UKEOF
ENVIRONMENTAL OBSERVATION FRAMEWORK

Data Management Planning for Citizen Science

This guidance note provides specific, practical advice to citizen science practitioners (specifically those involved in the planning, collection, storage or use of data) about the development of data management plans to support the value of datasets from citizen science projects. A good Data Management Plan – considering the whole lifecycle of data from its creation and storage to its use, publication and re-use – will add value to citizen science datasets and help them to have greatest possible use and impact.

UKEOF Citizen Science Working Group

Overview

Over the last decade there has been a sharp increase in the use of citizen science to generate data to support and inform scientific research[1]. This has also raised interest in the use of these data as a potential source of evidence to inform public policy and decision making[2]. This trend suggests a growing potential for citizen science to create data that can have a real impact in the way we understand, observe and manage our world and the challenges it faces.

Variation in project design, survey protocols, data models, data quality, and result dissemination mechanisms can result in project data being unpublished or siloed and therefore can be overlooked or rejected by evidence specialists or analysts. A good data plan will ensure that your data are FAIR[3] – meaning that they meet standards of findability, accessibility, interoperability and reusability.

This note builds on other guidance material produced by UKEOF, especially the Advice Notes from the Data Advisory Group[4] and the best practice guides from the Citizen Science Working Group[5].

Why are YOUR data important?

Data are one of the lasting legacies of a citizen science project. When data are properly collated and managed they form part of the historic record detailing the state of our environment – and can even be used to answer questions that were outside of the scope and purpose of the original project. Data should be created, managed and stored effectively to improve the chance that your project has lasting impact.

Citizen science utilises a range of different, often ingenious, survey or data collection techniques. The innovation and creativity of citizen science projects is one of its greatest strengths. Therefore, this guidance is not too prescriptive but rather seeks to help a project organiser work through key data-related issues as a part of the project planning process.

Small improvements in a data management plan can make huge impacts on data use.

The need for a Data Management Plan

In order for your data to have maximum value it should be well conceived, understandable and unambiguous, it should be discoverable (i.e. the users can find it and obtain a copy easily) and it should be consistent and well-structured. These guidelines have been arranged in the form of a checklist to assist in the creation of fit-for-purpose data outputs for your project (Fig. 1).

Data from your project are valuable. To maximise its value, you need to consider the lifecycle of the dataset: how it is created, recorded, stored, assured and, ideally, published. Making best use of data honours the efforts from volunteers and so is a moral responsibility for those developing citizen science projects.

Establish a data management plan at the planning stage of your project. Many of the pitfalls associated with the collection and eventual publication of digital data can be avoided through a well-conceived plan.

Figure 1: The Data Management Plan enables data to be used and reused accurately and efficiently. The Plan spans the life of the project, from inception through to final use and publication

1. Going (or starting out) digital

Digital creation and management of data has multiple advantages over the use of paper records. Even if non-digital media are used to capture data, they will be digitised at some point. We recommend capturing your data digitally from the outset (if possible).

Going digital will greatly improve the consistency, comparability and quality of your data. Capturing data digitally will:

- greatly reduce the possibility of errors creeping into data when it is transcribed into digital form.
- provide the opportunity to control the Data Model (see below), which is the format and range of data being input. When a Data Model is correctly used, this greatly reduces the risk of incomplete or missing records.

Digital also allows for data to be transmitted, or in some cases even published, during the project lifespan and allows data to be viewed, discussed and analysed rapidly and easily.

Digital can also help with engagement and uptake, by making data entry easy for the participant and giving them personalised or real-time feedback about their record. It can permit automated checking of data and rapid feedback to participants about possible errors. Digital can help engage particular target audiences.

2. Technology

Use mobile technology Consider how the use of technology can make data collection easier, more streamlined and even more enjoyable and informative for participants. The widespread use of smart phones and tablets provides a versatile method of collecting information digitally 'in the field'. Costs of creating bespoke mobile 'apps' can be relatively high, so investment in developing apps must consider both good interface design and usability, and also how data are structured and stored. Alternatively, a growing number of free to use apps are available online that enable users to create and configure their own data collection questionnaires, which can then be downloaded on the devices of project participants. This approach will be much cheaper than bespoke app development, and the technology will have been developed by experts and well-tested. However, they may be less cosmetically attractive to users and less versatile for project developers. It is important to ensure that your data model requirements will be met. Examples of platforms that are currently available include: iRecord and iNaturalist (for biodiversity sightings), or Epicollect and Citsci.org (which are more versatile platforms).

