

DCitizens Fostering Digital Civics Research and Innovation in Lisbon

DELIVERABLE 7.2: Data Management Plan

DCitizens has received funding from the European Union's Horizon Europe Framework Programme, project call HORIZON-WIDERA-2021-ACCESS-03, grant agreement 101079116



D7.2: Data Management Plan

Project Information

Grant Agreement	101079116
Title	Fostering Digital Civics Research and Innovation in Lisbon
Acronym	DCitizens
Funding Scheme	Twinning
Start date	01/12/2022
Duration	36 months
Call	HORIZON-WIDERA-2021-ACCESS-03
Website	https://dcitizens.eu/

Deliverable Information

ID	D7.2
Title	Data management plan
WP	7
WP Leader	IIT
Contributing Partners	IIT, IST-ID
Nature	R: Document
Authors	Valentina Pasquale (IIT), Alessio Del Bue (IIT), Matteo Taiana (IIT)
Contributors	Hugo Nicolau
Reviewers	Dina Dionísio
Deadline	M6 (31/05/2023)

Dissemination Level

PU	Public	<input checked="" type="checkbox"/>
PP	Restricted to other programme participants	<input type="checkbox"/>
RE	Restricted to a group specified by the consortium	<input type="checkbox"/>
CO	Confidential, only for members of the consortium	<input type="checkbox"/>

Document Log

Version	Date	Author	Description of Change
1	22/05/2023	Alessio Del Bue	First release
2	26/05/2023	Alessio Del Bue	Final Draft

Disclaimer

The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Executive Summary

This deliverable presents the first version of the Data Management Plan (DMP) for the DCitizens project and provides an analysis of the data management policy applied by the Partners to datasets generated within the Project. In particular, the DMP identifies the main datasets and describes research data management during the project, as well as how and what parts of the datasets will be openly shared, will be made accessible for verification and re-use, will be curated and preserved. The goal of this DMP is to facilitate effective internal data management and make data FAIR (Findable, Accessible, Interoperable, and Reusable). This document provides guidance to project partners on data management, and it is a useful tool to agree on data processing, facilitate the creation of a common understanding and, where possible, common practices.

Table of Contents

1	Introduction	1
2	Data summary.....	1
2.1	DCITIZENS datasets.....	1
2.2	Extended DCITIZENS datasets' information.....	3
3	FAIR data	9
3.1	Making data findable, including provisions for metadata	9
3.2	Making data accessible.....	10
3.3	Making data interoperable	11
3.4	Increase data re-use.....	11
4	Other research outputs	12
5	Allocation of resources.....	12
6	Data Security.....	13
7	Legal and Ethics	13
7.1	Protection of personal/special categories of data	13
7.2	Ethical issues (research data involving experiments with humans)	13

1 Introduction

This deliverable is submitted to the European Commission at M6 (31/05/2023) and represents a preliminary Data Management Plan (DMP) for the DCITIZENS project. The DMP is, in fact, a living document and it will be updated and further refined with the project's progress. It is also important to remark that this DMP reflects the provisions established by the project contracts and complements the planned project exploitation, dissemination and IPR procedures.

The document's first development and future updates mainly rely on the collection of information about datasets filled out by each partner responsible for producing such data (see Section 3.2). The form used to collect this information (i.e., Dataset Questionnaire Form, DQF) has been prepared and is constantly updated by IIT Offices involved in this process (RDM Office, Projects Office, Legal and ICT Directorates).

The DMP deliverable, including an editable copy and the DQFs, will be available to all partners via a sharable institutional OneDrive folder.

Formal revisions of the DMP will be submitted to the European Commission every 6 months if needed, starting from the initial submission at M6. Thus, the document will be updated as appropriate along the project duration. The different versions will be numbered and dated for identification. Official versions will be stored on a OneDrive folder accessible by all partners. Should a new dataset be identified during the project implementation, partners will submit a new form containing the newly identified dataset and will notify the coordinator. IIT will be in charge of updating the document and its annexes and notifying the Consortium through the project mailing list system.

2 Data summary

2.1 DCITIZENS datasets

The first version of the DCITIZENS DMP is based on the description of 2 datasets whose key details are summarized in the following table (Table 1).

