This project is co-financed by the European Union

Grant Agreement No.: 824603

Call: H2020-SwafS-2018-1

Type of action: RIA

Starting date: 1/02/2019

D1.1 - Data Management Plan (v2)

Contributors: All partners

Coordinator / Reviewer: Gefion Thuermer, King's College London

Deliverable nature	ORDP: Open Research Data Pilot
Dissemination level	Public
Work package and Task	WP1; T1.3
Contractual delivery date	31/07/2019
Actual delivery date	29/05/2020 (v2)

Authors

Author name	**Organization**	**E-Mail**
Gefion Thuermer	King’s College London	gefion.thuermer@kcl.ac.uk

Abstract	This document serves as the data management plan for the ACTION project and consists of individual data management plans for all datasets held across the ACTION consortium, and covering the activities of, both in terms of pilot projects and work package research activities.
Keywords	Data Management, Open Data, Open Science, Quantitative, Qualitative, Licenses

Disclaimer

The information, documentation and figures available in this deliverable, is written by the ACTION project consortium under EC grant agreement 824603 and does not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

How to quote this document

Thuermer, G. (2020) D1.1 - Data Management Plan (v2).

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This document serves as the data management plan for the ACTION project. It consists of a number of individual data management plans, covering the pilot projects that form part of the ACTION consortium, as well as additional data management plans for the datasets to be gathered and maintained as part of the work package activities in WP3 (Open Call), WP5 (Socio-Technical Toolkit) and WP6 (Impact Assessment). At the time of production of this DMP, no further datasets are foreseen. Should additional datasets be required or produced as part of ACTION, updated versions of this DMP will be produced and submitted as necessary.

Cefriel – TESS network survey dataset

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The dataset provides the structure (question content and survey structure) of the survey designed to collect quantitative information about the motivation to participate to citizen science campaign. The questionnaire is designed and administered in the form of conversational surveys. The dataset also includes the answers collected from 79 participants and the analysis done on the collected data.

The studied campaign is the TESS network initiative that involved people in measuring the sky darkness through sensors.

What types and formats of data will the project generate/collect?

1) Structure and content of the survey used to collect data, exported in RDF format. The structure of the survey in terms of questions, pre-defined answers and compilation flow is described with respect to the Survey Ontology (https://w3id.org/survey-ontology)

2) Answers to the survey, both in RDF and CSV format. The structure of the CSV is the following:

- *questionId*: universal identifier of the question
- *question*: text content of the question
- *questionType*: the visualization of the question (checkbox, text, star-rating, options, emoji, slider, select)
- *tag*: tag of the question, if assigned
- *option*: text value of closed answer (i.e. checkbox or multiple choice questions), if present
- *value*: numeric value of closed answers (i.e. checkbox or multiple choice questions)
- *freeAnswer*: text value of open-ended questions
- *points*: points assigned to the answer (default is 0, non-mandatory)
- *user*: id of the user that answered the question (empty if nobody did, u_xxxxx are anonymous ids)
- *language*: language of survey completion related to the user (empty if nobody answered)
- *date*: date of answer submission (dd/mm/yyyy)
- *time*: time of answer submission (hh:mm:ss)
- *session*: an identifier of the session of completion (i.e. if the same user fills the survey twice, the two will have different session ids)
- *totalDuration*: total time spent by the user filling the survey (hh:mm:ss)
- *projectId*: id of the project the survey is part of (if any)
- *projectName*: name of the project the survey is part of (if any)

The same data about the collected answers is also offered in RDF, again accordingly to the Survey Ontology.

3) Results of the analysis done on the collected data in CSV format. The structure of the CSV is the following:

- *variable*: motivating factor (or latent variable)
- *mean*: mean of the answers of all users to the question tagged with this motivating factor
- *stdev*: mean of the answers of all users to the question tagged with this motivating factor
- *correlation*: correlation between this motivating factor and the global motivation
- *pvalue*: p-value of the correlation
- *significance*: significance level of the correlation (p-value *** <0.001, ** <0.01, * <0.05)

Will you re-use any existing data and how?

We do not re-use data in this project, but part of the survey structure is inspired by related work on motivational studies: “Questionare for the Motivation for Citizen Science Scale” by Levontin, Chako and Gilad.

What is the origin of the data?

We generate the structure of the survey. The answers collected are generated by the survey compilers.

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

We have identified the following community that could be interested in using our data:

- survey designers that need statistics about citizen scientists' participation
- survey designer that want to design a new survey about motivation
- light pollution communities
- scientists that want to compare studies of motivations in different scenarios

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)? What naming conventions do you follow?

A Digital Object Identifier (DOI) will be generated using Zenodo to make it findable. This dataset will be findable from our Zenodo Community (http://zenodo.org/communities/actionprojecteu) and from our Data Portal

Will search keywords be provided that optimize possibilities for re-use?

We have identified the following keywords: conversational survey, citizen scientist motivation, light pollution, TESS network

Do you provide clear version numbers?

For this survey, we will produce a unique data set that will not be periodically updated.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

We will also include metadata as defined on Zenodo. This includes: upload type (publication, poster, Dataset, etc...), DOI, publication date, title, authors (full name, affiliation, ORCID), description, version, keywords, additional notes, access right (open, embargoed, restricted, closed) and license.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

All the datasets will be available by default under a CC-BY license.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

There is no provision to keep the data closed and no specific beneficiaries are foreseen.

How will the data be made accessible (e.g. by deposition in a repository)?

We will set up a web infrastructure and upload our data to the Zenodo public repository.

What methods or software tools are needed to access the data?

An ordinary web browser will be enough to download data from our data portal or public repository. Datasets will be in CSV and RDF. Optionally, the user could use our search facility of our data portal.

Is documentation about the software needed to access the data included?

The API of Zenodo can be used to consult the different datasets deposit in these repository. Plus, there are open software tools available to process datasets

Is it possible to include the relevant software (e.g. in open source code)?

Examples querying and using data will be uploaded to a public repository (i.e. GitHub).

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data will be stored in a public repository, our data portal providing search facilities. Documentation will be available on our website (http://actionproject.eu). The code will be uploaded to our public repository in Github. (https://github.com/actionprojecteu)

Have you explored appropriate arrangements with the identified repository?

No.

If there are restrictions on use, how will access be provided?

There will be no access restrictions.

Is there a need for a data access committee?

No.

Are there well described conditions for access (i.e. a machine readable license)?

The license of the data is CC BY

How will the identity of the person accessing the data be ascertained?

There is no need to identify the person to access and to download the data selected

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

This data uses a specific vocabulary to make the data interoperable.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Our data is described with respect to the Survey Ontology (https://w3id.org/survey-ontology)

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Yes

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The Survey Ontology already reuse some commonly used ontologies and vocabularies, such as PROV-O (http://www.w3.org/ns/prov#), the Data Cube Vocabulary (http://purl.org/linked-data/cube#), the Research Object suite of ontologies (http://purl.org/wf4ever) and the Research Variable Ontology (http://w3id.org/rv-ontology#).

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The data have been published with the following license: CC-BY

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be available as soon as it is uploaded our data portal or public repository.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

We have identified this community interesting in using our datasets.

- survey designers that need statistics about citizen scientists' participation
- survey designer that want to design a new survey about motivation
- light pollution communities
- scientists that want to compare study of motivations in different scenarios

There are no restrictions in the use of the data. We only nicely request to cite us using our DOI.

How long is it intended that the data remains re-usable?

Reusability of data is tied to the actual data sources to be catalogued.

Are data quality assurance processes described?

Not yet defined at this stage.

3. Allocation of resources

What are the costs for making data FAIR in your project?

The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held by UPM

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Data infrastructure cost during the project will be covered by the grant.

Who will be responsible for data management in your project?

The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories, data is hosted on UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories, data is hosted on UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

We do not foresee any legal or ethical issues as a result of this dataset. No personally identifiable or sensitive data will be gathered through these activities.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

N/A – no personal data is to be gathered as part of this dataset.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far

DBC – Dragonflies and Pesticides

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data on pesticide concentrations in different water bodies to relate to the local dragonfly community

What types and formats of data will the project generate/collect?

The data will consist of location information in coordinates (not exact), an identifier to link with dragonfly transects in the DBC database, date and concentrations for a large number of compounds (with CAS code).

Will you re-use any existing data and how?

In this project, we re-use data from our long term monitoring scheme that is stored on our own servers, this is not publicly accessible. The core data is also at https://www.ndff.nl/

What is the origin of the data?

At this moment, we are generating data in our project. This is stored on the DBC server, not publicly accessible. The core data is also at https://www.ndff.nl/

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

We have identified the following community that could be interested in using our data:

To people interested in the water quality of water bodies that are not covered in the Water Framework Directive regulations.

FAIR data

Making data findable, including provisions for metadata

A Digital Object Identifier (DOI) will be generated using Zenodo to make it findable. This dataset will be findable from our Zenodo Community (http://zenodo.org/communities/actionprojecteu) and from our Data Portal

Will search keywords be provided that optimize possibilities for re-use?

We have identified the following keywords: pesticides, surface water, NL, Dragonflies, DBC

Do you provide clear version numbers?

For this pilot, we will produce a unique data set that will be periodically updated. So we will use versioning for the dataset. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. We will also include metadata defined

We will also include metadata defined by CKAN. This includes (title, description, tags, license, source, version, author email and some another custom fields)

Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

The datasets on water quality will be publicly available. The data on dragonflies will be available through the portals where they are accessible already (NDFF.nl and GBIF.org, dedicated databases for biodiversity data)

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

The data on the occurrence of dragonflies will be available but only through the existing portals.

How will the data be made accessible (e.g. by deposition in a repository)?

We will set up a web infrastructure and upload periodically our water quality data to a public repository (i.e. Zenodo). Our data Portal (based on CKAN) will be used for searching but not physically store the data, rather it will keep a link to the public repository.

What methods or software tools are needed to access the data?

An ordinary web browser will be enough to download this data from our data portal or public repository. Datasets will be in CSV. Optionally, the user could use our search facility of our data portal.

Is documentation about the software needed to access the data included?

The API of Zenodo or our data portal can be used to consult the different datasets deposit in these repositories. Plus, there are open software tools available to process datasets

Is it possible to include the relevant software (e.g. in open source code)?

Examples querying and using data will be uploaded to a public repository (i.e GitHub).

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data on water quality will be stored in a public repository, documentation will be available on our website (http://vlinderstichting.nl). The code will be uploaded to our public repository in Github. (https://github.com/actionprojecteu)

Have you explored appropriate arrangements with the identified repository?

No.

If there are restrictions on use, how will access be provided?

There will be no access restrictions to the water quality data. There is for the NDFF.nl data as this includes detailed information on the distribution of rare and protected species. The NDFF has a dedicated portal where data is available to selected users. Data can be made available upon request. In GBIF.org it is publicly available but on a coarser spatial scale.

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

The license of the data is CC BY-SA

How will the identity of the person accessing the data be ascertained?

There is no need to identify the person to access and to download the data selected for the water quality data. For the distribution data accessibility i is already regulated by the NDFF portal.

Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

This data use a specific vocabulary to make the data interoperable.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

CAS Registry Numbers will be used to identify compounds

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

CAS Registry Numbers (unique identifiers for chemicals) will be used to identify compounds

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

CAS Registry Numbers will be used to identify compounds

Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The data have been published with the following license: CC BY-SA

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be available as soon as the dataset is generated.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

We have identified this community interested in using our datasets: people interested in the water quality of water bodies that are not covered in the Water Framework Directive regulations. Data will be available as soon as the dataset is generated with the periodicity configured by the coordinator

How long is it intended that the data remains re-usable?

The water quality data will not be updated but will be stored in a way that it will remain available for at least several decades. The distribution data is included in the NDFF and that guarantees it will be used for the foreseeable future.

Are data quality assurance processes described?

For the water quality data it is not yet defined at this stage. For the dragonfly data there are validation tools in place where algorithms flag suspicious observations. These are checked by experts.

Allocation of resources

What are the costs for making data FAIR in your project?

The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held by UPM

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Data infrastructure cost during the project will be covered by the grant.

Who will be responsible for data management in your project?

The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There are no ethically sensitive issues. The only concern is privacy, which is solved by anonymization of the data.

Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far

Drift

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of the data collection/generation is to answer our research questions. We aim to develop a toolkit for citizen science, to evaluate the impact of citizen science initiatives and to design policy master classes. For this, we will collect and generate data about the state of the art in citizen science research and practice, co-design, implement and document citizen science methods and practices, and co-evaluate impact.

What types and formats of data will the project generate/collect?

We will collect unstructured data mainly in Word, mp4, PDF, html, jpeg formats (raw data) and word, xls, Pdfs and html formats (processed data).

An overview of the datasets that we collect can be found in the table below. In principle, everything in this DMP refers to all the datasets. Exceptions are explicitly mentioned, using the dataset reference.

Dataset reference	Origin	Description
ACTION_INTERVIEW DATA_Cherishing	interviews	Semi-structured interviews with citizen scientists and with project coordinators to find our more about the motivation and participation of the volunteers. The data will be stored on the (secure) servers of the Erasmus University Rotterdam, and any personal data will be pseudonymised.
ACTION_SURVEY_Cherishing	Survey data	Surveys for citizen scientists to assess motivation and participation in their project. Any personal data will be anonymised.
ACTION_WORKSHOPS_Democratisation	Results from exercises during workshops	Feedback to exercises during workshops. This data will be used for a paper and/or blogs about the potential of citizen science to democratise science. Personal data was not collected.
ACTION_FEEDBACK_impact assessment	Interviews	Feedback given by pilots to the impact assessment materials and process. This data might be used for papers and for improving the impact assessment framework. Any personal data will be pseudonymised.
ACTION_FEEDBACK_MC	Feedback received during masterclasses	Feedback given by participants to the policy masterclasses. This data might be used for papers and for improving the masterclasses and policy brief. Any personal data will be pseudonymised.
ACTION_INTERVIEW DATA_Policy MC	Interviews	Semi-structured interviews with people that are knowledgeable about the policy context of the six countries in which a policy masterclass will be held. The data will serve as input for content development of that masterclass, possibly for a paper, and for the policy brief. Any personal data will be pseudonymised.

Will you re-use any existing data and how?

We will not re-use any existing data.

What is the origin of the data?

The origin of the data are interviews, case studies, ethnographic participant observation and personal data. Personal data includes name and e-mail to contact interviewees and possible participants of educational master classes, and employer and job position as well as gender, educational level, and language spoken at home to ensure diversity among participants.

What is the expected size of the data?

Expected size of the data are a few hundred MB.

To whom might it be useful ('data utility')?

The data is useful for citizen science initiatives, researchers on citizen science, policy makers that are interested in or dealing with citizen science, and the general public.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

What naming conventions do you follow?

Will search keywords be provided that optimize possibilities for re-use?

Do you provide clear version numbers?

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Our research data will mostly not be made openly accessible (ACTION_INTERVIEW DATA_Cherishing, ACTION_WORKSHOPS_Democratisation,ACTION_FEEDBACK_impactassessment, ACTION_FEEDBACK_MC, and ACTION_INTERVIEW DATA_Policy MC, ACTION_SURVEY_Cherishing). The interviews and feedback during workshops are collected with a quite instrumental research question and goal in mind and seem unfit to be shared beyond the project nor stored longer than necessary (10 years for verification). This along with the pseudonymisation would prevent asking anew for consent and seems to make reuse impossible. This also means that the data will not be FAIR, so this applies to all sections of 2.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

What methods or software tools are needed to access the data?

Is documentation about the software needed to access the data included?

Is it possible to include the relevant software (e.g. in open source code)?

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Have you explored appropriate arrangements with the identified repository?

If there are restrictions on use, how will access be provided?

Is there a need for a data access committee?

Are there well described conditions for access (i.e. a machine readable license)?

How will the identity of the person accessing the data be ascertained?

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

How long is it intended that the data remains re-usable?

Are data quality assurance processes described?

Further to the FAIR principles, DMPs should also address:

3. Allocation of resources

What are the costs for making data FAIR in your project?

No costs will be incurred.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?

The data management process of the ACTION consortium is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Julia Wittmayer, along with Kali den Heijer will be responsible for data management on the DRIFT part of the project.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Is the data safely stored in certified repositories for long term preservation and curation?

The data will be stored on the shared drive of DRIFT hosted by EUR IT, and/or on NextCloud.

The data is shared with DRIFT colleagues via the shared drive provided by EUR IT and GSuite. In addition, the consortium is currently settling down on using NextCloud.

For the purpose of this research and along with standards in the social sciences, a storage period of ten years is seen as appropriate. Since the field of citizen science is rapidly changing, and the research data is either focusing on policy lessons or on taking stock, it will very quickly be outdated and no longer storage will be necessary.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

We include informed consent for data sharing and long-term preservation in questionnaires dealing with personal data.

The data will be pseudonymised and where possible an effort will be undertaken to anonymize data.

In case of interviews, the use of a pseudonym will be discussed and agreed upon with the interviewee and reported on in the interview summary. This might mean that when quoting an interviewee s/he will be referred to in terms of his/her function or job title, rather than by name and the pseudonymization will be taken to a degree that is agreeable for the interviewee e.g. the Director of the Butterfly Association; a director of a Dutch environmental association; a member of a Dutch citizen initiative; etc.

In case of pictures that make it possible to identify people, we'll ensure to have the written consent of these people before using these in any kind of publications.

In case of participant observation, when we aim to quote somebody, we'll ensure that these quotes will not be linked to personal data, but rather that the person will be referred to as 'participant in xyz'.

In case direct quotes are used to illustrate the more popular outputs of the project (such as policy briefs or blogs) the interviewees will be asked whether they agree to the use including a reference to their name.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

EUR (Erasmus University Rotterdam) Data Management procedures

FVB-IGB – ACTION Accelerator

Data Management Plan

This document serves as the data management plan for the ACTION Accelerator, led by FVB-IGB. Each ACTION Pilot project has its own data management plan.

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data is collected through the ACTION Accelerator activities to allow for effective communication between consortium partners and Accelerator Pilot projects. Data is generated to assess the mentoring needs of the Accelerator Pilot projects, to gauge their progress through the Accelerator process, and to document Accelerator activities such as the Kick-off Workshops.

What types and formats of data will the project generate/collect?

Data is collected through the ACTION Accelerator activities to allow for effective communication between consortium partners and Accelerator Pilot projects. Data is generated to assess the mentoring needs of the Accelerator Pilot projects, to gauge their progress through the Accelerator process, and to document Accelerator activities such as the Kick-off Workshops.

Will you re-use any existing data and how?

We do currently not plan to re-use existing data. Personal data will not be re-used.

What is the origin of the data?

Data is generated by the Accelerator pilot projects (photos, completed progress statistics and assessment documents) or FVB-IGB (templates for assessments and progress statistics, workshop material, photos and webinar videos).

What is the expected size of the data?

A few GB.

To whom might it be useful ('data utility')?

Data monitoring the progress of the projects will be useful within the consortium. Aggregated data and workshop materials will be useful to the wider citizen science community interested in inclusion and diversity.

FAIR data

Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Where data is published it will include relevant metadata.

What naming conventions do you follow?

Where relevant, files are named with the file type and Accelerator Pilot name, with a version number as required.

Will search keywords be provided that optimize possibilities for re-use?

Where necessary keywords will be provided.

Do you provide clear version numbers?

Version numbers are used where necessary.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

In most cases elaborate metadata is not necessary. Automatically generated metadata on documents and photographs will be preserved.

Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

No datasets will be generated or made openly available by default. Most data generated is specific to how the pilot projects run and is not suitable for publication.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

Most of the data will not be public. Only videos and photos as well as templates for assessments and progress statistics and partly the workshop material will be published on the actionproject.eu website under CC-BY-SA 4.0 License. Videos and photos might be also published on open social media channels.

What methods or software tools are needed to access the data?

Data will be accessible through web portals with no special software or tools necessary.

Is documentation about the software needed to access the data included?

Where needed a short description will be provided but we don't expect many cases.

Is it possible to include the relevant software (e.g. in open source code)?

If available the relevant software will be included.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Documentation will be available on actionproject.eu.

Have you explored appropriate arrangements with the identified repository?

N/a

If there are restrictions on use, how will access be provided?

N/a

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

N/a

How will the identity of the person accessing the data be ascertained?

There is no need to identify the person

Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

The data generated in the Accelerator project are not interoperable due to their specificity.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

None

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

No

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

If necessary we will provide mappings for project specific ontologies and vocabularies.

Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Videos and photographs, and only not personalised workshop materials and templates will be published with a CC BY-SA 4.0 license.

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Videos and photographs will be published as required providing release forms have been signed by the involved persons if necessary. All other data that is to be published publicly will be made available 6 months after the close of the relevant Accelerator round at the latest.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Multimedia content and workshop protocols are reusable under CC-BY-SA 4.0 license.

How long is it intended that the data remains re-usable?

The data will remain re-usable for the duration of the project.

Are data quality assurance processes described?

There is no standard QA process for these data.

Allocation of resources

What are the costs for making data FAIR in your project?

There are no costs incurred barring the person time required to prepare materials suitable for publication.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Time costs are incorporated into the grant.

Who will be responsible for data management in your project?

The data management process is managed by FVB-IGB, as leader of WP2.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

No provision for long term preservation has been made at this time.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Data are hosted in consortium partner UPM's servers in a specially set up Nextcloud instance. Access to specific pilot data is granted to the pilot project team and to consortium partners who require access to complete tasks. Only personnel authorized can access these servers and periodical backups are being done. An opt-in mailing list, hosted by UPM, has been set up for communication between Accelerator projects and the consortium partners. Data are additionally stored on the IGB Nextcloud instance for access by ACTION members from FVB-IGB.

Is the data safely stored in certified repositories for long term preservation and curation?

Data are hosted in consortium partner UPM's servers in a specially set up Nextcloud instance. Only personnel authorized can access these servers and periodical backups are being done.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Yes – some of the data gathered during this process will be personally identifiable and therefore sensitive data. If needed participants will be asked for their consent for the use of their data. No data sharing and no long term preservation are required for this data.

Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No

FVB-IGB – Loss of the Night

1. Data Summary

The purpose of data collection is to allow citizen scientists to make precise observations of night sky brightness, in order to measure how this environmental parameter is changing over time.

Each data record contains the following:

- App version #
- Unique anonymous install identifier
- Language of application
- Language of device system
- Self-evaluation of weather conditions (tickboxes)
- Initial estimate of sky brightness (from app, based on data)
- Star information:
  - Unique star identifier
  - Magnitude of star (from table)
  - User decision on visibility
  - User classification for visibility or invisibility
  - Timestamps (UTC and local)
  - Geographical location (longitude and latitude)
  - Estimate of location accuracy
  - Magnetic field measured by device

Some data records also contain the following optional information:

- Self-chosen username
- Participant email address
- Age group (decade)
- Details about vision (e.g. wears glasses or contacts)
- Self-evaluation of experience (amateur, moderate, professional)
- Sky Quality Meter observation

The star identifiers and magnitudes come from the “Yale Bright Star Catalog”.

Data is provided by users of the application.

Up to several thousand observations per year. The current dataset has a size of 16 Mb when zipped.

Data will be used by light pollution researchers to study the change of light emissions over time. The data may also be useful for individual participants who live in regions experiencing rapid lighting change.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

We will use DOIs from Zenodo & CKAN URLs.

Search keywords will be provided that optimize possibilities for re-use.

We will produce a unique data set that will be periodically updated. So we will use versioning for the data set. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

We will also include metadata defined by CKAN. This includes:

- Title
- Description
- Tags
- License
- Visibility

- Source
- Version
- Author / Author email
- Maintainer / email
- Custom Fields (adding Zenodo DOI)

2.2. Making data openly accessible

Other than the following restrictions, all data will be open by default.

Restrictions: Email addresses and usernames will never be made public, and will not be stored on Zenodo. Geographic location will be rounded to 3 decimal places in order to partially obscure the true location, while still providing enough geographical precision for analysis.

We will set up a web infrastructure and upload periodically our data to a public repository (i.e. Zenodo). Our data Portal (based on CKAN) will be used for searching but not physically store the data, rather it will keep a link to the public repository.

An ordinary web browser will be enough to download data from our data portal or public repository. Datasets will be in xml. Optionally, the user could use our search facility of our data portal. A reduced dataset of “good” observations is also already available at: http://www.myskyatnight.com

There is open software available to open and process xml data in our software repository on Github (https://github.com/actionprojecteu).

Examples querying and using data will be uploaded to a public repository on Github (https://github.com/actionprojecteu).

Data will be stored in a public repository, our data portal providing search facilities.

Documentation will be available on our website (http://actionproject.eu).

The code will be available on a Github repository (https://github.com/actionprojecteu).

There will be no access restrictions.

CC BY license are both in human and machine readable versions.

2.3. Making data interoperable

The dataset is not interoperable nor can it be re-used between users from different institutions. Therefore we do not need to apply methodologies to make the data interoperable nor to generate project specific ontologies or vocabularies.

2.4. Increase data re-use (through clarifying licences)

Data is licensed with CC BY:

- Share — copy and redistribute the material in any medium or format
- Adapt — remix, transform, and build upon the material for any purpose, even commercially

Data are available about one day after acquiring it at http://www.myskyatnight.com. Data will be also available in a dataset in a public repository after a month.

Data will be used by light pollution researchers to study the change in sky brightness over time, and is expected to be used in scientific publications. However, it will be open to other environmental studies or commercial activities.

There is not a restriction in the use of the data. We only nicely request to cite us using our DOI.

There is no limit in data re-use planned.

Our method for QA will be explained in the documentation. Users will be able to set more or less stringent definitions based on their own analysis of the data.

3. Allocation of resources

The cost of making our data FAIR in public repositories like the open platform Zenodo is free. However, there is a cost on the data infrastructure to be held in the UPM.

Data infrastructure cost during the project will be covered by the grant.

The data management process is handled by UPM (Universidad Politécnica de Madrid) as leader of WP4 (Digital infrastructure for citizen science projects).

4. Data security

With the exception of data uploaded to public repositories, raw data is hosted in servers of GFZ (Germany), NOAO (USA), and UPM (Spain). Only personal authorized can access these servers, and backups are done periodically.

With the exception of data uploaded to public repositories, raw data will only be hosted on the servers mentioned above.

Only personal authorized can access these servers and periodic backups are being done.

5. Ethical aspects

As the data are anonymized and the location is obscured, we see no major issues. In cases when observations are submitted from large private lands (1 ha or greater), one could reasonably presume that the observer may have some association with the location (e.g. living or working at that location).

6. Other issues

None identified so far.

FVB-IGB – Tatort Streetlight data

1. Data Summary

The purpose of the data collection and generation is to scientifically monitor the behaviour of flying insects at street lights and to evaluate the improvements of a new street light design on insect behaviour.

Data collection will include insect behaviour data, such as occurrence of insects around street lights at different positions (at the lamp and in distance to the lamp) and taxonomic identification

Furthermore data will be obtained from our experimental field site and four field sites in communities. At each community an average of 15 traps will be used to monitor aquatic emergence of insects and flying insects (using 6 traps at the streetlights and 3 traps in the water per site and water body). Traps will be activated once per month at all four communities and at the experimental field site. We expect a huge database of regional insect fauna. The insects will be identified to the order and further, when human resources in form of citizen scientists allow the identification to species level. The data will show occurrence of the insect orders at the streetlights.

The data will be useful to scientists and amateur entomologists, to authorities for a broader knowledge of regional insect occurrence, to lighting planners for evidence of better lighting solutions for the environment and to schools for educational purposes. The insect decline is a severe environmental issue, which urgently needs data monitoring. The insect monitoring at the four communities will help to find out more about the occurrence of the local insect fauna.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

All data generated in this project will be uploaded and indexed in the ACTION data portal adding the metadata defined by CKAN. This includes:

Meta data: Date, weather and location of trap collection (including number of trap and collecting person), description of the streetlight; time and location of the identification; anonymized code for the identifying person and contact of the associated group / the supervising coordinator. Measurements of night time brightness, if available.

Scientific data: taxonomic identification, insect image

The datasets are accessible in the Zenodo data portal and can be re-used, since they are published under the cc-by 4.0 license. Furthermore, in the Zenodo data portal DOIs will be generated to archive the datasets and dedicate the identification to the person/group and coordinating supervisor. Based on CKAN, users can make searches based on the common metadata defined for all datasets (see above).

2.2. Making data openly accessible

Data is available in a dataset format through the Zenodo and our data portal.

Examples querying and using data will be uploaded to a public repository on Github (https://github.com/actionprojecteu).

The data can be used by entomologists and light pollution researchers to study the impact of the light pollution on biodiversity. Data is appropriate to be used in scientific publications. Plus, it could be used for environmental studies or commercial activities, e.g. promoting the area for tourism.

The only request for re-using the data is to cite the source using the associated DOI number.

Personal data will be anonymized using identification numbers.

The dataset will be made accessible and re-usable over the website of the German Federal Agency for Nature Protection.

2.3. Making data interoperable

The data will be accessible and interoperable as metadata will be listed accordingly (see above) and scientific names will be used for the insect taxonomy.

2.4. Increase data re-use (through clarifying licences)

The datasets are accessible in the Zenodo data portal/, and can be re-used since they are published with the cc-by 4.0 license.

Our method for QA will be explained in the documentation, SOP will be signed by participants and the methods will be supervised by the local coordinating supervisors as well as by scientists of FV-IGB.

There is no limit in data re-use planned.

3. Allocation of resources

The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held by UPM

Data infrastructure cost during the project will be covered by the grant. After the project is over, the STARS4ALL foundation will cover the costs if enough interest warrants it.

The data management process is handled by UPM (Universidad Politécnica de Madrid) as leader of WP4 (Digital infrastructure for citizen science projects).

4. Data security

With the exception of data uploaded to public repositories, raw data is hosted in servers of FV-IGB (Germany), and UPM (Spain). Only authorized personnel can access these servers, and backups are done periodically.

5. Ethical aspects

Published datasets will not contain personal data. Any personal information will be anonymized with codes. Image taking will only be allowed when the permit is granted by the person, parents/guardians.

6. Other issues

None identified so far.

KCL – Crowdsourcing data

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data is collected across several platforms and experiments, to generate statistics about the use of crowdsourcing tools and optimise their functionality, to study bias in the tasks performed on the tools. It may be in the used long-term to suggest tasks to people who are likely to produce the best results.

What types and formats of data will the project generate/collect?

We will collect:

- Id or the workers in the recruitment platform e.g., Amazon MT. This will allow us to study long term engagement in our platforms. This consists of a single number and does not allow access to additional information.
- We might ask the crowdworker for some general information about them, including nationality, location, gender, and age. This information is used for statistical purposes and not to identify individuals.
- The results produced by crowdworker throughout our platform and the log of their activities in the platform.

The data we collect will be in JSON, CSV, and SQL.

Will you re-use any existing data and how?

No data will be re-used.

What is the origin of the data?

The data is collected through several platforms:

- Qrowdsmith (developed by KCL)
- Virtual City Explorer (developed by KCL)
- Amazon Mechanical Turk

What is the expected size of the data?

The estimate will consist of annotations made by the crowd to given items. In terms of size, it should not exceed a few hundred megabytes.

To whom might it be useful ('data utility')?

The data is primarily useful to the developer of the tools (KCL), and used to refine their functionality and conduct research about users. It may also be useful to other researchers with an interest in crowdsourcing and HCI.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

No

What naming conventions do you follow?

We do not rely on specific naming conventions.

Will search keywords be provided that optimize possibilities for re-use?

We will use the following keywords: crowdsourcing, crowd workers, gamification, human computations.

Do you provide clear version numbers?

We will release the collected data in a format that facilitates a retrievement from IR systems, and if needed we will use versioning.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Where possible we will use metadata to properly describe resources.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

We will share the data obtained from the crowdsourcing experiment we will perform.

How will the data be made accessible (e.g. by deposition in a repository)?

Data will be made available on Zenodo.

What methods or software tools are needed to access the data?

Data will be made available in standard formats, e.g. JSON, or CSV. It can be accessed or manipulated by several open-source software.

Is documentation about the software needed to access the data included?

No, since they are common and widely used.

Is it possible to include the relevant software (e.g. in open source code)?

The source code of the prototypes of Qrowdsmith and the VCE will be available in Zenodo under Apache License 2.0.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

In Zenodo.

Have you explored appropriate arrangements with the identified repository?

Not yet.

If there are restrictions on use, how will access be provided?

At the moment we do not identify any restrictions to the date we will share.

Is there a need for a data access committee?

No.

Are there well described conditions for access (i.e. a machine readable license)?

No.

How will the identity of the person accessing the data be ascertained?

The data will be public, thus there is not need to identify the person who will access it.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Yes, data produced by the crowd will be stored in standard formats.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

We will use standard vocabularies where possible.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

We will use standard vocabularies where possible.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Yes, where possible.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Data will be published under a CC-BY license.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be published after being anonymised and analysed.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

The data will remain available on Zenodo after the project finishes.

How long is it intended that the data remains re-usable?

The data will remain available indefinitely. (?)

Are data quality assurance processes described?

We will assure the quality of the data in two way:

1. During the crowdsourcing task resolution through real-time quality checks. These allow for the identification of crowd workers who do not collaborate genuinely or produce too poor results.
2. After the termination of the tasks by analysing the results produced by the workers. For example, we will examine the agreement among workers and use it as an indicator of the reliability of the data.

3. Allocation of resources

What are the costs for making data FAIR in your project?

Making data FAIR in public repositories like Zenodo is free.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

n/a

Who will be responsible for data management in your project?

KCL as coordinator is responsible for data management. The data storage on Zenodo is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science).

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

n/a

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories, data is kept on KCL servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes, the data is stored on Zenodo.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

All ethics considerations are outlined in deliverable D8.1 – Ethics Requirements.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, all participants have consented to the data collection.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No.

KCL – Interview data

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The interviews are with crowdsourcing and citizen science requesters, to understand what they expect out of participants and which incentives they offer. It will be coded, summarised, and used to produce guidelines and a paper.

What types and formats of data will the project generate/collect?

The experiment will recruit researchers and practitioners who use paid or volunteer crowdsourcing in their work. These participants will be recruited through two mailing lists used heavily by the target audience, each of which is commonly used for recruitment purposes and for which study recruitment is an explicitly permitted usage of the mailing list. The participants will answer a series of questions aimed to determine how and why they choose particular platforms and tools to run their studies, as well as the motivational affordances and incentives which they commonly offer for this work, with a particular focus on the perceived fairness of rewards. The interviews will follow a semi-structured format. Additionally, a second focus of the interview process will be to understand how interview participants' views and opinions reflect best practice and agreed ethical considerations with regard to the use of paid crowdsourcing participants or volunteer participants (such as citizen scientists) who often have little power and may be in a position of vulnerability.

All data will be initially collected as audio recordings (mp3/wav-format), and later transcribed (.doc/.odt-format).

Will you re-use any existing data and how?

No data will be re-used.

What is the origin of the data?

The data is collected by KCL researchers.

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

The data is useful for researchers with an interest in crowdsourcing.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Yes – the transcripts will be assigned a DOI through Zenodo which can be used to find those transcripts.

What naming conventions do you follow?

N/A – since the data represent an accurate transcription of existing audio files, no adjustments will be made to follow a specific naming convention.

Will search keywords be provided that optimize possibilities for re-use?

Yes, keywords will be provided. We will use the following keywords: crowdsourcing, crowd workers, incentives, motivation, requesters, interview.

Do you provide clear version numbers?

No – we do not envisage that multiple versions will be required, as the transcripts will be published only when a complete and accurate representation of each audio file is complete.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

N/A – any metadata will be produced automatically by Zenodo and used for administrative purposes only (e.g., time and date of upload).

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Interview recordings will be destroyed once transcripts are completed. The transcripts will be anonymized, with names and any other sensitive or personally identifying information removed. Only after this is complete will transcripts be made publicly available.

How will the data be made accessible (e.g. by deposition in a repository)?

A Digital Object Identifier (DOI) will be generated using Zenodo to make it findable. This dataset will be findable on the ACTION Zenodo Community (http://zenodo.org/communities/actionprojecteu)

What methods or software tools are needed to access the data?

Document processing software, such as MS Word, Open Office, Adobe Acrobat or a PDF reader.

Is documentation about the software needed to access the data included?

No

Is it possible to include the relevant software (e.g. in open source code)?

No

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data will be made available on Zenodo.

Have you explored appropriate arrangements with the identified repository?

Arrangements have been made for the entire project with the Zenodo community.

If there are restrictions on use, how will access be provided?

There will be no restrictions on the use of the anonymised, published transcripts.

KCL – Interview data

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

We do not envisage any conditions for access – see section 2.4 below.

How will the identity of the person accessing the data be ascertained?

N/A – since there are no restrictions on access, we do not require any confirmation of identity or credentials prior to accessing the data.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

The documents will follow commonly agreed standards for the layout of interview transcripts. No further interoperability issues are foreseen.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

The interview process will follow a standard, semi-structured interview methodology. This process will be briefly summarized within Zenodo, to allow further re-use of the data if required.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

N/A

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

N/A

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Data will be published under a CC-BY license.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be made available with the publication of the research paper they were collected for.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

The data will remain available on Zenodo after the project finishes.

How long is it intended that the data remains re-usable?

The data will remain available indefinitely. (?)

Are data quality assurance processes described?

Yes – this process will involve proofreading and adjusting each transcript multiple times after the initial transcription and this will be explained within Zenodo.

3. Allocation of resources

What are the costs for making data FAIR in your project?

Given the simplicity of the gathered data and given that there is a need to transcribe the audio to facilitate the research, we do not foresee any cost in terms of time or resources. Furthermore, we note that making data FAIR in public repositories like Zenodo is free.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

N/A

Who will be responsible for data management in your project?

KCL as coordinator is responsible for data management. The data storage on Zenodo is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science).

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

We do not foresee any costs for long-term storage of the data.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories, data is kept on KCL servers. Only authorized personnel can access these servers. Periodic backups will be made.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes, the data is stored on Zenodo.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

All ethics considerations are outlined in deliverable D8.1 – Ethics Requirements; the data collection was approved by the KCL ethics committee under ID MRA-19/20-18224

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, all participants have been provided with an information sheet outlining the type of data collected and their use, and have consented to the data collection.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No.

KCL – Open call data

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

These data will be gathered through open call applications, in order to process and proceed with the open call review and accelerator process. The purpose of these data is predominantly to allow us to confirm the validity of the information given, to contact applicants to inform them if they are shortlisted or not, to conduct shortlisting interviews and to complete application paperwork.

What types and formats of data will the project generate/collect?

The data will take the form of names, contact addresses, emails and telephone numbers. This data will be formatted across multiple CSV spreadsheets, designed such that no one spreadsheet contains all of an applicant's personally identifiable information.

Will you re-use any existing data and how?

No data will be re-used.

What is the origin of the data?

This data will be collected from applicants as part of the open call process. The applicants will enter the necessary data into the EasyChair service.