Let the technology do the work where you can Smart devices (smartphones and tablets) can undertake a lot of the repetitive tasks associated with data capture on behalf of the users such as recording accurate date/time and location information without the need for manual input. Using the functionality of the technology that you employ to standardise data entry will also reduce the chance of input errors, enhancing its quality and usability. This can be helpful if you adopt a data standard (see below) where the way a measurement or observation is 'encoded' in a dataset is non-intuitive to a user (e.g. participants will not want to manually record the time of their record as "2018-03-08T14:07+0100", but a mobile device can record time in this way automatically).

The ability to capture images, video or audio records can be an incredibly useful function to provide contextual information about the subject matter. Such media can assist with quality assurance exercises or simply to help ascertain the confidence in a record that may be outside of expected patterns. When making this available, consider privacy issues and legislation (see box on data privacy regulations and GDPR requirements, page 5).

However, when designing technology, always remember to consider the usability of the solution – making sure that it is as simple and intuitive as possible.

3. Create a Data Model

A data model visually describes how the 'thing' you are observing and/or measuring is described by your data. In instances where there are complicated processes involved (e.g. multiple datasets or systems that manage your data) a data model can clearly define the relationships and dependencies between the different elements of the model. For projects that 'outsource' technical and IT support, take care to work through your data model with those parties.

The data model should include an understanding of the lifecycle (Fig. 2) of the data you are generating: from collection to curation and preservation[6]. This ensures that the data will be FAIR[3] – meeting the standards of findability, accessibility, interoperability, and reusability.

