PROJECT DELIVERABLE

PARCOS
Participatory Communication of Science

DELIVERABLE 1.1
Data Management Plan

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 872500.

“ParCos – Participatory Communication of Science”
A HORIZON 2020 RESEARCH AND INNOVATION ACTION

Consortium: Lappeenrannan-Lahden teknillinen yliopisto (FI, coordinator), Katholieke Universiteit Leuven (BE), Vlaamse Radio- en Televisieomroeporganisatie (BE), and Knowle West Media Centre LBG (UK).

Webpage: https://parcos-project.eu
Duration: 1/2020 – 12/2022
Grant: H2020-872500 (Call H2020-SwafS-2019-1)

Contact (co-ordinator):
Asst. Professor Antti Knutas & Dr. Annika Wolff
LUT University
e-mail: parcos.project@lut.fi

Disclaimer: This document’s contents are not intended to replace consultation of any applicable legal sources or the necessary advice of a legal expert, where appropriate. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user, therefore, uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors' view.

DESCRIPTION OF THE DELIVERABLE

Overview	Details
Authors	Annika Wolff, LUT University (FI) Antti Knutas, LUT University (FI)
Reviewers	Luke Sweeney, KWMC (UK)
Number of Deliverable	1.1
Title of Deliverable	Data Management Plan
License	CC BY 4.0, see https://creativecommons.org/licenses/by/4.0/
Attribution	CC BY 4.0 ParCos, http://parcos-project.eu H2020-872500
Dissemination Level	Public
Contractual delivery date	2020-06-30
To be cited as	Wolff A., and Knutas, A. (2020). Data management plan, deliverable 1.1 of the Horizon 2020 project ParCos, EC grant agreement no 872500, Lappeenranta, Finland.

SUMMARY

This report is the ParCos project Data Management Plan. It describes the practices that the project will adhere to regarding the use of data within the project. This includes data that supports project activities and data generated by project activities. The report will outline the responsibilities of the ParCos project in relation to the Open Research Data Pilot (ORDP). The report outlines how ParCos approaches will over time contribute to better understanding of how to curate these types of data for reuse. The DMP is a critical part of participation in the ORDP and contributes towards the Open ParCos Project data deliverable (D5.7) that will be delivered in M34 of the project. The DMP will be kept up to date and the final version will form part of this second deliverable.

TABLE OF CONTENTS

1 Introduction to ParCos Data 6

2 Open Research Data Pilot 8

2.1 Updating of Data Management Plan, based on ParCos Activities 8

2.2 Research Data Repository 8

2.3 Ensure third party use of data 9

2.3.1 Findable 9

2.3.2 Accessible 10

2.3.3 Interoperable 10

2.3.4 Reusable 10

2.4 Provide information and tools needed to use the raw data 11

3 Data Management Process 12

3.1 Licensing 12

3.1.1 Licensing of non-open data 12

3.1.2 Licensing of open data 13

3.2 Anonymisation 13

3.3 End of project procedures 13

4 Data consistency and quality 14

5 GDPR 15

6 Legal and Ethical Processes 15

1 Introduction to ParCos Data

ParCos is concerned with the participatory communication of science. It is founded on the principles that in addition to making science stories more easy and fun to engage with, it is critical to maintain links to data, as well as other sources of evidence, on which stories are based. This allows those who are engaging with the story to go and look at this evidence and interpret for themselves. However, in order to achieve this aim it, is necessary that the data is made available in ways that support re-use.

One of the goals for ParCos is therefore to consider what principles should guide the curation of data for reuse (**ParCos curator**) and subsequently how this can be presented to a user so that they can explore it for themselves (**ParCos Data Explorer**) and how these activities can be embedded within different types of science story (**ParCos storyteller**). A further aim is to support reinterpretation and creation of new stories based on (or inspired by) existing stories and evidence but integrating new sources of data for example for the purpose of making a story more meaningful to those who are retelling it.

These components will be delivered within a single platform (**ParCos platform**), and they will all be developed via co-design methods within three case studies, in UK, Belgium and Finland through which the methods will also be tested.

In the above scenarios, ParCos will need to manage data from the following sources, each of which may require different handling procedures:

1. Data which is utilised within the ParCos tools and used by end-users, which could be:
    a. Research data from resulting from science activities:
        i. Owned/procured by Parcos project team from activities external to the project
        ii. Collected by end-user researchers as part of activities they undertake with the ParCos tools
    b. Data obtained via open data portals or other sites
2. Data which is collected as part of evaluation of the tools

The data may also vary in the extent to which it is:

1. Already open: such as existing open data/open research data)
2. To become open: such as data collected by end-user researchers and collected as part of evaluation
3. Restricted: to be used only under certain licensing conditions, for example research data made available by external research for project purposes may fall under this category.

This Data Management plan describes the ParCos project practices in relation to these different types of data.

The data will be in a variety of formats, some of which will only be known when data is identified. Common formats include

Standard formats, e.g.: .txt, .csv .json, .xml

GIS formats, e.g.: .shp, .shx, dbf

Media formats, e.g.: .wav, .mov

Data may be converted to different formats for use by NLP and other statistical methods, using software such as WEKA (.arff), SPSS (.sav), R (.rdata), nVivo, python and others.

2 Open Research Data Pilot

ParCos project is participating in the Open Research Data Pilot (ORDP). The main deliverable related to the ORDP is the Open ParCos Project data deliverable (D5.7) that will be delivered in M34 of the project. As part of the ORDP ParCos project has promised to adhere to the following conditions in relation to all data (and metadata) that is needed to validate the results in scientific publications and other curated and/or raw data (and metadata) that is mentioned within this DMP:

1. Develop (and keep up-to-date) a Data Management Plan (DMP) (this document)
2. Deposit data in a research data repository.
3. Ensure third parties can freely access, mine, exploit, reproduce and disseminate your data.
4. Provide related information and identify (or provide) the tools needed to use the raw data to validate your research.

Given that one main goal of ParCos is to make science data more reusable, the ability to adhere to these principles is of particular importance.

2.1 Updating of Data Management Plan, based on ParCos Activities

The plan will be updated throughout the project to reflect improved clarity on the following issues:

1) The types of data that the ParCos project both uses and generates. This will be in the form of a data diary (described in more detail in the later section on data consistency and quality) added as an appendix to this plan, which will describe all datasets but with confidential information removed.
2) The controlled vocabulary used for creating metadata

2.2 Research Data Repository

The main public research data repository will be Zenodo, the European repository provided by OpenAIRE and CERN. Zenodo supports Digital Object Identifiers (DOIs) that make content easily and uniquely citable, and supports exporting data in the Dublin Core format, therefore supporting the metadata format selected for the project. Furthermore, Zenodo allows publishing documents, data, and software, and linking all these publications together with metadata. Zenodo has also been self-audited to be compliant with FAIR principles in regard to metadata and can export to Dublin Core, the selected metadata format [1].

As above, data will be deposited in Zenodo open data repository under the Creative Commons 4.0 BY license. However, related to licensing restrictions and use case requirements, there

[1] https://about.zenodo.org/principles/

can be alternative, open community repositories. Zenodo takes a static view for data: It is uploaded and published by one person only and afterwards it is available forever as-is, immutable.

When working with or contributing to shared community data, such as civic commons (Balestrini et al., 2017) where people are empowered as actors instead of data providers (Palacin-Silva et al., 2018), we will use shared databases and suitable licenses that enable community ownership of data. Licenses that enable collaborative access, sharing and use of data openly among individuals and organizations include the Community Data License Agreement (CDLA)[2]. This approach enables shared ownership and shared work on the data. Collaborative work on open science data is used for example in citizen science approaches (Alamoudi et al., 2020; Luther et al., 2009; Mazumar et al., 2018; Schade et al., 2016).

