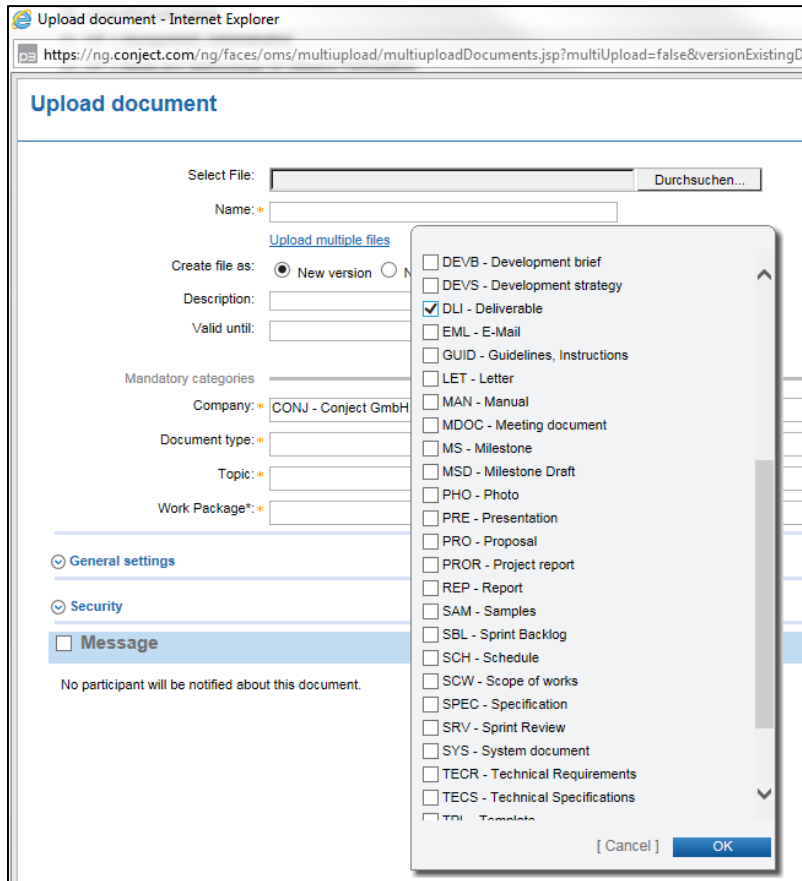


Fig. 4: Upload dialog with document type category selection

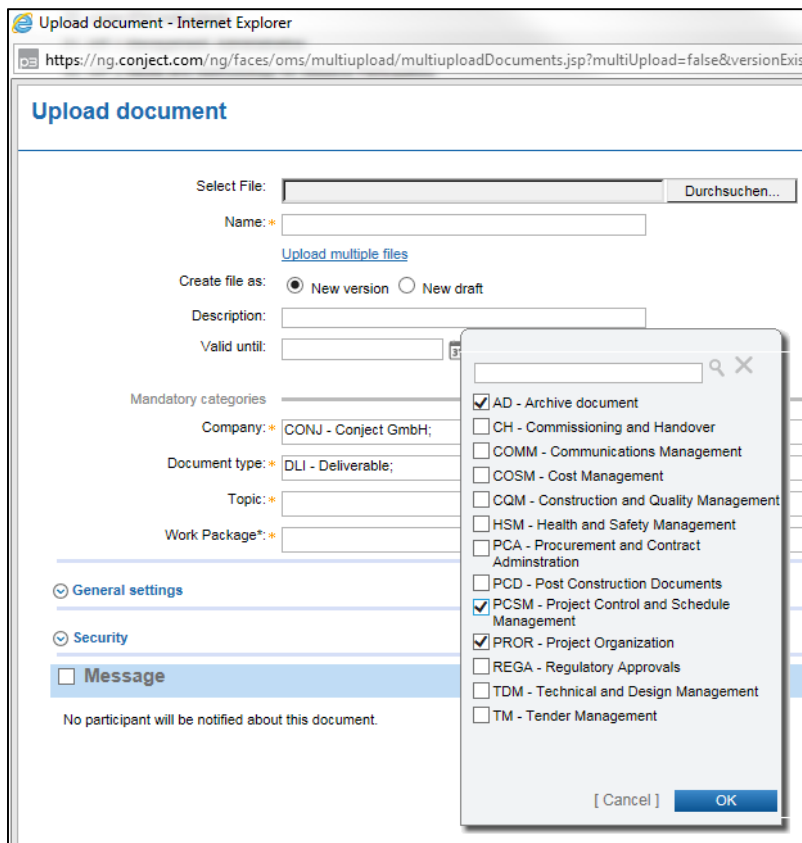


Fig. 5: Upload dialog with Topic category selections

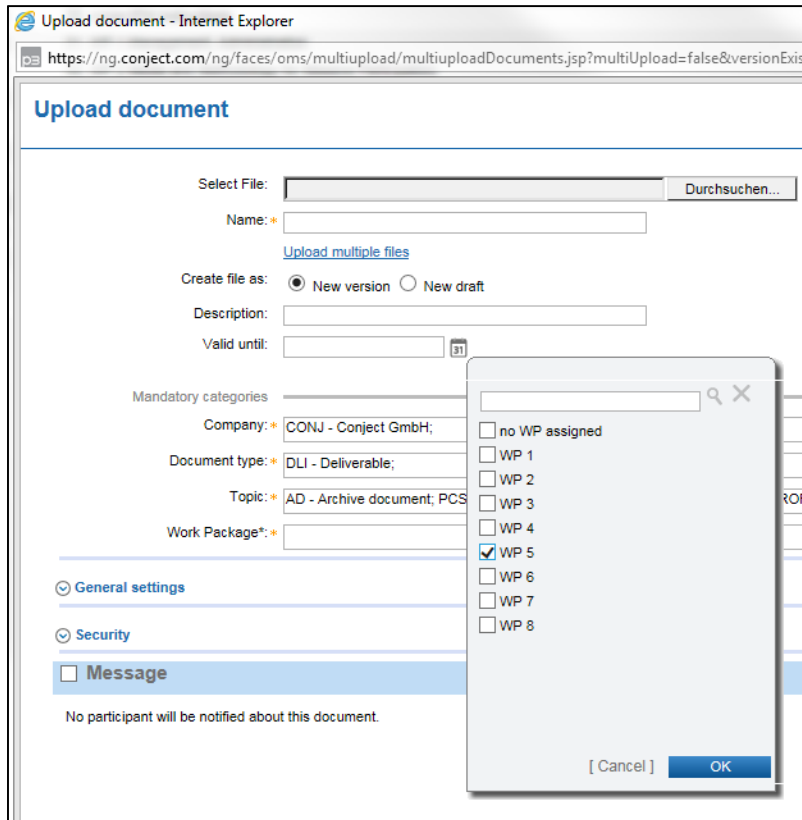


Fig. 6: Upload dialog with Work Package category selections

