NASA ESDS Citizen Science Data Working Group White Paper

Version 1.0 – 24 April 2020

Introduction

United States federal agencies have a long history of utilizing citizen science and crowdsourcing to supplement data collection in areas spanning a variety of scientific disciplines, from earth and planetary science, to public health and medicine, to disaster response. Legislation, such as the Crowdsourcing and Citizen Science Act (15 USC Section 3724), reinforces use of citizen science by Federal agencies, acknowledging such potential benefits as accelerating scientific research, addressing societal needs, providing hands-on STEM learning, and connecting members of the public to science agency missions. Realization of these benefits is directly related to the public's awareness of these participation opportunities and the ability to locate and utilize the data resulting from them.

The National Aeronautics and Space Administration (NASA) has incorporated the use of citizen science and crowdsourcing in numerous science projects within the astrophysics, heliophysics, earth, and planetary science fields over the last quarter century. Programs such as NASA's Research Opportunities in Space and Earth Sciences (ROSES) welcome investigators to incorporate citizen science into their research. In an effort to maximize the potential benefits of this research, this document presents a series of recommendations and best practices for data handling to researchers desiring to collaborate with NASA. The goals of these practices are to provide data that are Findable/discoverable, Accessible, Interoperable with complementary data sets, and Reusable (FAIR). NASA policies and legal considerations for data access, provisioning, and attribution are also included.

This document provides guidelines for legal, policy, and ethical issues; standards for citizen science data collection and management; information on ensuring usability of citizen science data and communication regarding its use; and best practices for long-term archival of citizen science data. Section 1 contains a detailed discussion of policy, ethical, and legal considerations influencing citizen science data collection. Section 2 considers standards for documentation, including documentation of instrumentation, procedures, and the data itself. It concludes with a discussion of how citizen science data should be attributed. Section 3 provides guidance about how to ensure citizen science data are collected and stored in a useable way. It also considers how NASA and data producers should notify the scientific community, including citizen scientists and the public, about citizen science datasets and the scientific conclusions reached using them. Finally, Section 4 provides detailed information regarding what should be archived from projects using a citizen science approach, including data and code. It provides guidance about archive location, process, and timeframe, as well as information about data access and distribution services provided by NASA that may be relevant to data producers working with citizen scientists.

Because this document cannot answer every question about working with citizen scientists, funded NASA investigators or those submitting proposals are urged to contact relevant NASA program managers with additional questions.

Table of Contents

1 Policy, Ethics, and Legal Issues

This section provides guidelines about standards relating to Earth Science citizen science data. Included guidelines address: policy, ethics, and legal considerations. The guidelines presented reflect best practices assembled by practitioners of NASA-funded Earth Science citizen science programs/projects and members of the NASA ESDS community. The guidelines are intended for use by pre-proposal and post-award data producers and providers.

1.1 Privacy Adherence to NASA Open Data Policy

NASA promotes full and open sharing of all data with the research and applications communities, private industry, academia, and general public. NASA's Open Data Policy (Data and Information Policy) is available at: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy.

The open data access policy applies to all data sources, including satellites, airborne and in-situ sensors, and data supplied by citizen scientists, as well as all supporting metadata, algorithms, source code, and documentation. All Projects/Proposers/Investigators seeking collaboration with NASA must agree to this Open Data Policy. Each project should create a Data Management Plan that facilitates conformance with NASA's data policy principles.

1.2 Privacy and Personal Information Concerns

Privacy policies generally follow NASA's web privacy policies (also see NASA Privacy Policy) unless otherwise noted. The basic premise is that the submission of information by a participant is strictly voluntary. When a participant submits data to a project, the individual gives the project permission to use the data in ways consistent with the stated purpose of the project. Individuals unwilling to grant this permission should refrain from submitting data.

If the project provides a mobile app in addition to a website, the mobile application is required to have a separate privacy policy (https://www.dhs.gov/sites/default/files/publications/047-01-003.pdf). The privacy policy disclosed to the participant should be easily accessible and available within the application itself. Elements to address in the privacy policy may include the collection, use, sharing, disclosure, protection, and retention of Personal Identifiable Information (PII), Sensitive Personal Identifiable Information (SPII), and sensitive content. Sensitive content is information that is not PII but raises privacy concerns, such as mobile device ID or metadata.