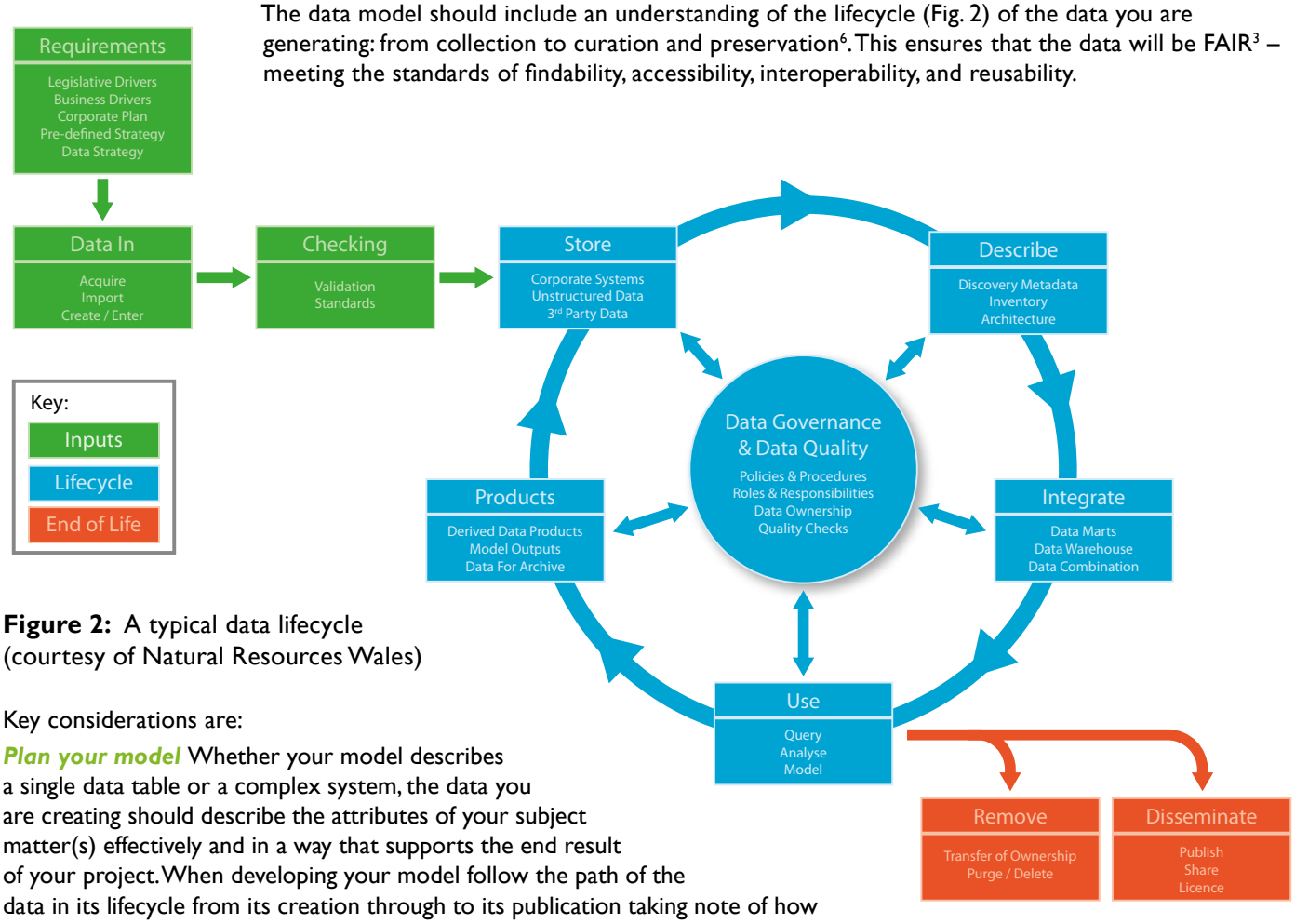

Figure 2: A typical data lifecycle (courtesy of Natural Resources Wales)

Key considerations are:

Plan your model Whether your model describes a single data table or a complex system, the data you are creating should describe the attributes of your subject matter(s) effectively and in a way that supports the end result of your project. When developing your model follow the path of the data in its lifecycle from its creation through to its publication taking note of how the data is to be used, processed, analysed, shared and synthesised.

Consider your audience Your data model needs to be sensitive to the interests and motivations of your data collectors. For example, highly complex survey protocols that require laborious data entry will not appeal to casual users and may impact on the quality and quantity of collected data. When considering your data requirements you need to balance them against what is acceptable for your participants. Consider the pros and cons of technology to assist in data collection tasks (see point 2 above).

Use Data Standards and Terminology Data standards describe the way in which the data are structured (these are like the row headers in a spreadsheet). This is valuable because data that are structured in expected and pre-defined ways are easily understandable by humans and by computers and in turn can be readily incorporated with confidence into others' systems or processes. (See box on DarwinCore[7] for more information.)

An accepted standard provides a useful template that can assist with deciding on which measurement system to use (e.g. feet or meters, lbs or kg) or how to record date/time. This will increase the utility of your data because the meanings and measurements are unambiguous and understandable to users all over the globe. This will increase others' confidence in your data and increase its chances of being re-used.

Use a Controlled Vocabulary A controlled vocabulary enhances data interoperability because it reduces the chance of errors, saves time, and makes it easier to integrate with other data. Also, data warehouses, which are used for sharing data, often require data to be submitted to them in a standardised format. Using data standards from the start of the project ensures that data are (or can easily be) aligned to the desired standard. It is problematic and time-consuming to align data after its creation, particularly if your dataset comprises multiple datasets with different collection techniques. (See box on Controlled Vocabularies for more information.)

Consistency Once you have defined a data model, stick to it to support the integrity and consistency of the data throughout the lifespan of the project. This should extend to all aspects of the data model that may impact on data's structure and content. If changes are proposed be sure to consider the potential implications of the changes on the integrity of the data and ensure that the changes are documented.

