ACTiCLOUD: ACTivating resource efficiency and large databases in the CLOUD

Project No: 732366

H2020-ICT-2016-1

D5.2: Data Management Plan

Due date of deliverable:	M6 (2017/06/30)
Actual submission date:	M6 (2017/06/30)

Executive summary:

The ACTiCLOUD Data Management Plan (DMP) describes what kind of data is generated or collected in the ACTiCLOUD project and outlines the Consortium's policy regarding how this data will be curated, preserved and made accessible. The DMP is a "living" document that will evolve through the ACTiCLOUD project lifespan, in order to reflect both the evolution in the type and form of generated data sets as well as changes in the Consortium's policies.

List of authors:

Author	Affiliation
Konstantinos Nikas	ICCS
Georgios Goumas	ICCS
Nectarios Koziris	ICCS

Dissemination Level	X	**PU (Public)**
		PP (Restricted to other programme participants)
		RE (Restricted to a group specified by the consortium)
		CO (Confidential, only for members of the consortium)
	Where restricted, access granted to:	
Nature	X	**ORDP (Open Research Data Pilot)**
		P (Prototype)
		D (Demonstrator)
		O (Other)

Review Status		Draft
		WP Leader accepted
		QA approved
	X	**Coordinator accepted**

Revision History:

Version	Author(s) (Affiliation)	Notes
0.1	Konstantinos Nikas (ICCS)	Initial Draft
0.2	Jim Webber (NEO)	Review
0.3	John Thomson (OnApp)	Review
1.0	Konstantinos Nikas (ICCS)	Final Release

ACTiCLOUD Consortium:

Participant No	Participant organisation name	Short name	Country
1 (Coordinator)	Institute of Communication and Computer Systems	ICCS	Greece
2	Numascale AS	NSCALE	Norway
3	Kaleao Limited	KALEAO	UK
4	OnApp Limited	ONAPP	Gibraltar
5	University of Manchester	UNIMAN	UK
6	MonetDB Solutions B.V.	MDBS	Netherlands
7	Neo Technology	NEO	Sweden
8	UMEA University	UMU	Sweden

Confidentiality:

This document contains proprietary and confidential material of certain ACTiCLOUD contractors, and may not be reproduced, copied, or disclosed without appropriate permission. The commercial use of any information contained in this document may require a license from the proprietor of that information.

THIS DOCUMENT IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES

(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS DOCUMENT, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Table of Contents

List of Abbreviation

Abbreviation / Acronym	Meaning
DMP	Data Management Plan
GA	Grant Agreement
CA	Consortium Agreement

1 Introduction

1.1 Purpose of this Document

The main purpose of the ACTiCLOUD Data Management Plan (DMP) is to provide a single point of reference on the policy that governs the data generated and managed by ACTiCLOUD as well as any data made available to the public. The DMP outlines the method and the relevant actions required for the curation, preservation and exploitation of the data generated and/or collected during the project.

1.2 Evolution of the DMP

The first version of the DMP is delivered in M6 of the ACTiCLOUD project, in line with the H2020 guidelines for the data management plan creation [1]. It is important to note however, that the DMP will keep evolving during the lifespan of the project.

More specifically, the ACTiCLOUD consortium anticipates revisions of the DMP to be delivered in:

1. **M14**: The first ACTiCLOUD prototype will be delivered in M12 and its initial evaluation will be concluded in M14. At this point, it will be clear what data can be generated and collected by the ACTiCLOUD prototypes.

2. **M18**: The initial versions of ACTiCLOUD results are expected to be delivered in M18 (deliverables D2.1, D2.2, D3.1, D3.2, D3.3 and D3.4). All these components are required for re-using and/or reproducing the ACTiCLOUD generated data. Therefore, at this point the ACTiCLOUD consortium will be able to specify how the necessary tools will be licenced to permit the widest reuse possible, whether the data and tools produced in the project are useable by third parties, as well as specify the length of time for which they will remain re-usable.

3. **M36**: At the end of the project, the final ACTiCLOUD prototype will be delivered together with the final versions of the ACTiCLOUD results. Therefore, at this point the ACTiCLOUD consortium will be able to update and finalize all the policies regarding data reuse that were described in older versions of the DMP, as well as finalize how data will be curated and preserved after the end of the project, in accordance with the rules set out in the GA and CA agreements that all the partners of ACTiCLOUD have signed.

As it is evident from the above timeline, the second and third revision of the DMP will be delivered in time with the interim and final evaluation/assessment of the project, as suggested in the H2020 guidelines [1].

2 ACTiCLOUD Data Management Plan

2.1 ACTiCLOUD strategy regarding the Open Data Research Pilot

The DMP is aligned and serves the ACTiCLOUD aim to contribute data to the open research pilot. Data sets which are candidates for sharing will be checked to ensure that:

1. They are not confidential and do not include commercially sensitive information.

2. They comply with the GA and CA agreements that all the partners of ACTiCLOUD have signed.

3. Sharing the data does not damage exploitation or IP protection prospects.

Any dataset approved for contribution to the Open Data Research Pilot will be appropriately licensed by the ACTiCLOUD consortium, using for example Creative Commons or public domain. ACTiCLOUD will then make the dataset available through the relevant and suitable open access repositories. Where data must be embargoed to support IP protection or exploitation, a timeline for its release will be provided.