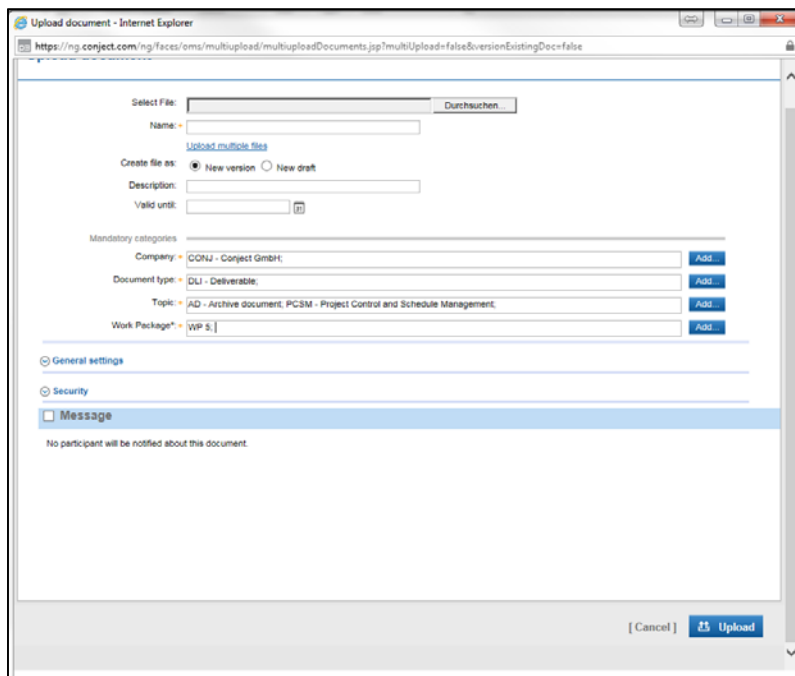


Fig. 7: Upload dialog with selected category choices (example)

Document retrieval can then be conducted on the basis of categories, as shown in the Advanced Search dialog.

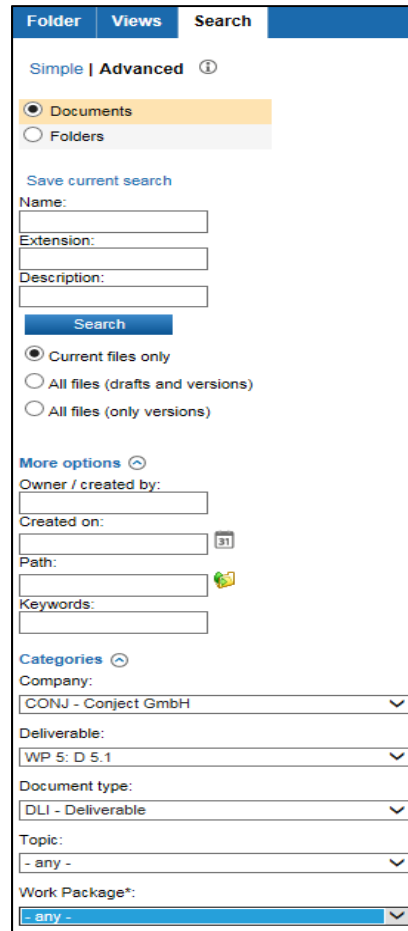


Fig. 8: search options (example)