If PII is collected, a notification statement to the participant should be presented at the point of collection. In addition, any changes to the collection, handling, or use of PII (e.g., access to sensitive content or tools on the mobile device) resulting from an app update or new release should be described in a contextual notice to the participant upon opening the updated app. The participant must provide an affirmative express consent. Participants should be granted the ability to opt out of features within the application, where appropriate (e.g., opting out of using certain sensors while choosing to utilize other application sensors).

SPII is information which, if lost, compromised, or disclosed without authorization could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual (Instruction 047-01-003 Privacy Policy for DHS Mobile Applications).

SPII includes:

- Social security number (including portions thereof)
- Government ID number
- Passport number
- Alien registration number
- Financial account information

- Biometric identifiers (including photographs with recognizable features of individuals)
- Full date of birth
- Parent's maiden, first, or middle name
- Legal and law enforcement records
- Educational information
- Performance ratings/appraisals
- Training records
- Any of the above about the individual's spouse, partner, children

It is strongly recommended that citizen science projects do not request or store SPII. Specific needs for the collection and storage of SPII should be reviewed for compliance with regulation, policy and IT storage best practices to ensure its protection, and the distribution of that information should be very limited.

Combinations of PII, SPII, and sensitive content that are linkable with the individual in order to specifically identify the individual with the data should also not be externally disseminated with the data collected by citizen scientists without their consent. The Department of Homeland Security's Handbook for Safeguarding Sensitive PII details policies and best practices for the collection and management of PII.

A citizen scientist should be allowed the option to opt in/out of correspondence, such as news, alerts, or events. For that entity it should be understood that correspondence related to doing business and quality control is necessary. Therefore, the citizen scientists may be asked to provide some PII, such as their name, telephone number, physical address, and/or email address to that entity. However, that information may not be shared with other data users.

The use of data contributed by citizen scientists are in the public interest. A citizen scientist should be able to correct or delete the data they contributed to the original entity in the event it was erroneously entered. However, the citizen scientists are asked not to remove or alter valid data they contributed, as it could have an impact on public health, public interest, historical research, or scientific research (GDPR, 2018).

1.2.1 Data Obfuscation

Data obfuscation is the practice of modifying data inputs or data outputs to disguise the data in some fashion. Data obfuscation may include scrambling stored data through a form of encryption or reducing the accuracy of data to minimize attribution. PII, and especially SPII, should be candidates for data obfuscation through encryption. Location data may be obfuscated to prevent data users from knowing the exact location of the data source. Children's Online Privacy Protection Rule (COPPA) regulations specify location obfuscation required if data collected is associated with children. Another example may include obfuscating the specific location of endangered species, nesting sites, or other sensitive areas. Participant information can be obfuscated with randomly assigned User IDs not directly traceable to specific participants. In general, accurate science data collection needs must be balanced with regulation and policy requirements, as well as participant privacy needs.

1.3 Liability Clauses

The United States Code is a consolidation and codification by subject matter of the general and permanent laws of the United States. It is prepared by the Office of the Law Revision Counsel of the United States House of Representatives and is available at: https://uscode.house.gov/browse.xhtml. Section 3725 of Title 15 Commerce and Trade, Chapter

63 Technology Innovation details laws applicable to the use of crowdsourcing and citizen science by the federal government. Part d8 under Section 3725 describes the Liability policy as:

> Each participant in a crowdsourcing or citizen science project under this section shall agree -
>
> (A) To assume any and all risks associated with such participation; and
>
> (B) To waive all claims against the Federal Government and its related entities, except for claims based on willful misconduct, for any injury, death, damage, or loss of property, revenue, or profits (whether direct, indirect, or consequential) arising from participation in the project.

Projects should reflect this liability policy in the Terms of Use such that all citizen scientists are made aware before agreeing to participate in the project.

In addition, as part of the Terms of Use agreement, projects should inform participants of any risks that may be encountered through project participation. For example, Figure 1 illustrates the Safety Statement for the GLOBE Observer mobile app that participants must acknowledge before they can use the app:

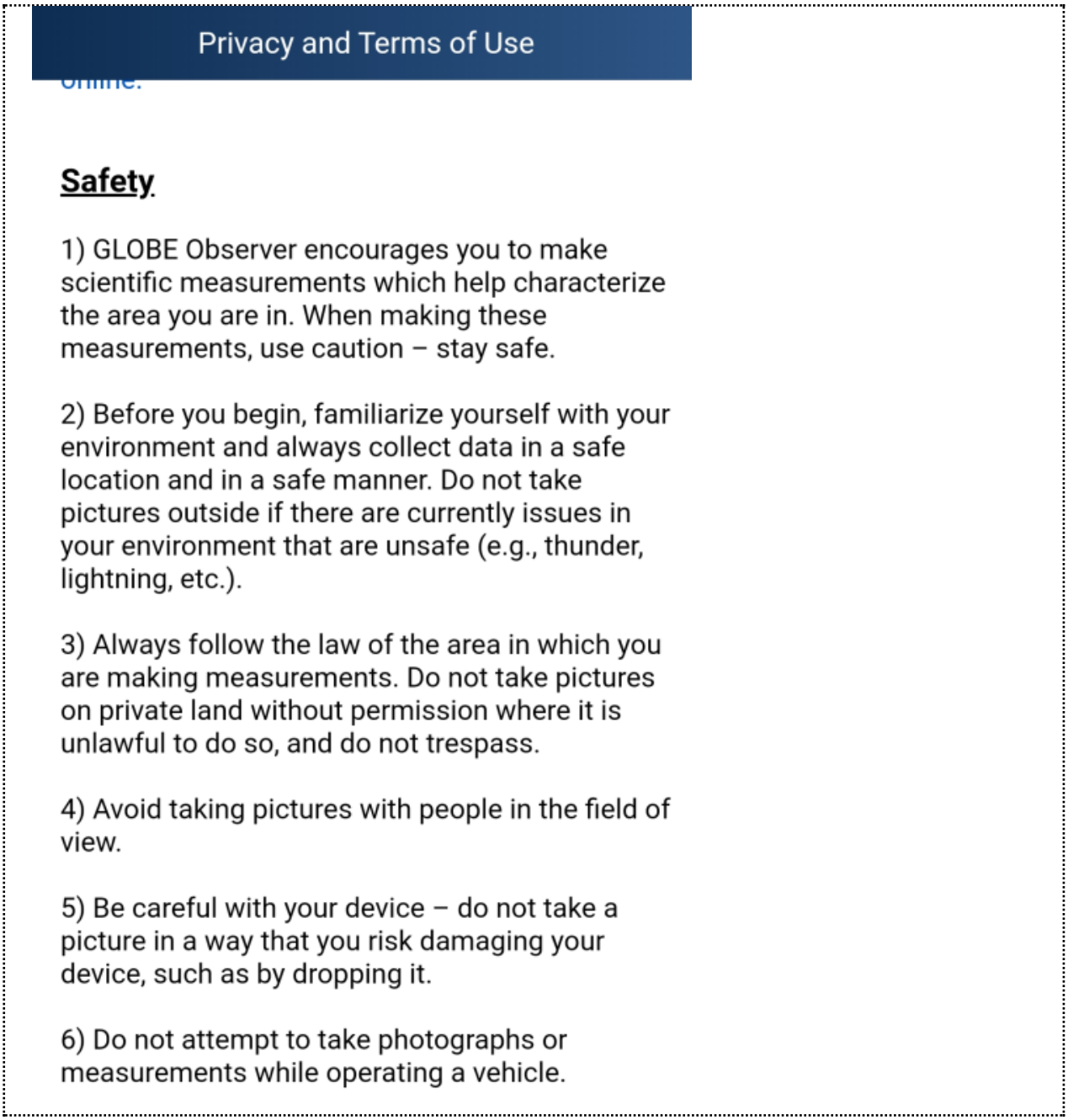


Figure 1: The GLOBE Observer Safety Statement

1.4 Data Ownership

As part of the project's consent, registration, and Terms of Use processes, policies regarding data ownership should be clearly stated. Aspects of data ownership include: ownership of individual data observations submitted, including images and photos; ownership of full data set.

By default, unless a project requires copyright assignment as a condition of volunteering, a citizen scientist retains ownership of their submitted works (e.g., photos) and could refuse to grant permission for NASA to publish the works. This has the potential to disrupt science processing or prevent the dissemination of data and findings. To avoid such disruption, NASA

recommends projects use “Creative Commons” licenses, which allow creators to retain their individual copyrights while permitting NASA to use their works. Item (2) of NASA’s Landslide Reporter Project’s “Contributor License Agreement” is an example that reflects the Creative Commons principles:

> 2. Contributor Grant of Copyright License. Subject to the terms and conditions of this Agreement, You hereby grant to NASA and to Recipients of the Cooperative Open Online Landslide Repository (COOLR), which includes Landslide Reporter and Landslide Viewer, distributed by NASA as a perpetual, non-exclusive, worldwide, royalty-free, irrevocable (except as stated in Section 4) copyright license to use, distribute, reproduce, modify, redistribute, prepare Derivative Works of, publicly display, publicly perform, and sublicense Your Contributions and such Distributed Works.