Data will be deposited after appropriate anonymisation procedures have been performed (see later section) and at an appropriate time to ensure that data is available to support claims made in publications whilst ensuring ‘protection of scientific information, commercialisation and IPR, privacy concerns, security as well as data management and preservation questions.’

Prior to this data may be locally stored in the eDuuni platform[3], which is maintained by CSC – the Finnish IT Center for Science. The CSC Eduuni workspace meets the Finnish government Vahti 2/2010[4] information security regulation criteria for heightened level security. Meeting these criteria enables the workspace to store and process secure research information that is defined in Finnish regulations as level III (confidential) and in EU as “EU CONFIDENTIAL.” Using eDuuni allows respecting research participants’ privacy and rights before anonymization and other necessary processing for open data publishing.

2.3 ENSURE THIRD PARTY USE OF DATA

In order to ensure third part use of data, the project will adhere to the FAIR principles (findable, accessible, interoperable and reusable), as specified by the H2020 Programme Guidelines on FAIR Data Management[5]. These are described in turn as follows.

2.3.1 Findable

Making data easily findable is a critical first step to ensure that it can be used. Findable in this context means primarily that it can be easily located by search, even if you do not know

[2] https://cdla.io/

[3] https://info.eduuni.fi/what-is-eduuni/

[4] https://www.vahtiohje.fi/web/guest/2/2010-ohje-tietoturvallisuudesta-valtionhallinnossa-annetun-asetuksen-taytantoonpanosta

[5] European Commission. (2016). Guidelines on FAIR Data Management in Horizon 2020. Accessed from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

exactly where to look for it, in other words making it easy not just for humans but also for computers to locate.

In ParCos we will use machine-readable metadata with ParCos data to facilitate the discovery of datasets.

ParCos will use the Dublin Core MetaData Initiative (DCMI) as the metadata format along with a controlled vocabulary [6] that is specified during the ParCos platform requirements specification (D6.1). DCMI satisfies the ISO Standard 15836:2009 and the ANSI/NISO Standard Z39.85-2007 on cross-domain resource description.

2.3.2 Accessible

The default stance in the project that the research data will be made openly available, using licenses specified in Section 3.1, after anonymization and making sure that the participants’ rights are respected.

Data will be made available through publishing in the Zenodo research repository by default and using metadata specified above. When working with communities that adhere e.g. to civic commons principles, data will be also contributed to shared databases where possible.

Required software tools and other documentation will be made available with the data.

2.3.3 Interoperable

Published data will be made interoperable where possible and machine readable, using standard software formats that can be opened with open source software applications. A controlled vocabulary will be provided along with the DCMI metadata to assist in interoperability.

2.3.4 Reusable

Data will be licensed using the permissive Creative Commons 4.0 BY license to allow as wide re-use as possible. In the case of dynamic, open databases, the Community Data License Agreement (CDLA) will be used. In cases where the researchers are working with communities, an effort will be made to publish the data using a license that is as open as possible and compatible with the community’s existing license.

Data curation principles - especially ones identified through ParCos activities and creation of the ParCos Data Curator - and use of metadata, as well as licensing that facilitates re-use and access to tools, which is directly related to the final condition of ORDP, open source tools.

2.3.4.1 Open Source Tools

To support reuse project tools (ParCos Data Curator, Explore, Storyteller and Platform) will be made available as open source, so that both the research data and research software will be

[6] https://dublincore.org/specifications/dublin-core/dcmi-terms/

openly available. Creative Commons license recommends against using it for software source code, another open permissive license will be used: The MIT License[7].

2.4 Provide information and tools needed to use the raw data

As part of achieving this aim, the project tools will be made available as open source and supported by openly licensed documents and tutorials.

[7] https://opensource.org/licenses/MIT

3 DATA MANAGEMENT PROCESS

A data management process has been devised to ensure that all data that comes into the project is handled appropriately. The processes ensure that all project data will a) licensing, b) anonymisation, c) *representation, d) storage*, and e) *end of project* (see Fig. 1). Since processes related to both FAIR principles and data storage are discussed elsewhere, this section mainly details procedures related to licensing, anonymization, and end of project.