Table 1. Summary of DCITIZENS datasets.

DATASET NAME	PARTNER(S)	ORIGIN	TYPE	FORMAT	SIZE	WP&TASK	ACCESSIBILITY ¹
DCitizens_KG	IIT	Both Existing and Generated	textual data/image/audio/video	.json, .txt, .jpeg, .MP4.	100GB	WP4: T4.3, T4.4	Open access
DCitizens_Community_led_Design_Data	IST-ID	Generated	Textual data	.txt	1 GB	WP4 (T4.1-T4.3)	Open access

¹ Legend: Closed access, Restricted access, Open Access

Most identified datasets are expected to have long-term value. Datasets shall be useful to several categories of research communities, and users from research, industry, and society, including:

- Research and scientific community: digital civics, human-computer interaction, social computing, computer science, ubiquitous computing, design, social innovation, and behavioural sciences.
 - Industry: Creative Industries and cultural sector.
 - Society: Local stakeholders active in education, digital transformation, culture and creativity, and climate action.
 - Datasets may have several re-uses such as novel similar research studies.
- Detailed expected utility for each dataset to be generated is reported in the tables of next section.

2.2 Extended DCITIZENS datasets' information

DATA SUMMARY		DATASET 1	DATASET 2
Main dataset (name)	DCitizens_KG		DCitizens_Community-led_Design_Data
Sub-dataset (name)	-		-
Responsible partner	IIT		IST-ID
Other partners involved	ITI		IIT, UNN, USi
Goal	Support the pilot activities with data related to the activities stored and linked in a structure format as in a Knowledge Graph.		Data collection is related to WP4 (T4.1-T4.3), which aims to support the collaboration with a diverse set of stakeholders and civic organisations to co-design digital technologies. The main goal is to understand community needs, challenges, goals, expectations, and local assets. The data will allow the consortium partners to co-design and develop novel digital technologies that better fit the socio-technical context of the local communities.
Data origin	Existing in some form in Knowledge bases and generated within the project from partners data entries.		Generated within the project
	<i>Please specify the source(s) of previously existing data</i>		<i>Please justify below the need of new data to be generated</i>
	Previous Knowledge Graph (KG) from MEMEX project: https://github.com/MEMEXProject/MEMEX-KG		Data does not currently exist and will be collected in cooperation with Lisbon's communities in a hyper-localised context.
	<i>Please justify below the need of new data to be generated</i>		
	To customize the KG given the custom pilots in DCITIZENS		
Data collection	Using digital tools on a smartphone App		Data will be collected using an ethnographic methodology and methods, such as field observations, focus groups, individual semi-structured interviews, design workshops, and design probes. The methodology is inherently longitudinal and open-ended, aiming to understand how communities function and why they function in a certain way. Raw data will be transcribed into textual format and data analysis will be mostly qualitative, using a reflexive thematic analysis approach. Reflexive thematic analysis aims to analyse qualitative data (e.g., text, images, video) to answer research

		questions about people's experiences, views and perceptions, and representations of a given phenomenon.
Dataset type	Other	Experimental data
	<i>Please specify here below the type of data</i>	
File formats	textual data/image/audio/video	We will use an interoperable data format for most qualitative data, i.e., plain text data, ASCII (.TXT).
Expected volume of data	100 GB	1 GB
Expected time of release	End of project, possibly an embargo period of maximum 6 months after the end of the project would be discussed among partners for finalizing project publications.	The dataset will be completed and released at the end of the project.
DOCUMENTATION AND DATA QUALITY		
Metadata and documentation	A github page will describe the type of data and how to access the repositories. Metadata will be aligned to WikiData format.	<p>We will organise data in a folder structure. Each main folder will correspond to a specific community and sub-folders will be divided by data collection method (e.g. interview, observation, workshop).</p> <p>For the main folders we will use the following name convention: YYYY-COMMUNITY, where YYYY is the year where we started collecting data, and COMMUNITY is the name of the community we will be working with.</p> <p>For sub-folders we will use the following name convention: YYYY-MM-METHOD, where YYYY-MM is the year and month where we started collecting data, and METHOD corresponds to the data collection method.</p> <p>Each .TXT file will use the following name convention: YYYY-MM-DD-NAME, where YYYY-MM-DD is the year, month, and day we collected the data, and NAME is a representative name of the subject (e.g., participant ID, group ID).</p> <p>Inside each folder we will create two additional .TXT files: readme.txt, and protocol.txt. The readme.txt will detail the folder, sub-folder, and file structure as well as the contents of each file. The protocol.txt will describe the data collection methodology (how different methods were used) or data collection</p>