1.4.1 Individual Observations

Under Creative Commons, individual citizen scientists retain the rights to the data they submit to the project. Projects must determine and clarify how/if citizen scientists may access their individual observations after they have been submitted to the project’s data store.

1.4.2 Full Data Set

A NASA citizen science project should claim ownership of the database containing all measurements submitted to it. Projects must clarify that the data in the database are openly and freely available to the public in accordance with NASA’s Open Data Policy (Section 1.1).

1.4.3 Photographs and Images

Should a project decide to use a submitted image or photograph in a publication or social media posting, it is required to treat such images as NASA treats all third-party created/credited content by giving proper credit and other elements that make it clear this is not a NASA-produced image. This scenario is already covered under the media usage guidelines for the agency that say "NASA occasionally uses copyrighted material by permission on its website. Those images will be marked copyright with the name of the copyright holder. NASA's use does not convey any rights to others to use the same material. Those wishing to use copyrighted material must contact the copyright holder directly" (Daines, 2015).

In many cases, the identity of the image’s owner may be unknown. Where the identity of the contributor is known, NASA requires that images be credited using the following recommended format:

____type of thing____ by ____person’s name____ (___license/copyright status___) based on ___data/images/video/thing___ provided courtesy of NASA/____other credits from the source____

An example citation:

Enhanced Image by John Doe (CC BY-NC-SA) based on images provided courtesy of NASA/JPL-Caltech/SwRI/MSSS

Images used in social media posts must also credit the owner and link to the full image. However, projects should not reference personal @handles for social media nor provide links to personal webpages.

1.5 Legal Policies

Federally funded projects utilizing crowdsourcing and citizen science data must comply with a number of laws and policies levied upon Federal Agencies. Projects should clearly communicate these policies to citizen science participants, establish mechanisms to ensure accountability to these policies, and provide audit mechanisms that verify these policies are reflected in the project's actual implementation. The following sections supply a short summary of each law or policy, along with guidance for compliance.

1.5.1 Children's Online Privacy Protection Act

COPPA regulates the collection, storage, maintenance, and use of personally identifiable information (PII) with regards to children under the age of 13. Geolocation data, collected by most citizen science projects, is one of several items that is considered PII. Collecting such information may only take place after receiving parental consent. Refer to the Federal Trade Commission guide to COPPA compliance: http://www.coppa.org/comply.htm.

Projects are highly discouraged from collecting such data, and in most cases this type of information will not be required. However, projects should maintain awareness of how this information may be collected if the project is used by a school, scout troop, etc. where children under 13 could be participants.

1.5.2 Paperwork Reduction Act

NASA has obtained an overarching PRA clearance from the Office of Management and Budget (OMB) for citizen science activities under the assigned control number 2700-0168. New projects are required to submit supporting documentation to have information collection approved under this overarching PRA.

1.5.3 Information Quality Act

NASA's implementation of the Information Quality Act (IQA) is available on the website for NASA's Office of the Chief Information Officer: https://www.nasa.gov/content/nasa-guidelines-for-quality-of-information. Information exempt from this policy includes information disseminated by NASA but not authored by NASA nor adopted as NASA's views. This includes information communicated by scientists and researchers via the "academic process," as defined in NASA's IQA Guidelines document. Review this document to ensure compliance with the policy.

1.5.4 Antideficiency Act

The goal of the Antideficiency Act is to control federal spending by limiting the ability of federal agencies to create financial obligations in advance of or in excess of appropriations. Included in this act is a restriction on the use of volunteers. However, as stated in the Wilson Center's report "Crowdsourcing, Citizen Science, and the Law, Legal Issues Affecting Federal Agencies":

> A well-planned, narrowly defined crowdsourcing activity that includes a written waiver of compensation signed/acknowledged by the volunteers seems unlikely to violate the antideficiency act.

Terms of Use should clearly state that citizen scientists are not compensated for their participation.