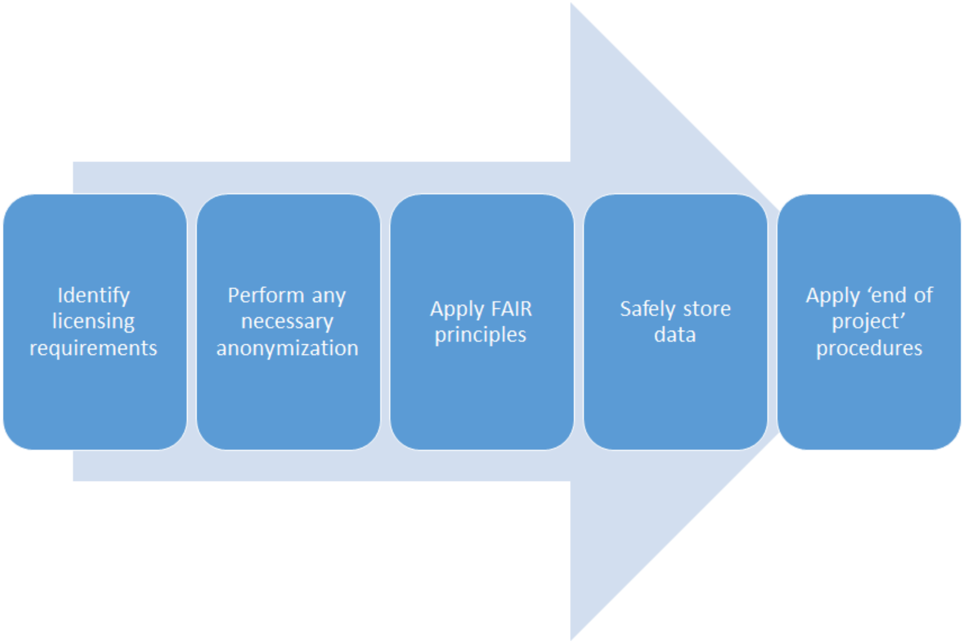
Identify licensing requirements
Perform any necessary anonymization
Apply FAIR principles
Safely store data
Apply ‘end of project’ procedures

Figure 1. A simple schematic of the Data Management Process

3.1 LICENSING

3.1.1 Licensing of non-open data

In accordance with the ORDP, the project will favour the opening of datasets created as a result of project activities in order to ensure replicability of the results. However, some of the datasets used within the project by end-users within case studies may come with licensing requirements. Whilst the ParCos approach has a vision of creating a process where it is possible to trace and use data ‘down the line’ and while the project will try to negotiate this and use open data to the largest extent possible, it could still be that based on old practices some data is made available with licensing restrictions that restrict further use. This may have implications for where/how data will be storied. In these cases, when data ia procured a data use agreement will be negotiated and agreed by both parties. This agreement will specify the terms of use relating to:

1. The licensing terms of the data and how the data can be used
2. Requirements related to data storage and agreement on where and how data will be stored

3. Duration of the agreement
4. What will happen to the data once the agreed period of use has ended, e.g. deletion from server

3.1.2 Licensing of open data

Open data and artefacts will be licensed using one of the three licenses listed below, depending on the data type and use purpose.

1) The default license is Creative Commons 4.0 BY [8]. This will be used in static, finished data that is published in Zenodo or other open data repository for open use.

2) If ParCos hosts a database or an API for open data, the CDLA 1.0 Permissive[9] will be used. CDLA includes additional considerations for databases that actively change, process data during its existence, and are used by or contributed to by several entities.

3) Open source software will be published under the MIT License[10], since Creative Commons license maintainer recommends against the use of Creative Commons for open source code.

3.2 ANONYMISATION

All personal information, as defined by GDPR and as interpreted by Data Protection Officer and a relevant Research Ethics Committee, will be removed from the data before open publication. If data cannot be anonymized, such as qualitative data or media files, the data will be kept private and only anonymized summaries will be published, such as anonymised transcript instead of audio recording or a description of a video. Alternatively if the research participant grants an explicit permission and there is a valid reason, as defined by GDPR, the data can still be deposited.