Protection & Security Your data model must consider how it manages personal or sensitive information. Be very clear what information you are collecting and storing about your participants within the scope of your project. Sometimes this will be obvious, such as a name or address or other contact information, but it could also be less obvious such as storing location information about a person's whereabouts at a certain point in time or digital records (photos, video or sound files) of a person. Any collection of personal information must comply with the General Data Protection Regulations, GDPR (see box on data privacy regulations and GDPR requirements for more information). Analysis of the participants could be an important use of the data, but be aware that ethical approval may be required[9].

It may be valuable to record and share the recorder's name. This allows the recorder to be appropriately credited and is often important for biological recording data, e.g. in iRecord, although this must be stated in the privacy notice for the project. Whether a name is associated with a record or not, it is good practice to make sure that personal data can be separated from the scientific data. Ideally this should be done early in the data life cycle so that unique anonymised codes link the survey data to the personal data. One option to doing this is to attach Universally Unique Identifiers (UUIDs) to each data point. UUIDs can be line numbers or can be randomly generated codes[11] that allow each data point to be uniquely identified, and so can provide a unique link between the public and sensitive components of a data point. The digital data collection platform can be designed so that UUIDs are created at the point of data collection.

Once public and sensitive components of the data are separated, you can:

- Freely share the survey data with anonymised recorder identity.
- Store personal data securely, so that you know who has access to the data.
- Consider blurring location data, so that specific locations (e.g. a person's garden) cannot be identified.

DarwinCore as an example of a data standard

DarwinCore[7] is a standard for biodiversity information, and is being extended to be used for other environmental data. It is used throughout the world. DarwinCore is organised with classes of information, e.g. 'Organism' or 'Location'.

There are properties within the classes, e.g. the class 'Location' has properties such as 'locationID', 'verbatimLocality' 'country', 'decimalLatitude' and so on. Each property has a description of the possible values. For example:

- 'decimalLatitude' can take numeric values between -90 and +90 inclusive
- 'individualCount' can take whole numbers from 0 upwards
- 'organismName' is a textual name
- 'eventDate' is a date-time
- A measurement, such as tree height, can be unambiguously described with the combination of 'sampleSizeValue' = 15.5 and 'sampleSizeUnit' = 'metre'

Dates can be particularly awkward: is "03-04-2018" the 3rd April or the 4th March? Best practice is to use the form YYYY-MM-DD (this conforms to the standard ISO 8601:2004(E)).

Therefore if you state that your dataset conforms to a particular standard, everyone will be able to understand your data accurately and without confusion. Further guidance is available[8].

Controlled Vocabularies

Digital data input and technology facilitate the use of a Controlled Vocabulary.

Selecting from lists ensures consistency, e.g. was the deer: ⊙ female ○ male ? This avoids people recording: "female", "fem.", "F", "doe" and so on.

Consistency can be 'hard coded' into the data collection platform to confirm to an accepted data standard, e.g. the user enters a value: "What was the height in metres? 2.6", but the database records 'sampleSizeValue' = 2.6 and automatically adds 'sampleSizeUnit' = 'metre'.

Data privacy regulations and GDPR requirements

In the UK it is essential to abide by the Data Protection Act 2018, which is the UK's implementation of the European Union's GDPR (General Data Protection Regulations). If you are running a citizen science project or using its data in the UK or anywhere in the EU, and personal data is collected, you are classed as a 'controller' or 'processor'. Personal data is anything that allows a living person to be identified (including name, address and their computer's IP address) and sensitive data includes many categories of additional information about the person.

- You must have valid grounds for collecting and using personal data. You should limit yourself to collect only what is adequate, relevant and necessary.
- You must be open and honest about the use of the personal data. Make sure that your policies at sign-up are clear and comprehensive.
- You must ensure that you store and process personal data securely.

Speak to an expert to discuss to discuss your GDPR requirements in advance of setting up your project and also see the ICO guidance[10].

Openness "Open data is data that anyone can access, use or share"[11]. It is good practice for citizen science data to be open data, where possible; this allows data to be viewed, used, shared and re-used without restrictions imposed from licensing or copyright. In situations that require rapid sharing of data – such as in alert systems for invasive, non-native species – lack of restrictions on sharing data is essential.