2.2 Data description & data types

The data collected, generated and used in the project includes but is not limited to the following types:

1. **Data center workloads**. Typical and nontypical cloud workloads to be used for the evaluation of the ACTiCLOUD prototypes.

2. **Performance related data.** Technical data related to the performance of the different components developed in ACTiCLOUD together with key performance indicators and aggregated results.

3. **Deliverables**. Written reports that describe the work performed in the project and its outcomes. The level of access to the deliverables produced is regulated by the Grant and Consortium Agreements.

4. **Reports & documentation.** Written reports and documentation such as written guidelines, meeting minutes, presentations, posters, promotional materials, surveys, interviews, etc. fall into this data type. The level of access to the reports produced is regulated by the Grant and Consortium Agreements.

5. **Scientific publications**. Publications in relevant scientific journals, books, and conferences which report on the work in the project. All project related publications will contain an explicit acknowledgment to the project, in which the name and the EU grant number will be mentioned.

2.3 Data formats, standards & metadata

ACTiCLOUD will strive to employ unified, widely accepted data formats and standards when making project data publicly available. Below, we present the selected data format for each of the

four data types identified in Section 2.2.

Each dataset will also be accompanied by metadata containing information about the origin of the data, its size, language, intellectual properties (where relevant), date of creation, version, and other important details. The exact metadata standard that will be followed, will be determined at a later point.

2.3.1 Data center workloads

The format will be defined in a later point.

2.3.2 Performance related data

The format will be defined in a later point.

2.3.3 Reports & Scientific publications

Generally, all data which falls under the Reports category will be available in PDF format. The partners have already agreed on templates for deliverables, presentations, and posters. This will assure that all reports produced by each partner have a unified format which is characteristic to the project.

2.3.4 Scientific publications

Scientific publications will follow the format required by the conferences or journals in which these publications will appear. Whenever possible, we will try to provide a stand-alone PDF version of the publication.

2.4. Data storage, sharing & reuse

2.4.1. Internal data storage & sharing

As described in deliverable D6.1, all data currently produced in ACTiCLOUD, i.e. reports & scientific publications, is stored in an access-restricted Google Drive. This repository has been selected as it minimizes the risk of crucial data loss and at the same time it is used through a modern web interface which makes it easy and attractive to use. The ACTiCLOUD Google Drive is managed by ICCS and every partner has equal access to it.

When the project has matured enough to reach the point where the rest of the data types, i.e. data center workloads & performance related data, are produced in ACTiCLOUD, ICCS will setup for them a secure data repository that will be backed up on a daily basis to minimize the risk of data loss. This repository is not expected to capture any sensitive or personal information.

2.4.2 Public data sharing & reuse

ACTiCLOUD intends to use several platforms to publish our results and datasets. The first platform is the project webpage (https://acticloud.eu) which was set up at the start of the project and is hosted and maintained by ICCS. The website contains a dedicated area for downloads, which will provide the necessary links to public data together with the necessary metadata.

ACTiCLOUD intends to share its datasets also in a publicly accessible disciplinary repository using

descriptive metadata as required/provided by that repository. Data that are shared will further include standards and notations needed for its interpretation, following commonly accepted practices in the field. Files and folders will be versioned and structured by using a name convention consisting of project name, dataset name, and ID.

The repository will be determined at a later point. Currently, the main candidate is Zenodo, a research data archive / online repository created through EC's OpenAIRE+ project that is now hosted at CERN using one of Europe's most reliably hardware infrastructures. Zenodo not only supports the publication of scientific papers or white papers, but also the publication of any structured research data (e.g. using XML).

2.5 Archiving & long-term preservation

At the time of this first version of the DMP, plans for archiving are preliminary and fluid. Near the end of the project, we will know more on the approximate end volume of the produced data and thus the associated costs for preservation, which will allow the definition of the procedures that will be put in place for long-term preservation of data.

After the end of the project, ACTiCLOUD will prepare an unalterable version of the data gathered to make available to any researcher by request. To ensure high-quality long-term management and maintenance of the dataset, the consortium will implement procedures to protect information over time. These procedures will permit a broad range of users to easily obtain, share, and properly interpret both active and archived information, and they will ensure that information is kept up to date in content and format so they remain easily accessible and usable.

3. Conclusions

This document is the initial DMP that provides a short, general outline of the consortium's policy for data management. The DMP is conforming to the EU directives and initiatives to promote research via the Open Access to Data. It describes the data to be generated or collected and provides an explanation of the different types of data which will be produced, plans for documenting this data and file formats where possible. It also discusses metadata as well as data sharing and long-term archiving preservation.

The DMP is a living document that will evolve through the ACTiCLOUD project in order to capture and reflect evolution in the form of dataset updates and/or changes in Consortium policies.

4. References

[1] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020, version 3.0, July 2016. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf