CS TRACK
Investigating Citizen Science

Horizon 2020 / Science with and for Society Programme
Grant agreement number: 872522

Data Management Plan
D7.1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 872522

Title of project	CS Track
Full title of project	Expanding our knowledge on Citizen Science through analytics and analysis
Title of this document	Data Management Plan
Number of this document	D7.1
Dissemination level	Public
Due date	31.5.20
Actual delivery	31.5.20
Versioning history	Several drafts during the last two months
Authors	Nils Malzahn (RIAS) and Raul Drachman (MOFET, responsible partner for this deliverable)
Executive summary	Our project is about data and information. Responding to the Call, we will expand our knowledge on Citizen Science (CS) and its impact. We will seek this increased knowledge by "observing" a large and diverse set of CS projects, gathering data from the web, questionnaires and interviews of involved players, etc., and from a more direct inspection of running activities. Vast amounts of data will thus be studied, relying for this on (1) web-based analytics, i.e. the use of computational analyses to study CS activities based on their manifestations on the web and social media, and (2) deepening and combining the obtained data and information through multi-perspective analysis and triangulation. Our data analytics and analysis will target both "own" aspects and developments of the CS projects (organizational/operational characteristics, scientific outcomes, good practices, individual/group learning, etc.), and societal aspects, related to the impact of those activities on society, such as gender, age, geographical and socioeconomic aspects; science role in society, changing attitudes to science, etc. All these reflect directly on questions pertaining to data and data management from all conceivable angles – collection, storage, safekeeping, preservation, accessibility, retrieval, ethical considerations, etc. –, which are addressed in this Plan. As not all the relevant questions have been solved (or even posed), this document has to be seen, at least for the next many months, as work in progress. We plan to update it on a current basis, issuing interim versions as important modifications will accumulate.

Table of contents

1 Data Summary

1.1 Purpose of the data collection / generation

Essentially, our project is about data and information. Responding to the Call, we will endeavor to expand our knowledge on Citizen Science (CS) and its impact. Overcoming present hurdles on the way to reach that knowledge will enable the potential benefits of CS – on individual citizens, organizations, and society at large – to be realized more effectively and frequently. This is the aim of our project, CS Track, which will seek this increased knowledge by "observing" a large and diverse set of CS projects, gathering data from the web, questionnaires and interviews of involved players, etc., and from a more direct inspection of running activities. Vast amounts of data will thus be studied, relying for this on (1) web-based analytics, i.e. the use of computational analyses to study CS activities based on their manifestations and traces on the web and social media, and (2) deepening and combining these analyses with approaches known from social studies through multi-perspective analysis and triangulation. Our data analytics and analysis will target both "own" aspects and developments of the CS projects (organizational / operational characteristics, scientific outcomes, good practices, individual/group learning, other success or failure indicators, etc.), and societal aspects, related to the impact of those activities on society, such as gender, age, geographical and socioeconomic aspects; science as a discipline and its role in society, changing attitudes to science, women in science, etc.

All the above reflect directly on questions pertaining to data and data management from all conceivable angles – collection, storage, safekeeping, preservation, accessibility, retrieval, ethical considerations, etc. –, which are addressed below in this Plan. To be sure, not all the relevant questions have been solved (and likely not all have been asked yet either). This document has to be seen, accordingly, as an evolving one, as work in progress, and will probably keep this character for most of the project's duration. We plan to update it on a current basis, delivering interim versions at different stages as important modifications will accumulate. In this sense, both the advance in the project work in general and, in particular, the specific needs arising in the context of PDP and other data-related ethical issues as our experience grows, are candidate sources of changes or refinements of this document in the future.

1.2 Relation of the data collection / generation to the objectives of the project

The main goal of the data collection in CS Track is to further our understanding of the ways CS activities can have an impact on society in a local and global perspective, how citizen scientists develop their scientific competences and knowledge and how they propagate these into society. This entails

1. exploring and characterizing the interplay (and possibly the overlap) of CS with official science by using computational methods of data mining and network analysis in combination with social research methods of summative and formative/participatory nature;
2. identifying distinguishing factors to characterize the specific types of discourse and approaches found in CS projects in terms of the inherent

knowledge-building strategies, targets of action and their relation to official science; and

3. translating the further knowledge and findings above into practical recommendations for actors at all relevant levels – policy makers, companies, NGOs, educational institutions – to raise the value of CS for science awareness and science literacy at all ages and for society in general.

1.3 Types and formats of data generated/collected

We will be using different formats for different types of data.

Data source	Type of data	Selected format	Justification of format
Questionnaires	**Tabular data with extensive metadata** variable labels, code labels, and defined missing values	Webropol / MS Excel (.xls / .xlsx) LimeSurveyl / MS Excel (.xls / .xlsx)	Format is widely known and established.
Twitter data	**Tabular data with extensive metadata** variable labels, code labels, and defined missing values	comma-separated values (.csv)	Easy to pass to scripts to perform analysis. Common format widely used as an input by other tools.
Google scholar - author network	**Tabular data with extensive metadata** variable labels, code labels, and defined missing values	comma-separated values (.csv)	Easy to pass to scripts to perform analysis. Common format widely used as an input by other tools.
Google scholar - institutions network	**Tabular data with extensive metadata** variable labels, code labels, and defined missing values	MPEG-1 Audio Layer 3 (.mp3) or Audio Interchange File Format (.aif) or Waveform Audio Format (.wav)	Easy to pass to scripts to perform analysis. Common format widely used as an input by other tools.
Web Scraping data **Data from Web extraction**	**NoSQL Database document collections** MongoDB	comma-separated values (.csv) Binary JSON (BSON)	Both formats are easy to pass to scripts to perform analysis. Common format widely used as an input by other tools.

Reports / Deliverables; **Transcripts / Excerpts;** **Survey documents** **Results from analysis processes**	**Documentation and scripts**	plain text (.txt) formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx)	The PDF/A format is widely used to archive documents. MS Office is widely used.
Images from results	**Image data**	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf) Scalable Vector format (.svg)	The project makes use of a variety of image formats as they are widely spread and depending on the generating tools (e.g. for networks). Since the images will mostly be used to explain phenomena to third parties, image formats that are accessible via wide-spread browsers are preferred.
Interview records	**Audio data**	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav) Free Lossless Audio Codec (FLAC) (.flac)	The interview data will be recorded for transcription. Thus, the format used needs to fit the requirements of the transcriptor. The raw interview records will most probably not be shared among other people than the interviewing party.
Statements, Demonstrators	**Video data**	MPEG-4 (.mp4)	Currently, the most wide-spread video data that can be converted to another format, if needed.

1.4 Re-use of existing data

Part of our data will be originated in previously prepared/published data – e.g., data of the CS projects distributed through their websites – and in this sense they will be "re-used", this time for the further analysis we will do on them. These are data that we will collect, gather. Some other data will be essentially produced by our teams, not re-used in the previous sense. These data will result from our interviews and questionnaires, observation of "the big picture", of trends, and their analysis including multi-disciplinary triangulation as well as network analysis, etc.

The project starts its research on existing data, principally from repositories of citizen science projects like Fablabs.io (https://www.fablabs.io/) and other citizen science activities. These indexes are used as an initial input for the project's database on Citizen Science. Data already provided to open research repositories are also taken into account. As an example, there are currently 56,221 research data sets listed in OpenAIRE (https://www.openaire.eu/) under the keyword "citizen science".

1.5 Data sources

The project seeks to increase knowledge about Citizen Science by "observing" a large set of CS activities (of all kinds and sizes) in different – alternative or complementary – ways.

CS Track's overall approach to data gathering and analysis will include the following elements:

- The different stakeholders involved in the CS Track activities (participants from different ages, guides, institutions etc.) will be part of the data gathering, either in a passive way (by means of the data automatically tracked about their products in the platforms) or actively (providing feedback e.g. via questionnaires or interviews). WP2 will document the activities, persons, user stories and processes involved in actions followed in order to perform the CS Track activities.
- To enable data triangulation, mixed data gathering and analysis methods will be applied. Quantitative analyses will be applied mainly to surveys (WP4) and the analytics of the data (WP3) tracked by web platforms. These will be based e.g. on existing surveys on CS and scientific literacy (Gormally et al., 2012) as well as from MoRRI indicators. Complementary, qualitative analyses (WP2) will be implemented to the content generated by the CS and the qualitative feedback gathered via surveys, interviews, videotaping CS activities or field notes.
- On the basis of these quantitative data, selected participants with different profiles or usage patterns will be selectively approached for a more qualitative assessment. This combination of unobtrusive data collection of system usage and the collection of targeted qualitative data will give a more precise and more informative picture of Citizen Science "activities".

1.6 Data size

Vast amounts of data, from all sources (see section 1.5), are collected and studied – through data analytics and multi-perspective analysis – converting them into information and knowledge that will shed new light on CS and on how its role in the

society and the economy could be optimized. Especially the automated web analytics algorithms will produce Gigabytes of data.

1.7 Data utility

The collected, aggregated, analyzed and derived data will be useful to all kinds of stakeholders. Without any particular ordering and with numerals just to facilitate reference from section 2.2.1 below, the following are the target users for our data:

1. Own project teams, for their work (including, as appropriate, dissemination activities and exploitation planning).
2. Policymakers in the field of CS, or in broader societal areas that may be only partially or less directly related to CS, such as education (and in particular, science education), informal learning, volunteer engagement, gender issues (women in science, women in society), youth (pursuing science careers), etc., will gain insights into Citizen Science, its value, its needs and limitations. We may include in this "policymakers" group also officers in organizations and agencies that fund or otherwise promote projects and other endeavors in the CS area.
3. CS entrepreneurs and managers/executives – actual and potential, individual and institutional (companies, NGOs, etc., and own staff of CS projects), in all areas of science and technology.
4. Citizen scientists that actively take part in CS activities, will get insights for their own activities, either by direct consumption of specific analysis results or by transferring good practices found in the project documents.
5. Other interested persons checking the existence of CS projects in any field and/or searching for appropriate CS projects to get involved in.
6. Researchers and academics from different disciplines will be able to re-use the data for their own research.
7. Science and technology (and related markets) analyzers and commentators.
8. Teachers and school principals

2 FAIR data

CS Track currently participates in the Pilot on Open Research Data in Horizon 2020, which aims to improve and maximize access to, and re-use of research data generated by actions. Thus, the project consortium is committed to the FAIR data principle.

2.1 Making data findable, including provisions for metadata

The disclosed research data will be made available through Zenodo (https://zenodo.org/):

"The Open AIRE project, in the vanguard of the open access and open data movements in Europe, was commissioned by the EC to support their nascent Open Data policy by providing a catch-all repository for EC funded research. CERN, an OpenAIRE partner and pioneer in open source, open access and open data,

provided this capability and Zenodo was launched in May 2013. In support of its research programme CERN has developed tools for Big Data management and extended Digital Library capabilities for Open Data. Through Zenodo these Big Science tools could be effectively shared with the long-tail of research." (Quoted from https://about.zenodo.org/)

2.1.1 Metadata standards

Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML (cf. https://about.zenodo.org/principles/ F2)

Zenodo is compliant to the OpenAIRE Guidelines v3.0. [https://about.zenodo.org/principles/]

2.1.2 Identifiability of data

Zenodo provides DOI-links. The DOI is a top-level and a mandatory field in the metadata of each record (cf. https://about.zenodo.org/principles/ F3).

2.1.3 Naming conventions

Following are Princeton's guidelines on naming conventions:

- Files should be named consistently
- File names should be short but descriptive (<25 characters) (Briney)
- Avoid special characters or spaces in a file name
- Use capitals and underscores instead of periods or spaces or slashes
- Use date format ISO 8601: YYYYMMDD
- Include a version number (Creamer et al.)

Thus, we use the following format to name our files that are shared among each other and via the open research data repository:
<workpackageno>_uniquefilename_YYYYMMDDVV.<filextension>
Example:
WP8_datamanagementplan_2020043001.docx

2.1.4 Search keywords

We will use the citizen science and CS track with every resource shared by the project. This allows to group the data within the repository apart from creating a community (a subgroup within Zenodo). Additional keywords will be derived from the specific data resource. For every resource we will make exploratory searches for related keywords and adopt them for the specific keyword set.

2.1.5 Versioning

All project internal documents will have a version number. The changelog and version numbering are done manually.

Software artefacts will be maintained in code repositories that will be referenced within the research object repository. The code repositories will provide their own version and release numbering system. Zenodo will provide version numbers for releases of any kind of research object as well using the well-established system of major and minor version numbers based on the effect and size of changes.

2.1.6 Standards for metadata creation

Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML (cf. https://about.zenodo.org/principles/ F2)

Additionally, we will use a project-specific metadata standard, making re-use of e.g. the Citizen Science Ontology created by the COST project (originally funded by the EU).

2.2 Making data openly available/ accessible

2.2.1 Openness of Data

In general, the data generated within the project including deliverables of WP1 to WP5 will be open w.r.t. to the license that accompanies each particular research data object.

The project will also share all data collected during its research. At least it will share all aggregated/processed data and analysis results, if there is personal (raw) data involved.

We can presently perceive the following data categories and main channels through which our data will be accessible to the users:

- Pre-processed, raw data.
- Data/information in the processed-analyzed spectrum:
    - retrievable via the project's platform.
    - shown in the project's e-magazine.
- Studies and elaborated results of our analysis; and
- Knowledge-based policy recommendations.

The above channels/categories presuppose a different (increasing) degree of elaboration, aggregation, processing and (multi-perspective) analysis, readily associable to the path data → information → knowledge, which will ultimately determine (1) the respective typical users, (2) the necessary (or possible) degree of openness or availability. Our current plans in this regard can be summarized in the following table:

Row #	**Data/info channel/category**	**Likely typical user** (category # in section 1.7 above)	**Open/not open**/(conditions)
1	Pre-processed, raw data	1.	Not open. All or most of the relevant data will be originated in interviews, questionnaires and web-analytics and, accordingly, subject to restrictions from ethical considerations. Other considerations may nevertheless apply for not freely releasing these data.

2	Data and information retrievable via the project's platform	In principle, all, but especially 2., 3. and 4.	Open
3	Data and information shown in the project's e-magazine	In principle, all, but especially 2. to 8.	Open
4	Studies and elaborated results of our analysis	Especially 6.	Open [e.g., publication of a paper by project's staff; etc.]
5	Knowledge-based policy recommendations	2.	Open

2.2.2 Availability

The research data will be made available through Zenodo. The items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental program defined for the next 20 years at least. https://about.zenodo.org/principles/. Additionally, the project makes its data available to a wide audience through the e-magazine. The former is more data-oriented and the latter more information-oriented. Both will be fully and openly available, offering the data and information that the project decided to release (see 2.2.1 above).

2.2.3 Accessibility

To be accessible, the Zenodo repository provides a standardized communications protocol (OAI-PMH and REST API), harvestable by search engines and crawlers (see https://about.zenodo.org/principles/). The protocols also allow for authentication and authorization procedure, where necessary, but the metadata are publicly accessible and licensed under public domain. No authorization is ever necessary to retrieve it. Metadata is even accessible if the data is no longer available. There is no need to register to access the openly data shared via the repository.

Source code created within the project that is openly accessible will be provided in open code repositories like GitHub will be referenced within Zenodo.

Both the project-created platform and the e-magazine will be freely accessible for any user from a web browser. Registration and identification may be required for accessing data of a particular citizen science activity, but no cost or other impediments will apply.

2.2.4 Storage

The open research data will be stored at Zenodo. We will either store the actual dataset or a link to the data (e.g. in the case of open archive publications).

Zenodo has assessed itself against the Plan S requirements for Open Access Repositories (as published October 2019) and succeed to fulfil all requirements https://about.zenodo.org/principles/

Data that cannot be shared openly such as personal data that cannot be shared either for legal reasons or ethical reasons has to be stored outside of Zenodo, as Zenodo does not guarantee that its employees are not able to access the restricted data.

The confidential data will be stored safely under the conditions and obligations of GDPR by the partner who is collecting this kind of data and afterwards by the partner who stores the data to grant access for collaborative analysis within the project consortium (see also deliverable 8.2).

2.2.5 Access control

Open research data: the Zenodo-way

- **Access to data objects**: Files may be deposited under closed, open, or embargoed access. Files deposited under closed 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.
- **Use and re-use of data objects**: Use and re-use is subject to the license under which the data objects were deposited.
- **Embargo status**: Users may deposit content under an embargo status and provide an end date for the embargo. The repository will restrict access to the data until the end of the embargo period, at which time the content will become publicly available automatically.
- **Restricted Access**: Users may deposit restricted files with the ability to share access with others if certain requirements are met. These files will not be made publicly available and sharing will be made possible only by the approval of depositor of the original file.
- **Metadata access and reuse**: Metadata is licensed under CC0, except for email addresses. All metadata is exported via OAI-PMH and can be harvested.
- **Confidential data:** Confidential data will be anonymized as soon as possible and will stored securely following the guidelines of Finnish National Board of Research Integrity (TENK; https://www.tenk.fi/en) as well as the analysis will do.

2.3 Making data interoperable

2.3.1 Compliance to standard methodologies

Zenodo provides an API for accessing its data by electronic means. While this provides basic accessibility, we will base our data set formats on widely known and used data format standards like, COST Citizen Science Ontology, YSO The general Finnish Ontology and other well-defined data models. Whenever the project decides to stray from the standard or establishes a new data model, the documentation of the syntax and semantics of the data will be referenced by the data set and the reference manual will be published openly.

2.3.2 Compatibility to standards

The stored data will make use of metadata standards wherever possible. To further the convergence of terminology on Citizen Science we will especially make use of

the Citizen Science Ontology (Ceccaroni et al (2018). Citizen-science ontology. https://doi.org/10.5281/zenodo.3721074.)

But of course, we will also follow Dubin Core and other standard formats to create metadata as required by Zenodo.

2.4 Increase data re-use (through clarifying licenses)

Whenever the data is stored in the open data repository, a license will be specified to clarify the extent of allowed usage. Zenodo provides links to a large amount of well-known open licenses (and their different variants and versions). Thus, the project team may decide on an appropriate license for each data object. Choosing well-known open (source | document | etc.) licenses ensures a widespread re-use of the data.

Currently, the project consortium does not consider any embargos on publishing the open research data, but if the need arises, e.g. to protect the interests of third-parties (including other research groups within the partners' institutions), the chosen repository allows for pre-defining embargo deadlines.

2.4.1 Licenses

We will have a close look at open source licenses. In general, the project will use one of the open source licenses that is most fitting to the data object (e.g. GPL licenses for software, Creative Common licenses for textual data).

The e-magazine as well as the data inside of the platform will also clearly state an appropriate license for their content.

For the database entries (including questionnaire formats), the project currently considers using the Open Data Commons Open Database License as a default license.

2.4.2 Availability of data for re-use

As per Section 2.2 above. Allowed with the above conditions (2.4.1).

2.4.3 Data re-use by third parties

Third parties that may need special attention w.r.t. licensing are those referred in rows no. 4 and 5 in the table in section 2.2.1.

2.4.4 Data quality assurance

To ensure the quality of the delivered data, mostly peer reviews among the researchers of the work packages will take place. Apart from the finally delivered data to the repository each step of the collection and analysis of the data will be conducted in a peer review process. This is enforced by the fact that the data is "moving" through the work package so that each work package will validate the correctness and plausibility of the received data.

Furthermore, a thoroughly discussed experiment design (survey design, target groups, participants) will enable high quality data – also for re-use by other researchers.

The extensive use of controlled vocabularies will further advance the re-usability.

A loss of data quality at any stage of the data manipulation process reduces the applicability and uses to which the data can be adequately put (Chapman 2005). Thus, the metadata of the collected data will comprise the following documentation:

1. time of gathering of the data
2. Information about data manipulation prior to digitization (e.g. label preparation), if applicable
3. How the data was digitized, if applicable
4. Metadata created to enrich the data object
5. DOI: to indicate where the data is stored
6. How the data is used (analysis & manipulation)
7. Where the data has been presented (e.g. paper, electronic publication, databases), if applicable

Since the project aims at gaining deep insights in Citizen Science and characterize these activities along various dimensions a taxonomical approach seems feasible. Thus, the data created and collected within the project should have entries mapping categories to standardized vocabularies and nomenclatural status (synonym, accepted, typification), reference (author, place and date of publication), determination (by whom and when the record was identified), quality fields (accuracy of determination, qualifiers)

2.4.5 End of re-use

The data stored in the open research data repository will be stored for as long as the repository continues its service. Currently, the chosen repository has secured funding for approximately 20 years. https://about.zenodo.org/policies/

Articles that may be published in open archives and cannot be stored in the open research repository may have other time frames.

3 Allocation of resources

3.1 Estimated costs

We have thought of three approaches to estimate the costs, all of which with a different conceptual basis (say, answering a different question, which could be more or less relevant depending on the context). All the approaches (especially the 2nd and 3rd below) yield tentative estimations that should be taken as a preliminary attempt to give an order of magnitude. Both the approaches and the numerical results will be reviewed in future versions of this Data Management Plan.

One approach to estimate the costs is looking at our data as "the project's output". In principle, the whole CS Track project is aimed at producing data (and information) and making them available and useful for/to the foreseen uses/users (say, as detailed in section 1.7 above). Even our management and dissemination / exploitation costs can be associated with the production and/or provision of the data to all foreseeable users (if users were not aware of the existence of our data, it

would not make much sense to produce the data…). Accordingly, we can see the whole project cost – 2.3 MEuros – as the total cost of all our data.

A second approach looks at the additional, or specific, cost incurred to warrant the FAIRness of the data (as part of the project's total cost). Which of our costs may be attributed more or less directly to the fact that we want to make our data FAIR?

The existence of the Zenodo data repository, with all its affordances, saves much (not specifically budgeted) cost. As long as Zenodo is made available for free for usage by the project, there is no additional cost included for the storage of open research data. Costs for publication of e.g. journal articles and conference proceedings during the funding period of the project are allocated in the budget of the project already (a limited amount preliminarily deemed appropriate). The personnel costs of making data FAIR is also covered by the project funding, as it entered our estimations of the overall work needed.

The costs of publishing the initial community platform and the e-magazine are also covered by the project funding. The sustainable preservation of the created content comes down to the cost of hosting and maintaining the website (e.g. security updates).

Assuming different tentative assumptions about the percentage of each partner's costs, for each of the budget sub-categories we considered in the budget preparation, we can estimate the total cost of FAIR compliance at about 450 KEuros, i.e., some 20% of the project's total cost (budget). This is a very tentative and potentially inaccurate estimation and, as such, subject to correction in future releases of this DMP. It probably (much) overstates the cost, as it considers FAIRness very broadly, including some elements (or portions thereof) that could be considered "normal" rather than FAIR.