1.5.5 Accessibility/Section 508

NASA Citizen Science projects are required to make their websites accessible ("Section 508" compliant), and to the greatest extent possible, any apps developed to support the project must also be compliant. NASA's implementation of Section 508 is described at: Section 508 Standards. Citizen Science Project leads are encouraged to contact the Agency coordinator (contact information is located on the Section 508 website) with questions regarding accessibility.

1.5.6 Terms of Service for Mobile Apps

Projects developing a mobile app for data collection and display must have it approved by NASA's Strategic Partnership Office (ITPO) before it can be released to any commercial distribution platform, such as the Apple Store (for iOS devices) or Google Play (for Android devices). The Strategic Partnership Office ensures that the Terms of Service for the app distribution platform(s) are specific to NASA/federal agencies, as the standard Terms of Service may be in conflict with federal law.

The multi-step process for releasing a NASA app includes assessing the app for Section 508 compliance (Section 1.5.5), completing a Global Concerns Statement, and an Export Control Disclosure form.

1.5.7 Adherence to Local and International Laws

In addition to complying with Federal law, NASA Citizen Science projects should also ensure they are in compliance with all applicable international, state, and local laws. The Harvard Law Clinic provides a state-by-state listing that addresses relevant topics, including:

- Trespassing
- Loitering
- Stalking
- Privacy
- Drone use
- Critical infrastructure
- Agency regulations

The European Citizen Science Association (https://ecsa.citizen science.net/) provides numerous resources related to its six components of Responsible Research and Innovation (RRI): Governance, Science Education, Ethics, Open Access, Gender, and Public Engagement. Citizen Science Projects that include European participants are encouraged to consult the ECSA website for information regarding European policy, ethics, and legal issues.

2 Standards

This section provides guidelines about standards relating to Earth Science citizen science data. Included guidelines address: documentation; data content, format, metadata, and quality; and attribution (i.e., credit to authors and significant contributors). The guidelines presented here reflect best practices assembled by practitioners of NASA-funded Earth Science citizen science programs/projects and members of the NASA ESDS community. The guidelines are intended to be considered for pre-proposal and post-award data producers and providers, the former to inform programmatic expectations while proposals are prepared and the latter to assist in the conduct of the awarded projects.

2.1 Introduction

NASA's Earth Science Data Systems (ESDS) process, archive, and distribute a large volume and variety of data products. Historically, the data products being archived have originated from observations by space-borne, air-borne, or in situ instruments – all which have been scrutinized by the scientists creating the datasets. ESDS standards, best practices, and archival processes and architectures accommodate these data products. NASA-sponsored Earth Science citizen science is a new paradigm, requiring unique considerations to harmonize those data within ESDS.

NASA is active in citizen science (https://science.nasa.gov/citizenscience). "In citizen science, the public participates voluntarily in the scientific process, addressing real-world problems in ways that may include formulating research questions, conducting scientific experiments, collecting and analyzing data, interpreting results, making new discoveries, developing technologies and applications, and solving complex problems" (citizenscience.gov). All federal agencies, including NASA, are granted authority to do citizen science by the Crowdsourcing and Citizen Science Act of 2016 (15 USC 3724). "Through citizen science and crowdsourcing, the federal government and nongovernmental organizations can engage the American public in addressing societal needs and accelerating science, technology, and innovation" (citizenscience.gov).

While the NASA ESDIS (Earth Science Data and Information System) Standards Office (ESO) releases and continues to lead the review, approval, and adoption of standards for NASA Earth science data (https://earthdata.nasa.gov/esdis/eso), the broader citizen science community has been developing its own standards and best practices for citizen science data (PPSR_CORE 2013; DataONE, 2013). Material from ESO and the broader citizen science community are drawn upon here.

By incorporating the guidance in this section, NASA Earth Science citizen science data "will be self-describing to the extent that a future data user will be able to decode and use the data… while needing to consult few, if any, external resources" (Evans et al., 2016). This document establishes a basic set of guidelines such that NASA Earth Science citizen science data can meet FAIR data guidelines, as described in Wilkinson et al. (2016). For example, can the citizen science data support professional-, student-, and citizen-driven research? Section 1 provides guidance to ensure that the collection, management, and sharing of citizen science data conforms to international, federal, NASA, and local laws and policies.