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

Within the project, these data will only be useful for carrying out the open call process. Due to the nature of this data as personally identifiable and the potential for misuse, these data will not be shared beyond the ACTION consortium and will only be shared with those consortium members who will be responsible for contacting and engaging with call applicants - currently KCL as the call managers, and UPM who will manage the call infrastructure. A small subset of the data may be made available to external reviewers of applications.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

No. The data will not be reused nor accessible to those outside of the call process and therefore, there will be no need for the data to be findable.

What naming conventions do you follow?

N/A – No naming convention will be necessary.

Will search keywords be provided that optimize possibilities for re-use?

N/A – No re-use will be necessary or permitted beyond the open call process.

Do you provide clear version numbers?

N/A

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

N/A – No metadata will be produced from the call data.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

None of these data will be made openly accessible. Since all data will be personally identifiable, these data will be used strictly for the open call process only and not made openly available to anyone not involved with the call.

How will the data be made accessible (e.g. by deposition in a repository)?

These data will not be widely accessible. Only those involved with the call will have access to the data, which will be stored on a password protected server, with the credentials known only to those involved with the call.

What methods or software tools are needed to access the data?

Secure log-in credentials will be required to access the server on which the data will be stored.

Is documentation about the software needed to access the data included?

We do not foresee documentation to access the data being necessary.

Is it possible to include the relevant software (e.g. in open source code)?

N/A

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

N/A

Have you explored appropriate arrangements with the identified repository?

N/A

If there are restrictions on use, how will access be provided?

Access will be provided by direct contact with those responsible for running and managing the server.

Is there a need for a data access committee?

No – we do not predict access will be necessary except to those members of the consortium and review board indicated above.

Are there well described conditions for access (i.e. a machine readable license)?

N/A

How will the identity of the person accessing the data be ascertained?

Only those with the necessary log-in credentials will be able to access the data and thus, possession of these credentials will be seen as indicative of identity as an authorised member of the consortium.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

N/A – these data will be personally identifiable and therefore no interoperability will be required or permitted.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

N/A

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

N/A

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

N/A

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

No licensing will be necessary as no re-use of these personally identifiable data will be required or permitted and the data will be destroyed upon completion of the open call process.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

These data will be not be made available for re-use.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

None of the data produced as part of the open call will be useable by third parties. All of these data are personally identifiable and are subject to the European Union General Data Protection Regulation 2016/679.

How long is it intended that the data remains re-usable?

N/A – These data will not be re-usable.

Are data quality assurance processes described?

We do not foresee such processes as being necessary. When completing the initial data submission process, applicants will be asked to confirm the details entered.

3. Allocation of resources

What are the costs for making data FAIR in your project?

KCL – Open call data

We do not foresee any costs, largely as there will be no need to make these particular data FAIR.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

N/A

Who will be responsible for data management in your project?

Data management will be carried out by KCL as the call administrators and UPM as the institution responsible for the data storage infrastructure and technologies.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

N/A – long term preservation will not be required.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

These data will be stored on a secure server. We will restrict the transfer of sensitive data wherever possible. Data will be backed up during the open call application and review process, but this will be brief - approximately 6 months - and after this is complete, these data will be destroyed.

Is the data safely stored in certified repositories for long term preservation and curation?

N/A

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Yes – all of the data gathered during this process will be personally identifiable and therefore sensitive data. Please see deliverable D8.1 – Ethics Requirements, for details.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Participants will be asked for their consent for the use of their data strictly for the purpose of the open call, during the application process. No data sharing and no long term preservation are required for this data.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

These data will be stored in line with the Spanish "Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal" (LOPD)" which incorporates and complies with the requirements of the European Union General Data Protection Regulation 2016/679.

NILU datasets

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The objective of the Norwegian pilot is to raise awareness amongst adolescents about air pollution and its impacts and to provide education. By designing their own measuring campaign, building their own air quality sensor and carrying out air quality measurements themselves, the students will look deeper into the topic and learn more than they would have without the practical exercises.

What types and formats of data will the project generate/collect?

The Norwegian pilot will produce concentration data of $PM_{2.5}$ and PM_{10} in the air. Some students will also collect anonymised GPS data and data of other components, such as temperature and relative humidity, noise or CO_2. The data will be delivered in csv format. We will also collect different Arduino codes the students used for programming their sensors.

Will you re-use any existing data and how?

Some students will use concentration data from the official air quality monitoring stations in the greater Oslo area. Upon request, NILU will provide access to the data and/or share the API.

The students will also use an Arduino code that is already openly available on Github from a previous project (Air:bit).

What is the origin of the data?

Air pollution monitoring

What is the expected size of the data?

Several MB

To whom might it be useful ('data utility')?

To other CS projects to have an Arduino code available and to get an indication of what levels of air pollution have been measured and e.g., to high schools that are planning similar activities.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Yes, data (and associated metadata) will be uploaded to Zenodo and findable via CKAN. As such, each dataset will be given a DOI.

What naming conventions do you follow?

Given the nature of the data (numerical values and location coordinates), we do not predict that specific naming conventions will be required.

Will search keywords be provided that optimize possibilities for re-use?

We use the following keywords: "air pollution", "education", "schools" and "DIY sensor".

Do you provide clear version numbers?

We intend to only make the gathered data available after the student conference, when finalised. We do not therefore foresee that version numbers will be necessary.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Metadata will include the location and time at which the measurements were gathered.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

All datasets can be shared since they are not connected to any personal data.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

These data will be published in the ACTION data portal using CKAN. This will include the data themselves and any associated metadata. Data will be published manually upon completion of the student conference.

What methods or software tools are needed to access the data?

Measurement data can be downloaded as .csv files. Accessing Arduino codes requires Arduino software.

Is documentation about the software needed to access the data included?

N/A – We do not foresee specific documentation being required for the CKAN software, which includes help and guidance for users.

Is it possible to include the relevant software (e.g. in open source code)?

We will include links to relevant programming guidance where possible, but will not upload the software for each individual sensor, due to the number of quality monitoring sensors involved.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

These data will be made available both on the ACTION web pages and Zenodo.

Have you explored appropriate arrangements with the identified repository?

The first data is already available on Zenodo. It will also be made available at the ACTION data portal.

If there are restrictions on use, how will access be provided?

We do not foresee restrictions on data access and use. Data will be openly available.

Is there a need for a data access committee?

No data access committee is required.

Are there well described conditions for access (i.e. a machine readable license)?

The license of the data is CC BY 4.0 International.

How will the identity of the person accessing the data be ascertained?

N/A – No restriction will be necessary and as such, the identity of the person accessing the data is not a concern and does not need clarifying.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Yes, interoperability will be achieved through the use of the common CSV file and through the inclusion of full metadata.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Given the relatively simple nature of the data to be released, we do not anticipate that specific vocabularies or standards will be required.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

As above.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

N/A - we do not foresee specific ontologies or vocabularies being required.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The license of the data is CC BY 4.0 International.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be made available for re-use upon completion of the student conference, when final versions of the data are available .

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Yes - no restrictions are foreseen.

How long is it intended that the data remains re-usable?

No restriction is foreseen in terms of how long data will remain re-usable.

Are data quality assurance processes described?

Data will be provided as is – although quality assurance processes will form part of the data gathering and analysis process that students will carry out, we cannot guarantee the accuracy of the data gathered by sensors.

3. Allocation of resources

What are the costs for making data FAIR in your project?

The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held by UPM.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Data infrastructure costs during the project will be covered by the grant within the scope of WP4 by UPM.

Who will be responsible for data management in your project?

Data management will be carried out by UPM within the scope of WP4 (Digital Infrastructure for Citizen Science Projects).

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

We are actively seeking solutions for the long-term preservation of data – should it prove necessary – for all project pilots within ACTION, including those funded as part of the open call and accelerator process.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

No sensitive data will be stored for this dataset. Data will be stored on secure servers managed and provided by UPM and hosted in Spain. Only authorised personnel will be able to access these servers, using secure login credentials known only to those involved within the ACTION project. All data will be backed up periodically.

Is the data safely stored in certified repositories for long term preservation and curation?

The data will be accessible through Zenodo for long term.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

We do not foresee any legal or ethical issues as a result of this dataset. No personally identifiable or sensitive data will be gathered through these activities. Although these data are being gathered by school students, no data regarding those students will be gathered or released as part of the dataset.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

N/A – no personal data is to be gathered as part of this dataset.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

N/A

SINTEF data

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Analyse how citizen science projects (to what degree they implement data science principles and how they manage data)

What types and formats of data will the project generate/collect?

Collect: data from citizen science projects websites and related sources (mostly websites)

Generate: tabular data / spreadsheets

Will you re-use any existing data and how?

List of citizen science projects from various sources (e.g. Wikipedia)

What is the origin of the data?

Citizen science projects websites and related sources (mostly websites)

What is the expected size of the data?

Order of MB.

To whom might it be useful ('data utility')?

Wider community to understand how citizen science projects deal with data.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Data used: projects URLs

Data produced: will have DOI

What naming conventions do you follow?

DOI

Will search keywords be provided that optimize possibilities for re-use?

Yes: citizen science projects, data management, data science

Do you provide clear version numbers?

Yes.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Date, authors, description of fields in the generated spreadsheet.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

SINTEF data

Data will be made openly available.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?

Deposition in a repository.

What methods or software tools are needed to access the data?

Spreadsheets software

Is documentation about the software needed to access the data included?

No

Is it possible to include the relevant software (e.g. in open source code)?

No

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Repository that supports open access

Have you explored appropriate arrangements with the identified repository?

Not yet

If there are restrictions on use, how will access be provided?

Dataset will be published under CC-BY

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

Dataset will be published under CC-BY

How will the identity of the person accessing the data be ascertained?

n/a

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Yes, spreadsheet / CSV

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Export to CSV

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Probably not.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Probably not.

SINTEF data

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Flexible license.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

During the project.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Yes, useable by 3rd parties

How long is it intended that the data remains re-usable?

Forever

Are data quality assurance processes described?

Yes. They are documented in D4.1

3. Allocation of resources

What are the costs for making data FAIR in your project?

Limited cost.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

From the project budget.

Who will be responsible for data management in your project?

Dumitru Roman

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Not discussed, but data will be freely available.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Data is replicated in various places (Dropbox, Google drive, locally stored).

Is the data safely stored in certified repositories for long term preservation and curation?

Not yet.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

No.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

SINTEF data

No.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No

T6 – Impact assessment data

Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

We are collecting data for research purpose, with the aim of investigating the socio-economic, political and environmental impacts of ACTION pilots and of the overall project

What types and formats of data will the project generate/collect?

We will collect:

- demographic data of respondents such as age, gender, level of education, nationality and income level
- data on participants' opinions
- data on participants behaviors and related changes resulting from ACTION project activities

Will you re-use any existing data and how?

We do not re-use data in this project

What is the origin of the data?

At this moment, we are generating data through semi structured questionnaires: this are distributed online in some cases, used in online interviews or distributed in paper form in face to face events.

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

Data can be useful for ACTION partners in order to improve their activities and for social scientists and researchers from other disciplines interested in citizen science processes and impacts.

FAIR data

Making data findable, including provisions for metadata

The dataset will be uploaded on Zenodo. A Digital Object Identifier (DOI) will be generated using Zenodo to make it findable. This dataset will be findable from our Zenodo Community (http://zenodo.org/communities/actionprojecteu) and from our Data Portal

Will search keywords be provided that optimize possibilities for re-use?

We have identified the following keywords: impact, survey, social, economic, environmental, political

Do you provide clear version numbers?

For this pilot, we will produce a unique data set that will be periodically updated. So we will use versioning for the dataset. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

We will include metadata defined by CKAN. This includes (title, description,tags,license, source, version, author email and some another custom fields)

Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

All the datasets will be available by default.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

There is no provision to keep the data closed, DRIFT will be co-author of this dataset and data processor jointly with T6.

How will the data be made accessible (e.g. by deposition in a repository)?
thought the ACTION open knowledge space, the open data portal developed by ACTION project and based/linked to Zenodo.
What methods or software tools are needed to access the data?

An ordinary web browser will be enough to download data from ACTION data portal or public repository. Datasets will be in CSV. Optionally, the user could use our search facility of our data portal.

Is documentation about the software needed to access the data included?

The API of Zenodo or our data portal can be used to consult the different datasets deposited in these repositories. Plus, there are open software tools available to process datasets

Is it possible to include the relevant software (e.g. in open source code)?

This dataset does not include code

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data will be stored in a public repository, in this case the ACTION open knowledge space providing search facilities. Final analysis of the dataset will be available on ACTION website (http://actionproject.eu).

Have you explored appropriate arrangements with the identified repository?

No.

If there are restrictions on use, how will access be provided?

There will be no access restrictions.

Is there a need for a data access committee?

No

Are there well described conditions for access (i.e. a machine readable license)?

The license of the data is CC BY-NC-SA

How will the identity of the person accessing the data be ascertained?

There is no need to identify the person to access and to download the data selected

Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

No

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

None.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

No.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

We are not planning to use ontologies or vocabularies to describe the data generated.

Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The data will be published with the following license: CC BY-NC-SA

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be available as soon as the dataset is generated an updated following the DoA and internal to the project arrangements timeline

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Data can be re-used. We have identified these communities as interested in using our datasets. ACTION consortium and citizen science projects engaged in project pilots, but potentially also for researchers (especially social scientists) in the area of citizen science and pollution and potentially for reproducibility purposes in the event of academic publishing.

How long is it intended that the data remains re-usable?

Reusability of data is tied to the actual data sources to be catalogued. We expect the data to remain relevant for 2-3 years

Are data quality assurance processes described?

Specific quality assurance processes will not be required as these will largely result from individual respondents being asked to confirm the accuracy of their responses, but the data set will be cleaned in order to eliminate errors, typos or repetitions.

Allocation of resources

What are the costs for making data FAIR in your project?

The cost of making our data FAIR in public repositories like Zenodo is zero. However, there is a cost on the data infrastructure to be held in the UPM

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Data infrastructure cost during the project will be covered by the grant.

Who will be responsible for data management in your project?

The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
See D8.1

Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far

UPM – AZOTEA data

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

AZOTEA is aiming to monitor the evolution of brightness and color of night sky at cities during the coronavirus lockdown period using raw (not JPEG) DSLR images. Citizens take images regularly from their rooftops (=azoteas), upload their images into a NextCloud server and we reduce the data

What types and formats of data will the project generate/collect?

Azotea data & metadata:

- RAW DSLR images, with their EXIF metadata. Relevant to us are:
  - Image timestamp
  - Camera model
  - Focal length
  - F Number (i.e f/2)
- Image data:
  - Image type (LIGHT frame or DARK frame) [metadata]
  - Region of Interest (ROI) around image center (rectangle coordinates) [metadata]
  - ROI Average and stddev in counts. This is the scientific measurement.
- Observer data, to give credit to observation [metadata]:
  - Name
  - Organization (typically a local astronomy club)
  - Optional profile picture for the website.
  - Email (to get access to NextCloud) and contact observers during the lockdown period.
  - Approximate location (location name or neighbourhood in large cities)
- Other data
  - Raw observer image for the AZOTEA web site (a NextCloud repository)

Will you re-use any existing data and how?

No.

What is the origin of the data?

RAW DSLR images of the sky.

What is the expected size of the data?

We expect around 20-30 users at most, taking images over a night spaced by 6-10 minutes, every day. Each image is about 15-30 MBytes. The processed CSV file contains between 50-200 lines per user and night

To whom might it be useful ('data utility')?

Data will be used by light pollution researchers to study the impact of the light pollution in the environment.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)? What naming conventions do you follow?

We will use DOIs from Zenodo.

Will search keywords be provided that optimize possibilities for re-use?

Yes. These keywords will be searchable in Zenodo.
A provisional list will include:

- Public Lightning (Alumbrado Público)
- Citizen Science (Ciencia ciudadana)
- Light Pollution (Alumbrado público)
- Photometry (Fotometria)

Do you provide clear version numbers?

We will produce a unique data set that will be periodically updated. So we will use versioning for the data set. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
For the dataset as a whole, we will also include metadata. This includes:

- Title
- Description
- Tags
- License
- Visibility
- Source
- Version
- Author / Author email
- Maintainer / email
- Custom Fields (adding Zenodo DOI)

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
All the datasets will be available by default.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
There is no provision to keep the data closed and no specific beneficiaries are foreseen.

How will the data be made accessible (e.g. by deposition in a repository)?
We will set up a web infrastructure and upload periodically our data to a public repository (i.e. Zenodo). Our data Portal (based on CKAN) will be used for searching but not physically store the data, rather it will keep a link to the public repository.

What methods or software tools are needed to access the data?
An ordinary web browser will be enough to download data from our data portal or public repository. Datasets will be in CSV. Optionally, the user could use our search facility of ACTION data portal.

Is documentation about the software needed to access the data included?
No. There is available open software available to open and process CSV data.

Is it possible to include the relevant software (e.g. in open source code)?
Yes, the reduction software is open (i.e GitHub).

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
Raw DSLR images will be stored in an UCM-based NextCloud server
Reduced data will be stored in a public repository, our data portal providing search facilities.
Documentation will be available on an UCM webpage https://guaix.ucm.es/azoteaproject).
The pipeline reduction code will be uploaded to our public repository in Github. (https://github.com/actionprojecteu)

Have you explored appropriate arrangements with the identified repository?
No.

If there are restrictions on use, how will access be provided?
There will be no access restrictions.

Is there a need for a data access committee?
No.

Are there well described conditions for access (i.e. a machine readable license)?
Creative Commons license are both in human and machine readable versions.

How will the identity of the person accessing the data be ascertained?
We have no provision to identify persons using the data.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
No as far as we know.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
None.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
No.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
We are not planning to use ontologies or vocabularies to describe the data generated.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?
Data is licensed with CC-BY 4.0 that allows to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material for any purpose, even commercially

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Data will be available as soon as it is uploaded on our servers. An API will be available to query data in real time through an API. Some form of authentication will be required to access our real time API, but Data will be also available as a dataset in a public repository after a month.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
Data will be used by light pollution researchers to study the impact of the light pollution in the environment and to be used in scientific publications. However, it will be open to other environmental studies or commercial activities.
There is not a restriction in the use of the data. We only nicely request to cite us using our DOI.

How long is it intended that the data remains re-usable?
This kind of measurements do not expire, on the contrary it will become a reference for the "normal" activity period.

Are data quality assurance processes described?
Not yet defined at this stage.

3. Allocation of resources

What are the costs for making data FAIR in your project?
The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held in the UPM. The raw images NextCloud server is hosted at facultad de Fisicas-UCM and we don't have associated costs

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project?
The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories and the raw DSLR images, date are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.
Raw DSLR data NextCloud server is only accessible by UCM authorised staff.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories and the raw DSLR images, data are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Personal data in the CSV datasets is deemed necessary to give credits to participating science in research papers. For the same reason, we include a profile photo in the AZOTEA web site.
Email is being used to manage NextCloud account and to contact participants for the campaign.
All these raise a GDPR issue that need to be solved with informed consents.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Yes, informed consent is necessary from the citizens. In addition, we will also include a layman, localized versions of the ACTION clause below

- Recipients of personal data

UCM is bound by an agreement with the ACTION Consortium, a research project operating under the EU Horizon 2020 framework. This data collection is conducted within the scope of the consortium activities, for research purposes as described by art. 89 of the GDPR. It will include several activties for which hereby ask you to give your consent:

[ ] ________________
[ ] survey on motivations in participating in activites
[ ] survey related to the impacts of activites
[ ] ________________

The results of this data collection may be shared with the aforementioned ACTION Consortium as a whole, or with single partners in the consortium, for scientific and research purposes, after an anomymization / pseudononimization process as required from art. 89(1) of the GDPR. This process will be completed by UCM. For any requests regarding this issue, please contact the ACTION Consortium contact at gefion.thuermer@kcl.ac.uk.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far.

UCM – Street Colors

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?
The purpose of data collection is to assess the spectral contents and geographical distribution of light pollution generated by the public lighting systems.

What types and formats of data will the project generate/collect?

Street Colors pilot project (educational project for schools):

- Timestamp, ISO8601 data format [metadata]
- Measured values in 6 bands (450nm, 500nm, 550nm. 570nm, 600nm & 650nm) [data]
- Sensor exposure time [metadata]
- Sensor gain [metadata]
- Sensor temperature [metadata
- Geographical coordinates (longitude, latitude) [metadata]
- Observer.
    - School name & grade, never individual students [metadata]
    - Teacher or educator e-mail (privately held)

Will you re-use any existing data and how?
We have an extensive library of spectral sources used by public lighting systems to serve as reference data to match.

What is the origin of the data?
The origin of data is the public lightning lampposts.

What is the expected size of the data?
We expect around 20 student groups from several at most, each classifying about 10 lampposts. Each measurement will require less than 1 Kbyte.

To whom might it be useful ('data utility')?
Data will be used by light pollution researchers to study the impact of the light pollution in the environment.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
What naming conventions do you follow?
We will use DOIs from Zenodo.

Will search keywords be provided that optimize possibilities for re-use?
Yes. These keywords will be searchable in Zenodo.
A provisional list will include:

- Public Lightning (Alumbrado Público)
- Citizen Science (Ciencia ciudadana)
- Light Pollution (Alumbrado público)
- Spectrum (Espectro)
- Smartphone

Do you provide clear version numbers?
We will produce a unique data set that will be periodically updated. So we will use versioning for the data set. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
For the dataset as a whole, we will also include metadata. This includes:

- Title
- Description

- Tags
- License
- Visibility
- Source
- Version
- Author / Author email
- Maintainer / email
- Custom Fields (adding Zenodo DOI)

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
All the datasets will be available by default.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
There is no provision to keep the data closed and no specific beneficiaries are foreseen.

How will the data be made accessible (e.g. by deposition in a repository)?
We will set up a web infrastructure and upload periodically our data to a public repository (i.e. Zenodo). ACTION data Portal will be used for searching but not physically store the data, rather it will keep a link to the public repository.

What methods or software tools are needed to access the data?
An ordinary web browser will be enough to download data from our data portal or public repository. Datasets will be in CSV. Optionally, the user could use our search facility of our data portal.

Is documentation about the software needed to access the data included?
No. There is open software available to open and process CSV data.

Is it possible to include the relevant software (e.g. in open source code)?
Examples querying and using data will be uploaded to a public repository (i.e GitHub).

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
Data will be stored in a public repository, our data portal providing search facilities.
The documentation for building the device will be published in Instructables, a popular DYO platform. The teaching materials will be available on our website (http://actionproject.eu).
The code will be uploaded to our public repository in Github. (https://github.com/actionprojecteu)

Have you explored appropriate arrangements with the identified repository?
No.

If there are restrictions on use, how will access be provided?
There will be no access restrictions.

Is there a need for a data access committee?
No.

Are there well described conditions for access (i.e. a machine readable license)?
Creative Commons license are both in human and machine readable versions.

How will the identity of the person accessing the data be ascertained?
We have no provision to identify persons using the data.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
Not as far as we know.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

None.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
No.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
We are not planning to use ontologies or vocabularies to describe the data generated.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?
Data is licensed with CC-BY 4.0 that allows to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material for any purpose, even commercially

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Data will be available as soon as it is uploaded on our servers. An API will be available to query data in real time through an API. Some form of authentication will be required to access our real time API, but Data will be also available as a dataset in a public repository after a month.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
Data will be used by light pollution researchers to study the impact of the light pollution in the environment and to be used in scientific publications. However, it will be open to other environmental studies or commercial activities. There is not a restriction in the use of the data. We only nicely request to cite us using our DOI.

How long is it intended that the data remains re-usable?
Reusability of data is tied to the actual light sources to be catalogued. Update campaigns should be carried out to keep our data up to date.

Are data quality assurance processes described?
Not yet defined at this stage.

3. Allocation of resources

What are the costs for making data FAIR in your project?
The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held in the UPM

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
Data infrastructure cost during the project will be covered by the grant. After the project is over, the STARS4ALL foundation will cover the costs if enough interests warrants it

Who will be responsible for data management in your project?
The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?
Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Personal information about users will not be published in our public repositories.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Published datasets will not contain personal data.

The schools enrollment list (teachers & educators) with e-mail contacts will be held in a private spreadsheet, not to be shared nor published. However, we will also include a layman, localized versions of the ACTION clause below, aimed at the teachers involved in the educational activities.

- Recipients of personal data

UCM is bound by an agreement with the ACTION Consortium, a research project operating under the EU Horizon 2020 framework. This data collection is conducted within the scope of the consortium activities, for research purposes as described by art. 89 of the GDPR. It will include several activties for which hereby ask you to give your consent:

[ ] ________________
[ ] survey on motivations in participating in activites
[ ] survey related to the impacts of activites
[ ] ________________

The results of this data collection may be shared with the aforementioned ACTION Consortium as a whole, or with single partners in the consortium, for scientific and research purposes, after an anomymization / pseudononimization process as required from art. 89(1) of the GDPR. This process will be completed by UCM. For any requests regarding this issue, please contact the ACTION Consortium contact at gefion.thuermer@kcl.ac.uk.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far

UCM – Street Spectra

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of data collection is to assess the spectral contents and geographical distribution of light pollution generated by the public lighting systems.

What types and formats of data will the project generate/collect?

Street Spectra pilot project:

- Geolocalized JPEG images as the raw data. Spectral content and lighting source classification will be derived from analyzing the raw data [data]
- Lamp types classified by users [data]
- Lamppost geographical coordinates (latitude + longitude) [metadata]
- mobile phone model [metadata]
- grating model (i.e. “Edmunds Optics 1000 lines/mm linear diffraction grating”) [metadata]
- Timestamp (metadata)
- e-mail (metadata)

Will you re-use any existing data and how?

We have an extensive library of spectral sources used by public lighting systems to serve as reference data to match.

What is the origin of the data?

Data will be the images and classifications done by Street Spectra users through the application developed in this project.

What is the expected size of the data?

We expect around 1000 users, each classifying about 100 lamppost. Image sizes vary between 2MB and 10MB for the Street Spectra pilot, so the expected size will vary between 200GB and 10TB.

To whom might it be useful ('data utility')?

Data will be used by light pollution researchers to study the impact of the light pollution in the environment.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
What naming conventions do you follow?

We will use DOIs from Zenodo.

Will search keywords be provided that optimize possibilities for re-use?

Yes. These keywords will be searchable in Zenodo.
A provisional list will include:

- Public Lightning (Alumbrado Público)
- Citizen Science (Ciencia ciudadana)
- Light Pollution (Alumbrado público)
- Spectrum (Espectro)
- Smartphone

Do you provide clear version numbers?

We will produce a unique data set that will be periodically updated. So we will use versioning for the data set. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

For the dataset as a whole, we will also include metadata. This includes:

- Title
- Description
- Tags
- License

- Visibility
- Source
- Version
- Author / Author email
- Maintainer / email
- Custom Fields (adding Zenodo DOI)

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
The dataset will be available by default.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
There is no provision to keep the data closed and no specific beneficiaries are foreseen.

How will the data be made accessible (e.g. by deposition in a repository)?
We will set up a web infrastructure and upload periodically our data to a public repository (i.e. Zenodo). ACTION data Portal will be used for searching but not physically store the data, rather it will keep a link to the public repository.

What methods or software tools are needed to access the data?
An ordinary web browser will be enough to download data from our data portal or public repository. Datasets will be in CSV. Optionally, the user could use our search facility of our data portal.

Is documentation about the software needed to access the data included?
No. There is available open software available to open and process CSV data.

Is it possible to include the relevant software (e.g. in open source code)?
Examples querying and using data will be uploaded to a public repository (i.e GitHub).

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
Data will be stored in a public repository, our data portal providing search facilities.
Documentation will be available on our website (http://actionproject.eu).
The code will be uploaded to our public repository in Github. (https://github.com/actionprojecteu)

Have you explored appropriate arrangements with the identified repository?
No.

If there are restrictions on use, how will access be provided?
There will be no access restrictions.

Is there a need for a data access committee?
No.

Are there well described conditions for access (i.e. a machine readable license)?
Creative Commons license are both in human and machine readable versions.

How will the identity of the person accessing the data be ascertained?
We have no provision to identify persons using the data.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
Not as far as we know.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
None.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
No.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
We are not planning to use ontologies or vocabularies to describe the data generated.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?
Data is licensed with CC-BY 4.0 that allows to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material for any purpose, even commercially

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Data will be available as soon as it is uploaded on our servers. An API will be available to query data in real time through an API. Some form of authentication will be required to access our real time API, but Data will be also available as a dataset in a public repository after a month.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
Data will be used by light pollution researchers to study the impact of the light pollution in the environment and to be used in scientific publications. However, it will be open to other environmental studies or commercial activities. There is not a restriction in the use of the data. We only nicely request to cite us using our DOI.

How long is it intended that the data remains re-usable?
Reusability of data is tied to the actual lightning sources to be catalogued. Update campaigns should be carried out to keep our data up to date.

Are data quality assurance processes described?
Not yet defined at this stage.

3. Allocation of resources

What are the costs for making data FAIR in your project?
The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held in the UPM

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
Data infrastructure cost during the project will be covered by the grant. After the project is over, the STARS4ALL foundation will cover the costs if enough interests warrants it

Who will be responsible for data management in your project?
The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?
Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?
With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Personal information about users will not be published in our public repositories. All personal information will be previously anonymized.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Published datasets will not contain personal data.