- Consider how data can be made open, without contravening the principles of appropriate privacy and security (see sections above). Often it will be best to remove all personal (and individually-identifying) information from the data, but sometimes, and where the project's privacy policy permits this, a person's name should be shared with the record.
- Consider the degree of openness that is possible. An Open Government Licence[12] is best for making public sector information openly available. Alternatively, Creative Commons licenses makes it clear from the point of submission that the data will be open data. CC0 or CC BY licences are best. Be aware that putting a 'non-commercial' restriction on the use of the data (e.g. CC BY-NC) can inadvertently restrict its use.
- Do not create restrictions on re-use inadvertently by including Intellectual Property from third parties. This could be using background mapping products, use of imagery or even using an analytical model or technique created by others.
- Plan in advance for any potential conflicts, especially where spatial data could be contentious e.g. the location of rare species or sampling locations on private land. You could allow blurred location data to be openly viewed and shared, but only share high precision data with authorised users.

4 Data Quality

From the start of the project, consider the quality of your data, both in terms of the accuracy of the data points and the quality of the dataset. Data quality has to be considered in context – are the data suitable for their intended use (and potential re-use)?

- Accuracy: is accuracy known (e.g. the range of measurement errors could be determined from pilot studies, or accuracy could be determined from verifiable records with photographs included), and is accuracy sufficient? Consider measures you can employ to maintain quality levels – such as training, support or guidance documents. Can you provide automated checks for quality, e.g. through appropriate use of methods and technology?
- Completeness: are values missing?
- Validity: does the data match the rules?
- Uniqueness: is there duplicated data?
- Consistency: is the data consistent across various data stores?
- Timeliness: does the data represent reality from the required point in time?
- Information content of the dataset: is the dataset be sufficient to address your needs? Consider whether your project design,the motivations or location of participants might encourage over or under-representation of features you are interested in and whether these can be accounted for in analysis.

It is important to consider quality control of the dataset. Some quality issues (e.g. uniformity of measurements and use of key terms) can be managed to some degree with technology but some factors are harder to control (e.g. the quality of the observation or judgement in the field). Also record if the data has undergone any quality assurance checks or corrective work. This gives the user (be it a scientist, student, land manager or decision maker) confidence to use the data in an appropriate way.

5 Data Publishing

Plan for what will happen to the data created by your project, so that it can easily be shared and re-used. This maximises the value gained through the volunteers' participation. Plan for the time and cost required to publish the data.

- Share data with repositories of similar data e.g. records of species occurrences should be shared with the National Biodiversity Network Atlas in the UK[16] or The Global Biodiversity Information Facility (GBIF)[17]. Marine data is coordinated in the UK by the Marine data Information Network (MEDIN)[18] which has a data archive structure based on data themes. The archives can provide help on data design, archiving and routes to sharing data.
- Find a suitable data repository where the full dataset and metadata can be securely stored and made accessible to others, such as the Environmental Information Data Centre (EIDC: http://eidc.ceh.ac.uk/), Dryad (https://datadryad.org/) or Zenodo (https://zenodo.org/). Many repositories provide a Digital Object Identifier (DOI)[19], giving your dataset a persistent reference. The Registry of Research Data Repositories (re3data.org) lists additional data repositories.
- At the very least, data should be available for sharing via a file sharing facility where requests for data can be handled manually.
- Publication of data is an important recognition of the benefit given by volunteers and it can help people feel valued and so remain engaged for future activities.

6 Metadata

Metadata gives information about a dataset. This includes aspects of the method of data collection (e.g. the data collection protocol or the quality assurance of the data).

In simple terms, a range of values organised into a table or viewed as a scatter of points on a map are not meaningful unless they are placed within the context of what they represent, how and why they were collected and what the data content means. Good metadata allows the context of the data to be documented and therefore increases the chances that your data is used appropriately by others.

Metadata usually conforms to data standards, but can be descriptive in nature, e.g. the 'Methods' metadata for a dataset could be a long paragraph describing the protocol for site selection, data collection, measurement techniques and quality assurance of the data. Metadata provides a potential user of your data with enough information to ensure that your data can be used appropriately and with confidence. When creating metadata be sure to be accurate and comprehensive.