		protocol (for each research method) for main folders and sub-folders, respectively.
Keywords	Knowledge Graph, Pilot Data	We derived keywords from ACM's Computing Classification System: Human-Computer Interaction, Human-centered Computing, Empirical Studies, Collaborative and Social Computing.
Data quality		All textual data will reviewed by a human to ensure the accuracy of transcriptions.
STORAGE AND BACKUP DURING THE RESEARCH PROCESS		
Storage and backup solutions	IIT storage servers, Zenodo after project ends	Data will be stored in a private, password protected, University of Lisbon cloud service. These services offer automatic backup and recovery options. Store options are up to 1 TB.
Data security and protection	Ensured by internal ICT department according to the institutional Information Security policy	Access to storage server is password protected and only accessible to the research team via individual credentials. All data stored in the private cloud service will be pseudonymised. When applicable, the mapping between participant personal information and ID will be stored in a paper sheet and stored in a secure locker only accessible to the PI.
LEGAL AND ETHICAL REQUIREMENTS, CODES OF CONDUCT		
Personal/special categories of data	NO	YES
Protection of personal/special categories of data	There are no particular special categories of data	<p>We will submit and seek approval of all research activities involving human participants to the Instituto Superior Técnico's Ethics Commission. Moreover, all activities will be submitted to the University Data Protection Officer for guidance on data management.</p> <p>All data will be pseudonymised. The mapping between participant personal information and ID will be stored in a paper sheet and stored in a secure locker only accessible to the PI. Access to storage server is password protected and only accessible to the research team via individual credentials.</p>

		In accordance with Portuguese data protection laws, any personal data will be deleted after 5 years. Additionally, participants may request access and/or deletion of their data at any time.
Ethical issues issues (research data involving experiments with humans)	NO	YES
	NA	<p>We will submit and seek approval of all research activities involving human participants to the Instituto Superior Técnico's Ethics Commission. Personal information consists of names and e-mail addresses of participants collected during the recruitment process. We will inform the University Data Protection Officer (DPO, dpo@ist-id.pt) of all research activities and follow their instructions to guarantee best practices in collecting and processing personal information. Moreover, we will disclose the DPO contact details to the research participants in the informed consent.</p> <p>Research data will be stored in the University private cloud services, password-protected with encrypted communication protocols, and only accessible to the research team. All data will be pseudonymized. We will not associate data from any particular participant in a personally identifiable way (name or e-mail). Each research participant will be assigned an arbitrary (anonymous) code number, and this code number will be used when collecting data. All paperwork with identifiers and matching code numbers will be kept in lockers available only for the research team. The paperwork will be destroyed after 5 years or at participants' request. No written documentation other than the Informed Consent will contain participants' identifiers. Participants' names and email addresses will not be associated with any collected data from the study.</p> <p>Informed consent: we will collect informed consent from participants in all data collection activities (e.g., interviews, observations, co-design sessions) and user studies. Inform consent will inform participants about: who is the research team; the nature and purpose of the project; study procedures; potential risks and discomforts; potential benefits; reimbursements;</p>

		<p>alternatives to participation; how to terminate participation; data confidentiality procedures; IRB, DPO, and PI contact information for further questions. Participants may request access and/or deletion of their data at any time by sending an email to the PI or DPO.</p> <p>Details of recruitment: we will collaborate with civic organisations to recruit participants. We will send flyers to be distributed by the organisations' collaborators to fill in an online survey, which will also serve as a screener. After the survey screener, the volunteers who meet the above eligibility criteria will be invited by e-mail to participate in the study. All remaining participants' data will be deleted.</p> <p>Informed consent procedures: informed consent and information sheets will be sent out to volunteers via e-mail when invited to participate in the study. They will be instructed to read the informed consent and sign it.</p> <p>Potential risks to participants: minimal risk. The research activities will not have an adverse effect on the participant's rights and welfare.</p>
Other ethical issues (e.g. involving animal subjects)	NO	NO
	NA	NA
Intellectual property rights		No IPR issues to report.
DATA SHARING AND LONG-TERM PRESERVATION		
Data sharing	Open access	We will share the data by depositing it in a long-term trustworthy data repository. The depository will provide a persistent identifier (i.e., DOI) so that data can be reliably and efficiently located and referred to. Data will be preserved for as long as the repository guarantees.

		Data may be reused in the future to conduct additional qualitative analysis under multiple different perspectives/lenses (e.g., cross-cultural comparisons, longitudinal analysis).
Data repository	Zenodo or IIT Dataverse	Zenodo or IIT Dataverse
Restrictions on sharing	Possibly an embargo of maximum 6 months after the end of the project for writing scientific publications	The dataset will be released at the time of publication or otherwise at the end of the project.
Data curation	All data will be retained	All pseudonymised data will be preserved for as long as the repository guarantees. All paperwork with identifiers and matching code numbers will be destroyed after 5 years or at participants' request.
Requirements for reusability	Software for access will be open source on a github platform	All data will be stored in an interoperable format, i.e. .TXT; thus, no special software is necessary to access and use the data.
Licensing	CC BY - Attribution	CC BY - Attribution
DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES		
Roles and responsibilities	IIT	The coordinating institution (IST-ID), namely its PI - Hugo Nicolau - will be responsible for data collection and management, including metadata production, data quality, storage and backup, data archiving, and data sharing. Consortium partners will be granted access to research data by IST-ID. Overall, IST-ID will be responsible for implementing the DMP with guidance from IIT. IIT will ensure the DMP is reviewed/updated every 6 months, if needed.
Resourcing	Sustained by project	IIT dedicated 1 person-month to monitor, review, and request updates to the DMP every six months. Other partners dedicated 0,5 person-month to review the DMP for the workpackages they are leading every six months. The coordinator (IST-ID) dedicated 3 person-month to implement the DMP. Staff time has been costed to implement the DMP. Data storage and backup services are guaranteed by the host institution.

3 FAIR data

3.1 Making data findable, including provisions for metadata

In order to make each public dataset findable and citable, a Digital Object Identifier (DOI) will be assigned to each dataset. Final datasets in DCITIZENS will be uploaded and publicly shared through ZENODO or other institutional repository (IIT DATAVERSE for datasets generated at IIT), which provide DOIs to all publicly available uploads. The DOI of each dataset will be added to the datasets' tables reported in Section 2.2 in future updates of the document.

To facilitate datasets' findability and reusability, file naming conventions agreed among partners will be used and clearly explained in associated "readme.txt" files. File naming conventions have been detailed in Section 2.2.

Meaningful search keywords will be provided in the datasets' metadata to optimize the possibility for discovery and potential re-use. Search keywords will be chosen according to a standard nomenclature or vocabulary, namely ACM's Computing Classification System.

Zenodo is compliant with the EU definition of "trusted repository". As regards metadata, Zenodo allows to set extensive citation metadata, including DOIs, authors, contributors, keywords, funding, related or alternate identifiers, and references to scientific articles or other type of publication. Zenodo is indexed in OpenAIRE Explore and registered in re3data.org and FAIRsharing.org. Zenodo is compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, <http://www.openarchives.org/pmh/>), a widely used protocol for harvesting metadata. Following OAI-PMH, the available standard formats for metadata harvesting in Zenodo are DataCite and DublinCore.

IIT Dataverse (<https://dataverse.iit.it/>) is the institutional research data repository of the Istituto Italiano di Tecnologia, for both preservation and sharing of research datasets. IIT Dataverse is compliant with the EU definition of "trusted repository". It is based on the Dataverse software, developed at Harvard University (www.dataverse.org). Dataverse assigns persistent DOIs to all data uploads for findability and it is accessible through standard HTTPS protocol. IIT Dataverse is indexed in OpenAIRE Explore and registered in re3data.org. Moreover, it provides APIs to search and access datasets, including a SWORD API. Each data upload includes 1) citation metadata, 2) optional and customizable domain-specific metadata (e.g., for Life Sciences), and 3) file-level metadata. Metadata can be exported in different standard formats (DataCite, OpenAIRE, JSON, JSON-LD, OAI, etc.) for maximal interoperability. Dataverse ensures reusability of datasets, by supporting open licenses, like Creative Commons licenses, and offers the possibility to customize specific data usage agreements. IIT Research Data Management service oversees

dataset publication, by providing basic data curation to ensure dataset quality and FAIRness.

3.2 Making data accessible

The project complies with the Open Science and Research Data Management requirements about openness and accessibility of research data, metadata, and other outputs resulting from HE grants, as detailed in the DCITIZENS Grant Agreement (art. 17) and described in the HE Annotated Model Grant Agreement (Annex 5), and HE Programme Guide. Therefore, research data generated during the project, including raw and processed data as described above in Section 2.2, will be deposited in trusted repositories and made open, with the exception of datasets that support unfinished peer-reviewed publications, patent applications, or information that cannot legally be made openly accessible (e.g., personal or sensitive data, following “as open as possible, as closed as necessary”). The last column of Table 1 summarizes the accessibility level foreseen for each produced dataset. Datasets that support peer-reviewed scientific articles will be made open at the time of publication with no exceptions. Possibly an embargo of maximum 6 months after the end of the project would be applied to ensure the possibility of including data in scientific publications.

Along the project’s duration, whenever possible project’s documents and research data will be shared within the consortium through an institutional shared OneDrive folder, requiring user authentication to keep confidentiality of data until required.

Appropriate and comprehensive documentation (e.g., extensive, and complete *readme.txt* files), together with relevant metadata, will be prepared and attached to the data before sharing.

Software codes needed to access / read / visualize the data will be made openly accessible through dedicated GitHub / GitLab repositories, which provide open access and long-term storage of source codes.

Noteworthy, GitHub repositories can be linked in Zenodo and code releases can be assigned DOIs for findability and greater reproducibility.

All metadata in Zenodo is licensed under Creative Commons Zero (<http://about.zenodo.org/terms>), while the data files may be either open access or subject to a license described in the metadata. Zenodo metadata will contain references to related materials and tools (e.g., codes) by citing and linking DOIs. All data stored in Zenodo will remain accessible for the lifetime of the repository, which is currently warranted for a minimum of 20 years. Metadata will remain available also after data is no longer available.

By default, data and metadata in Dataverse are licensed under Creative Commons Zero (<https://about.zenodo.org/terms/>). Specific licenses for data will be clearly specified in the metadata (e.g., CC-BY instead of CC0). Dataverse metadata will contain references to related materials and tools (e.g., codes) by citing and linking

DOIs. Research data stored in IIT Dataverse will remain accessible online after the end of the project with no specific deadline and until it is required. Metadata will remain available also after data is no longer available or transferred to an offline storage location (e.g., tapes).

3.3 Making data interoperable

To facilitate exchange and re-use of data, datasets will be exported and stored in formats that are commonly accessible (e.g. .txt, .csv), including the associated metadata and additional comments and descriptive text to aid the interpretation of the data.

To ensure the interoperability of data and metadata, standard or community-endorsed vocabularies (e.g., metadata structure of Wikidata) will be used when applicable. In case new vocabularies are generated or uncommon vocabularies are used, mappings between custom and community-endorsed vocabularies will be associated to the dataset for interdisciplinary interoperability. Abbreviations, codes, and variables' names will always be clarified at first use or defined in "readme.txt" files.

3.4 Increase data re-use

As mandated in art. 17 of the Grant Agreement, the digital (or physical) access to the results needed to validate the conclusions of scientific publications will be provided, including access to all the information about the research outputs/tools/instruments needed to validate publications and enable the re-use of data. The access to research data will be provided through deposition in trusted repositories, as detailed in the previous paragraphs. Whenever possible, research datasets will be available under the **Creative Commons Attribution 4.0 International (CC BY 4.0)** license, allowing third parties to share and adapt data with no restrictions as long as attribution (i.e., citation) is provided. All the Consortium partners will ensure that proper licenses are attached to the deposited data to define all conditions under which the work is provided and can be reused, also in case data are not released under an open license.

Along with data any information about methods, protocols, models, software, algorithms, workflows, simulations, electronic lab notebooks' records, etc. will be also provided and linked in the repository through the use of persistent identifiers (e.g., DOIs) (see also "Other research outputs"). For instance, source codes will be version-controlled and deposited in GitHub / GitLab and made available in open access with suitable licenses (GNU GPL, MIT, etc.). Whenever useful, information about data cleaning, data quality assurance procedures, methodology, as well as variables' definitions, units of measurements, software dependencies, and in general data structure will be included in "readme.txt" or "readme.md" files, deposited together with data.

4 Other research outputs

In Table 2, a list of other research outputs generated or re-used within the project is reported.

Table 2. Summary of other DCITIZENS research outputs.

OUTPUT NAME	PARTNER(S)	ORIGIN	TYPE	ACCESS TYPE	WP&TASK	ACCESSIBILITY ²
MEMEX digital App	IIT	Existing	code and software	Digital	WP4: T4.3/ T4.4	Restricted Access

The consortium will deliver open-source software by making the full source code available with appropriate licences (e.g., GNU GPL/LGPL, MIT License). Collaboration and versioning will be managed via GitHub/GitLab: to improve reproducibility, a DOI will be assigned to code releases via GitHub/Zenodo integration whenever useful or required. Appropriate documentation will also be included with key information such as system requirements, dependencies, installation instructions, description of how to run and use the code (inputs, outputs, parameters, etc.), software citation, tutorials, and/or API documentation. Particular attention will be devoted to follow recommendations for the development of FAIR research software, e.g., FAIR4RS principles and recommendations elaborated by the Research Data Alliance WG (<https://www.rd-alliance.org/group/fair-research-software-fair4rs-wg/outcomes/fair-principles-research-software-fair4rs>).

5 Allocation of resources

The coordinating institution (IST-ID), namely its PI - Hugo Nicolau - will be responsible for data collection and management, including metadata production, data quality, storage and backup, data archiving, and data sharing. Consortium partners will be granted access to research data by IST-ID. Overall, IST-ID will be responsible for implementing the DMP with guidance from IIT. IIT will ensure the DMP is reviewed/updated every 6 months, if needed.

IIT dedicated 1 person-month to monitor, review, and request updates to the DMP every six months. Other partners will dedicate 0,5 person-month to review the DMP for the workpackages they are leading every six months. The coordinator (IST-ID) dedicated 3 person-month to implement the DMP. Staff time has been costed to implement the DMP. Data storage and backup services are guaranteed by the host institution.

² Legend: Closed access, Restricted access, Open Access

The external repository chosen for long-term data sharing and preservation (Zenodo) offers free data archiving up to 50 GB per dataset, which is enough given the expected data volume.

6 Data Security

IIT: According to the information security risk level associated to datasets (i.e., low and/or medium), only IT assets approved for the corresponding risk level will be used, as detailed in the IIT Information Security policy. Laptops, workstations, and servers are managed by institutional ICT Service and all procedures in place for storage and backup comply with IIT Information security policy. Where necessary, authentication with institutional login will be required to protect data confidentiality and integrity.

IST-ID: Access to storage server is password protected and only accessible to the research team via individual credentials. All data stored in the private cloud service will be pseudonymised. When applicable, the mapping between participant personal information and ID will be stored in a paper sheet and stored in a secure locker only accessible to the PI.

7 Legal and Ethics

7.1 Protection of personal/special categories of data

Information about protection of personal data collected in Dataset 2 is reported in Section 2.2 (Extended DCITIZENS datasets' information).

7.2 Ethical issues (research data involving experiments with humans)

Information about ethical issues related to the collection of personal data in Dataset 2 is reported in Section 2.2 (Extended DCITIZENS datasets' information).