Category settings can be adapted during the project if necessary. However, the addition or removal of categories is not downward compatible (additions) and might render existing documents invisible via category search. Hence the removal of existing categories is not advisable. However, adding more choices to any existing category is entirely non-critical. Besides project-specific categories the default categories Owner/created by, Creation date, Path in the document tree and Keywords are always in place. Keyword search is made available by full-text scanning of the entire document on document upload.

## 2.5 Archiving and preservation

### 2.5.1 Storage, backup, replication and versioning in U\_CODE

Intermediate data generated by the U\_CODE partners will be stored in the U\_CODE collaborative platform conjectPM. This repository can be easily accessed by all partners. It includes all the publications, raw data, reviews, all Deliverables and the management of the U\_CODE project.

### Data Security at conject Data Centres

The conjectPM systems are located in two self-sufficient and geographically separated facilities. During normal operation the system load is balanced across the two locations. In



the unlikely event that either one of the data centres becomes unavailable the remaining one can take over full operation and guarantee the availability of all customer data. The coniectPM file system consists of an array of independent storage units. It maintains at least three copies of each file spread across the two locations. Failures of storage units are automatically detected and handled by recreating the data on other storage units. Storage units can be added or replaced while the system stays fully operable, ensuring that sufficient capacity is always available when required.

Core system components are secured against failure by duplication of power supplies, CPUs, storage devices and network connections. All hardware components have secondary devices in place for failover contingency. Due to the high levels of resiliency in place coniectPM guarantees a 99.5% availability SLA to all of their clients around the globe.

The U\_CODE project on the coniectPM platform has been configured to match the overall U\_CODE work package structure. Access rights to documents have been set according to the work package leader. A general section in the project folder structure is set up for administrative purposes and information exchange between U\_CODE partners.

### *2.5.2 Long term data sharing platform*

Selected data from the coniectPM repository will be shared publicly during or after the life time of the project. All long term data collected or generated will be deposited in a repository. If required, the entire information content of the U\_CODE project can be stored on disk for archiving. This functionality can also be used to transfer U\_CODE content to another system. The final repository has not been chosen yet. The choice of repository will depend on:

- location of repository
- research domain
- costs
- open access options
- prospect of long-term preservation.

#### **ZENODO repository:**

One of the repositories considered is ZENODO <https://zenodo.org/>. This is online, free of charge storage created through the European Commission's OpenAIREplus project and is hosted at CERN, Switzerland. It encourages open access deposition of any data format, but also allows deposits of content under restricted or embargoed access. Contents deposited under restricted access are protected against unauthorized access at all levels. Access to metadata and data files is provided over standard protocols such as HTTP and OAI-PMH.

Data files are kept in multiple replicas in a distributed file system, which is backed up to tape every night. Data files are replicated in the online system of ZENODO. Data files have versions attached to them, whilst records are not versioned. Derivatives of data files are



generated, but the original content is never modified. Records can be retracted from public view; however, the data files and records are preserved. The uploaded data is archived as a Submission Information Package in ZENODO. Files stored in ZENODO will have MD5 checksum of the file content, and it will be checked against their checksum to assure that a file content remains correct. Items in the ZENODO will be retained for the lifetime of the repository which is also the lifetime of the host laboratory CERN which currently has an experimental programme defined for the next 20 years. Each dataset can be referenced at least by a unique persistent identifier (DOI), in addition to other forms of identifications provided by ZENODO.

### **OpARA repository**

Another option is provided by the Technische Universität Dresden, which is currently setting up an institutional, inter-disciplinary repository with long-term archive in the project OpARA. It will provide open access long-term storage of data, including metadata and will go into production in 2017.

Other institutional and thematic repositories will be considered and evaluated in the next months.

## **3. Budget**

The costs of preparing the data and documentation will be borne by the project partners. This is already budgeted in the personnel costs included in the project budget.

The permanent costs of preserving datasets on the ZENODO repository will be free of charge as long as the single dataset storage is no greater than the maximum 2GB of data.

The permanent costs of preserving datasets on the OpARA repository are planned to be free of charge for TUD members. But the final decision on costs has not been taken.

## **4. Attachment 1: Initial Datasets in U\_CODE**

Initial Datasets in U\_CODE sorted by U\_CODE partners<sup>3</sup>

Initial Datasets in U\_CODE sorted by DATA SET Type<sup>4</sup>

---

<sup>3</sup> Not intended for print on A4 paper.

<sup>4</sup> Not intended for print on A4 paper.