The type of metadata is likely to be decided by the nature of the data or subject matter your project involves. For example you may wish to use Ecological Metadata Language (EML)[14] for ecological datasets or GEMINI2[15] for geospatial datasets.

If you are publishing data to an online data storage facility there may be a predefined metadata standard or template required to be compatible with their systems. Good, comprehensive metadata will also increase the chances of your data being 'discovered' through search functions. It is valuable for metadata to be compliant with a standard (i.e. where compliance can be achieved by populating key fields), but good metadata is also informative and comprehensive.

Overall, while creating metadata can be complicated, the most important aspect is that metadata allows other people to understand and appropriately re-use your data, so avoiding mistakes that can come from uninformed re-use of data.

Further reading

- UKEOF guides on data:
  - The principles of planning, collecting and using citizen science data http://www.ukeof.org.uk/documents/DataAdviceNote2_ single.pdf
  - The principles of good data and information management http://www.ukeof.org.uk/documents/DataAdviceNote1_single.pdf
  - Data Governance http://www.ukeof.org.uk/documents/DataAdviceNote3_single.pdf
- Open data licenses: http://opendefinition.org/licenses/

REFERENCES

1. Haklay, M., 2015. Citizen Science and Policy: A European Perspective, Washington DC: Woodrow Wilson Centre & Pocock et al. (2017) The diversity and evolution of ecological and environmental citizen science. PLOS ONE 12, e0172579. https://doi.org/10.1371/journal.pone.017
2. Higgins 2016 https://sciencenode.org/feature/is-citizen-science-living-up-to-the-standard.php
3. Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. https://doi.org/10.1038/sdata.2016.18
4. http://www.ukeof.org.uk/our-work/data-advisory-group
5. http://www.ukeof.org.uk/our-work/citizen-science
6. http://www.dcc.ac.uk/resources/curation-lifecycle-model
7. http://rs.tdwg.org/dwc/
8. https://docs.nbnatlas.org/share-species-occurrence-records-with-the-nbn-atlas/
9. http://www.ethicsguidebook.ac.uk/
10. https://ico.org.uk/for-organisations/guide-to-data-protection/
11. https://en.wikipedia.org/wiki/Universally_unique_identifier
12. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
13. http://theodi.org/what-is-open-data
14. https://knb.ecoinformatics.org/#tools/eml
15. https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini/40-gemini/1037-uk-gemini-standard-and-inspire-implementing-rules
16. https://nbnatlas.org
17. https://www.gbif.org
18. https://www.medin.org.uk/
19. https://en.wikipedia.org/wiki/Digital_object_identifier

AUTHORS

Colin Chapman (Welsh Government), with input from Michael Pocock (UKEOF & UKCEH), Jon Parr (MBA), Mike Brown (UKEOF & UKCEH), Barnaby Letheren (Natural Resources Wales) and the UKEOF Citizen Science Working Group

UKEOF works to improve coordination of the observational evidence needed to understand and manage the changing natural environment. It is a partnership of public sector organisations with an interest in using and providing evidence from environmental observations. Contact us at office@ukeof.org.uk

UKEOF
ENVIRONMENTAL OBSERVATION FRAMEWORK

UKEOF PARTNERS

British Geological Survey; UK Centre for Ecology & Hydrology; Department for Agriculture, Environment and Rural Affairs (Northern Ireland); Department for Business, Energy and Industrial Strategy; Department for Environment, Food and Rural Affairs; Economic and Social Research Council; Environment Agency; Forestry Commission; Joint Nature Conservation Committee; Met Office; Natural England; Natural Environment Research Council; Natural Resources Wales; Office of National Statistics; Scottish Environment Protection Agency; Scottish Government; Scottish Natural Heritage; UK Space Agency; Welsh Government.

Image credits: *Cover:* © Michael Pocock. Sampling water quality with a visiting member of the Water Warriors citizen science team from University of Malaya*: Page 3:* Paul Fisher © UKCEH; *Page 6:* Annie Spratt via Unsplash.com; *Page 7:* Andrew Sier © UKCEH

www.ukeof.org.uk

@UKEnvObs

Published February 2020