The current version of StreetSpectra pilot gathers data using the EpiCollect V App. The Epicollect back-end do not store any personal information and relies on the user having previously a Google account to authenticate against. Our StreetSpectra Epicollect Form contains a nickname field which the user freely fills-in and it allows to gather observations from a given anonymous citizen. It is theoretically possible that the users may use their real name in nickname field but Epicollect V app do not allow to validate user entries and neither we or the Epicollect platform have way to contact these users.

Our next-generation mobile APP and platform (ACTION servers by UPM) will follow the same idea of not storing authentication data. However, it will include an email field that will be used by the ACTION partners for motivation surveys. This requires us to include in the APP a layman, localized versions of the ACTION clause below in the Terms & Conditions.

Permission will be granted from final users to access third-party systems (OpenStreetMap) to upload lamppost information on their behalf if we develop such functionality later on.

- Recipients of personal data

UCM is bound by an agreement with the ACTION Consortium, a research project operating under the EU Horizon 2020 framework. This data collection is conducted within the scope of the consortium activities, for research purposes as described by art. 89 of the GDPR. It will include several activties for which hereby ask you to give your consent:

[ ] _________________
[ ] survey on motivations in participating in activites
[ ] survey related to the impacts of activites
[ ] _________________

The results of this data collection may be shared with the aforementioned ACTION Consortium as a whole, or with single partners in the consortium, for scientific and research purposes, after an anomymization / pseudononimization process as required from art. 89(1) of the GDPR. This process will be completed by UCM. For any requests regarding this issue, please contact the ACTION Consortium contact at gefion.thuermer@kcl.ac.uk.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far

UPM – DMP Tool (internal)

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of this data is to give access to the tool and organize in projects the data generated by the tool.

What types and formats of data will the project generate/collect?

These are the fields of our data:

- username: string
- password: string (cipher with SHA-1 algorithm)
- project: string, represents the name of the project in which the user participates
- email: string, represents the email of the user.

Will you re-use any existing data and how?

We do not re-use data in this project

What is the origin of the data?

The data is generated by users registering for the tool.

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

We have not identified any community interested in our data

2. FAIR data

2.1 Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
What naming conventions do you follow?

Data will not have metadata and therefore identifiable and locatable because we are not going to publish it.

Will search keywords be provided that optimize possibilities for re-use?

We are not going to publish this dataset so it does not make sense to assign a set of keywords.

Do you provide clear version numbers?

N/A

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. We will also include metadata defined

N/A

2.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Data produced will not be shared and it will not be made openly available. The reason is because this data is used for the proper functioning of the tool and contains personal information. Also, this data is not relevant for researching purposes. In the future, If we want to implement a notification mechanism via email, we will generate a consent form to do it.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

N/A

How will the data be made accessible (e.g. by deposition in a repository)?

This data will not be accessible.

What methods or software tools are needed to access the data?

N/A

Is documentation about the software needed to access the data included?

N/A

Is it possible to include the relevant software (e.g. in open source code)?

N/A

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

N/A

Have you explored appropriate arrangements with the identified repository?

No.

If there are restrictions on use, how will access be provided?

This data is not going to be published, and it will not be accessible.

Is there a need for a data access committee?

N/A

Are there well described conditions for access (i.e. a machine readable license)?

N/A

How will the identity of the person accessing the data be ascertained?

N/A

2.3 Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

No

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

None.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

No.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

We are not planning to use ontologies or vocabularies to describe the data generated.

2.4 Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

Data will not be published

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data is not going to be published.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

None

How long is it intended that the data remains re-usable?

N/A

Are data quality assurance processes described?

Not yet defined at this stage.

3. Allocation of resources

What are the costs for making data FAIR in your project?

Data will be stored in UPM servers.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Data infrastructure cost during the project will be covered by the grant.

Who will be responsible for data management in your project?

The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

This data will be removed after 3 years after the end of the project. Also, it will be able to be removed after the request of a user, as established in the GDPR.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

These data are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

5. Ethical aspects

This data contains personal information although we are not going to share/publish them. However, we plan to implement some notification mechanism in future releases. In that case, we will generate a consent form to use the email address to send messages to the users.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far

UPM – DMP Tool

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The objective of this data is to generate automatically a Data Management Plan following the official template of the H2020 projects.

What types and formats of data will the project generate/collect?

The data collected will contain the responses of the users. These are the elements:

- user: indicates the username of the DMP (string)
- name: indicates the name of the DMP (string)
- project: indicates the project associated to a DMP (text)
- purpose: indicates the purpose of the data collected (text)
- description: indicates the description of the data generated (text)
- reuse: indicates if the user is reusing any previous data (boolean)
- reuse_url: indicates the url of the data reused (string)
- use_data: indicates if the project is collecting data at this moment (boolean)
- use_data_url: indicates the url of the data collected (string)
- interest: indicates if your data can interest to any community (boolean)
- community: indicates the community interested in using your data (string)
- sharing: indicates if the user wants to share its data (string)
- keywords: indicates the list of keywords used to publish the data (list of strings)
- embargo: indicates if the user wants to establish an embargo period (boolean)
- embargo_date: indicates the date when the data is going to be released. (date)
- reason: indicates the reason because the data is not going to be released. (text)
- license: indicates the type of license used for the data (string)
- conditions: indicates the conditions to use the data (text)
- vocabulary: indicates if data is using a vocabulary to describe its fields (boolean)
- vocabulary_text: indicates where and how data is using the vocabulary (text)
- quality: indicates if the user has used any quality assurance procedure for the data (boolean)
- quality_text: it is a description of the quality procedure (text)
- personal: indicates if the data contains personal data
- personal_text: it describes the procedure used to deal with the personal data
- protected_geolocation: indicates if your data contains geographical information that can not be published
- protected_geolocation_text: it describes the procedure used to deal with geographical data

Will you re-use any existing data and how?

We do not re-use data in this project

What is the origin of the data?

At this moment, we are not generating data in our project

What is the expected size of the data?

The estimated size for the dataset is Small (less than 100MB)

To whom might it be useful ('data utility')?

We believe that this data can be interesting for some working groups related with the generation of DMPs and data in general. To be more specific, this data could be used by the “DMP Common Standards WG” of the Research Data Alliance (https://www.rd-alliance.org/groups/dmp-common-standards-wg)

2. FAIR data

2.1 Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
What naming conventions do you follow?

After the embargo period, we will publish the data generated in our Zenodo’s community(http://zenodo.org/communities/actionprojecteu), where a DOI will be generated automatically.

Will search keywords be provided that optimize possibilities for re-use?

We have identified the following keywords: dmp, pollution, citizen science.

Do you provide clear version numbers?

For this tool, we will produce a unique data set that will be periodically updated. So we will use versioning for the dataset. Platforms like Zenodo could provide us the data versioning, so that we do not have to track versions ourselves.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. We will also include metadata defined

We will also include metadata defined by CKAN. This includes (title, description,tags,license, source, version, author email and some another custom fields)

2.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Datasets in our project will not be openly published and shared. Nevertheless, after the date 2022-01-31 will be openly available.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

There is no provision to keep the data closed and no specific beneficiaries are foreseen.

How will the data be made accessible (e.g. by deposition in a repository)?

We will set up a web infrastructure and upload periodically our data to a public repository (i.e. Zenodo). Our data Portal (based on Zenodo) will be used for searching but not physically store the data, rather it will keep a link to the public repository.

What methods or software tools are needed to access the data?

An ordinary web browser will be enough to download data from our data portal or public repository. Datasets will be in CSV. Optionally, the user could use our search facility of our data portal.

Is documentation about the software needed to access the data included?

Our data portal can be used to consult the different datasets deposited in these repositories. Plus, there are open software tools available to process datasets

Is it possible to include the relevant software (e.g. in open source code)?

Examples querying and using data will be uploaded to a public repository (i.e GitHub).

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data will be stored in a public repository, our data portal (https://data.actionproject.eu) providing search facilities. Documentation will be available on our website (http://actionproject.eu). The code will be uploaded to our public repository in Github. (https://github.com/actionprojecteu)

Have you explored appropriate arrangements with the identified repository?

No.

If there are restrictions on use, how will access be provided?

Data have an embargo period. After 2022-01-31, the restriction will be wiped out from the data.

Is there a need for a data access committee?

No.

Are there well described conditions for access (i.e. a machine readable license)?

The license of the data is CC BY 4.0 International

How will the identity of the person accessing the data be ascertained?

It is not contemplated to implement an authenticate process to access the data.

2.3 Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

No

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

None.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

No.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

We are not planning to use ontologies or vocabularies to describe the data generated.

2.4 Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The data have been published with the following license CC-BY 4.0 International

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Data will be available after the date 2022-01-31.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

We think that the community of data management can find our datasets interesting. The data produced in this project can be interesting for sociologists, who can analyze the difference between the answers and the final version of the DMP. Data will be available after the date 2022-01-31.

How long is it intended that the data remains re-usable?

We will publish our data in Zenodo with a license CC-BY 4.0. The data will be able to used as long as it remains published on the platform.

Are data quality assurance processes described?

Not yet defined at this stage.

3. Allocation of resources

What are the costs for making data FAIR in your project?

The cost of making our data FAIR in public repositories like Zenodo is free. However, there is a cost on the data infrastructure to be held in the UPM

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Data infrastructure cost during the project will be covered by the grant.

Who will be responsible for data management in your project?

The data management process is managed by UPM (Universidad Politécnica de Madrid), as leader of WP4 (Digital infrastructure for citizen science)

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Regarding costs, we are seriously considering Zenodo as the platform of choice for preserving our data, since this is an open platform. Our data should be readily available once the quality check by our backend application (or by user validation) is made.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

Is the data safely stored in certified repositories for long term preservation and curation?

With the exception of data uploaded to public repositories, are hosted in UPM's servers. Only personal authorized can access these servers and periodical backups are being done.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

n/a

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

None identified so far