3.3 END OF PROJECT PROCEDURES

The majority of data will remain open beyond the end of the project and no change will need to be made. Where project restricted data sets have been used, data will be deleted based on the conditions stipulated in the agreed terms of use.

[8] https://creativecommons.org/licenses/by/4.0/
[9] https://cdla.io/permissive-1-0/
[10] https://opensource.org/licenses/MIT

4 DATA CONSISTENCY AND QUALITY

Data quality will be supported through the activities already described in this document, relating to the use of consistent metadata, the use of protocols for collecting data during experimentation (e.g. through survey, interview, observation). Consistency and quality of data will be controlled by maintaining a **data diary of project data**, both data used within the project and data collected during project activities. The data diary will indicate and maintain a list of preferred formats although others may be used and added as needed. The data diary will describe, for each dataset:

1. Owner
2. Format
3. Conversions applied
4. Metadata
5. License type
6. Storage location
7. Deposit date
8. Removal date
9. Version history

The data diary activities are transversal to the Data Management Process of Fig. 1. Each process generates new information to be recorded.

5 GDPR

According to European General Data Protection Regulation (GDPR), data collected from people in the EU is subject to certain rights and anyone collecting or using data within the EU must adhere to the GDPR regulations. Working within the GDPR regulations, ParCos will not collect or utilize data without first receiving the informed consent of participants. The process for obtaining informed consent is described in D8.2 and data protection in more detail in D8.3.

Once data has been collected, the data subject has the right to request access to the data that is held about them, whether that data was provided directly by the subject or data derived through indirect means such as via observation, unless the data has been fully anonymised. That is, rendered anonymous in such a way that the data subject is not, or is no longer, identifiable. When data is requested, it should be provided in a common format so that the data subject can reliably access it. The data subject also has the right to have their data rectified or erased. This may happen if, for example, the participant wishes to withdraw consent for participation in the project after some data collection has already taken place.

Information on data rights will be made available to all participants as part of the informed consent procedure, via a separate document. Participants will be made aware of their right to withdraw consent and it will be explained what happens to their data afterwards.

When online services and other tools that collect data are created, the implementation is first self-evaluated with checklists[11] and then evaluated by the legal and ethics committee. When necessary, external review is sought from the project DPO and the appropriate organization’s research ethics committee.

6 Legal and Ethical Processes

The ethical processes of the project are described in detail D8.1, 8.2 and 8.3. These documents outline the procedures for gaining ethical approval for project activities and collection of data, the use of informed consent and the anonymisation procedures that will be in place and the roles of a) the legal and ethical committee, to define and oversee adherence to ethical aspects. This is a project level responsibility of a) Annika Wolff (LUT) and Bieke Zaman (KUL), and b) the data protection officer, who is a legal representative from LUT.

[11] For example, using the self-evaluation checklist by the GDPR.EU project, https://gdpr.eu/checklist/

REFERENCES

Balestrini, M., Rogers, Y., Hassan, C., Creus, J., King, M., & Marshall, P. (2017). A city in common: a framework to orchestrate large-scale citizen engagement around urban issues. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2282-2294).

Luther, K., Counts, S., Stecher, K. B., Hoff, A., & Johns, P. (2009). Pathfinder: an online collaboration environment for citizen scientists. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 239-248).

Mazumdar, S., Ceccaroni, L., Piera, J., Hölker, F., Berre, A., Arlinghaus, R., & Bowser, A. (2018). Citizen science technologies and new opportunities for participation. UCL Press.

Palacin-Silva, M., & Porras, J. (2018). Shut up and take my environmental data! A study on ICT enabled citizen science practices, participation approaches and challenges. *EPiC Series in Computing*, *52*, 270-288.

Schade, S., & Tsinaraki, C. (2016). Survey report: data management in Citizen Science projects. *Publication Office of the European Union: Luxembourg*.