A further, third approach is inspired by some available sources (e.g., https://datawizkb.leibniz-psychology.org/index.php/before-my-project-starts/what-should-i-know-about-costs-of-data-management/ ; https://www.ukdataservice.ac.uk/manage-data/plan/costing (and then https://www.ukdataservice.ac.uk/media/622368/costingtool.pdf), which look at two alternative perspectives that are much related to the above two approaches:

1. Estimating costs for all data related activities (data collection, data processing, data analysis, data sharing and data management);
2. Estimating only additional costs that are necessary for data management and data sharing procedures and go beyond the costs of standard procedures in research projects.

These brought us to consider focusing the question on whether the effort is causing additional costs or if it is part of "good research practice". Intuitively, and as a first approximation, the additional work may be caused by

- Additional quality procedures (e.g. peer review before storing) and
- Documentation
- Metadata creation
- Uploading and perhaps cleaning.

Without using the proposed costing tool (referred in the last link quoted above) it sounds a priori sensible to assume 1PM, on average per partner, for the whole

project's life for the above. This means that, tentatively, about 9 PMs, of a total of 322 PMs to be invested, overall, in the project, would be devote to the above tasks. In money terms, this is approximately 50-60 KEuros (depending on the basis taken and on details of the calculation).

3.2 Responsibility for data management

The data arises in the various work packages within the project. The work package leaders will manage the data creation and management process for their respective work packages. Consistent with the project's organizational/managerial structure, the Impact Assurance Coordinator, the Citizen Science Committee (CSC) and the Enabling Technologies & Analytics Committee (ETAC) will take care of any frictions between the work packages in all matters related to data, and set up the general management principles. To operationalize this responsibility pattern, we established a small team – the DMP Task Force – comprising four persons that were nominated in a recent meeting of our Project Management Board. They include the two authors of this Data Management Plan (representing project management / coordination and the ETAC), the impact Assurance Coordinator and, representing the CSC, the Ethics and Gender Coordinator. The DMP Task Force will further the impact of and awareness to the DMP subject both at project macro level and at the level of the actual and concrete tasks of data gathering and handling. We plan that discussions, decisions and managerial actions related to data types, storage, protection/ethics, accessibility, links with third parties if any, etc., will be coordinated by this team. In case of need it will pass the matter to the PMB decision, after its study and pre-analysis.

3.3 Long term preservation

Zenodo does not charge any membership or other fees. Its funding is guaranteed for the next 20 years funding by public bodies. Thus, no additional costs are expected for preserving the open research data.

The confidential data will be stored at one of our co-partner's facilities as long as the project is funded and will be deleted according to the storage policy of the storing partner. Currently, we do not expect any benefit from preserving this data any longer as it will not be shared. Otherwise, it would have been public data to begin with. In any case, all options will be considered in due time.

The community platform as an active entity needs either a successful business model including paid memberships or continuous funding by another entity. Otherwise, the content will age and the community will become inactive. In the latter case, it does not seem reasonable to preserve the platform as such. It would be more reasonable to transfer the data into the open research data repository as part of the project's data legacy then.

4 Data security

There are two types of data to be distinguished in the project. Open research data and data that needs confidential handling because of legal regulations (e.g. GDPR) or ethical considerations. Open research data is shared via Zenodo that provides a

professional data security services like data replication, access control etc. https://about.zenodo.org/infrastructure/

The confidential data will be stored safely under the conditions and obligations of GDPR by the partner who is collecting this kind of data and by the partner who stores the data to grant access for collaborative analysis within the project consortium (see also deliverable 8.2).

5 Ethical aspects

Ethical aspects and those related to the protection of personal data (PDP) were initially addressed in the two deliverables of WP8 (Ethics) – D8.1 and D8.2 – released in March 2020. The reader is referred to them, as they, for the moment, condense our available/usable information on the subject. Those deliverables, as well as this Data Management Plan are live documents that quite surely will evolve with the project and, accordingly, their contents are subject to change and adaptation.

[This last page is left empty. For technical reasons, it could not be deleted.]