This document refers to the type of citizen science in which participants are actively involved in conducting observations. They are assumed to be knowingly contributing to a citizen science project – whether they are making actual physical measurements (for example, by installing and setting up a low-cost air quality monitor) or providing information about specific phenomena (e.g., recording whether there is a cloud in the sky at a particular time). This document does not

focus on the case where an investigator is “mining” public data from, for example, social media (e.g., monitoring Twitter feeds).

Legal issues (e.g., personal safety, personally identifiable information, etc.) related to citizen science are detailed in Section 1.

Section 2.2 provides recommendations about collecting the data, whereas Section 2.3 provides recommendations on how data should be archived. Section 2.4 details how the data should be attributed to enable provenance and scientific integrity. Due to the relative maturity of NASA’s Global Learning and Observations to Benefit the Environment (GLOBE) program in providing data through citizen science efforts, GLOBE documents are used as examples.

2.2 Citizen Science Data Collection: Documentation and Protocols

2.2.1 Documentation

Projects are strongly encouraged to provide documentation to the participants doing the data collection and the end users using the data who may or may not have been involved in the initial data collection.

Measurements are more accurate and consistent when citizen scientists receive training and have an opportunity to practice using the instrument or device (Freitag et al., 2016) or in projects where multiple citizen science observations are aggregated to find agreement (Rosenthal et al., 2018). Additionally, volunteers or citizen scientists collect better data when they have a vested interest in the project (Lewandowski & Specht, 2015). In order to attain the highest quality measurements of Earth Science variables, the best practice is to standardize the protocols and to include the following information in clear terms:

1) Data Quality Objectives and Indicators such as precision, bias, accuracy, representativeness, completeness, sensitivity, measurement range.
2) Sampling Design
3) Sample handling and Custody
4) Equipment/Instrument Maintenance
5) Testing, Inspection, and Calibration
6) Field and Laboratory Quality Control: Verification and Validation
7) Data Usability

2.2.2 Documentation for Instrument Operation

This section provides guidelines for a document that should be aimed at individuals operating the instrument(s) used in a project to produce data. Instrument documentation primarily consists of drawings, diagrams, specifications, and operating procedures. This should include the following topics for successful operation of the instrument and production of a quality dataset:

1) Design and diagram of the instrument
2) Design criteria, standards, specifications, vendor lists
3) Manufacturing details of the instrument
4) Software used and version number
5) Commissioning
6) Operating instructions
7) Maintenance
8) Feedback mechanism

2.2.3 Documentation for Data Use

This section outlines guidelines for documentation that should be aimed at end users of the data. Projects should provide sufficient documentation such that end users can understand, access, and use the project's available data. Elements in this "data user guide" should include:

1. Project objectives and scientific questions guiding data collection (the "why")
2. Project participants
3. Citation for the data
4. Information about how to report issues in the data
5. Methods and materials for collecting data
6. Steps used in data processing
7. Definition of data variables and units
8. Identification of variables that are estimated versus directly measured
9. Derivations for estimated variables (including assumptions made)
10. Quality assurance
11. Terms of use
12. Instructions on how to access data
13. Updates to the data including version control
14. References

Helpful data use documentation examples include the GLOBE Data User Guide and GES DISC ReadMe. DataONE's Best Practices for Describing Data is also a useful resource when assembling project data documentation.

2.2.4 Measurement Protocols or SOPs

Collecting relevant observations largely depends on the methodologies used to acquire the observations. The goal is to mitigate individual biases and human error while obtaining the observations. Thus, the methodology should be described clearly and concisely in a Standard Operating Procedures (SOP) document.

SOP items critical to meeting measurement objectives include:

A) Measurement location or site. The SOP should detail selecting measurement sites based on the optimal requirements of the geoscience variable. Sites should be representative of the surrounding area or observing conditions. For example, a soil moisture measurement site should represent the surrounding soil and vegetation type. Likewise, a water quality sample should be collected from the water body in question. Similarly, for other geoscience variables, the standards should be clearly indicated in the SOP.

B) Repeatability. The SOP should indicate the minimum sampling required to represent the average of the geoscientific variable. Repetition helps address the imprecision of the device and mitigate human error.

C) Device calibration and condition. The SOP should outline the steps used to calibrate or prepare the device used in measuring the geoscience variable to collect observations to ensure they are of high quality.

D) Sample handling. The SOP should indicate whether the sample or data needs to be taken to a lab or otherwise receive post-processing to lead to high-quality or useful data. The SOP should discuss sample handling during collection and, if necessary, the lab procedures used to obtain the observation.

E) Recording data. The SOP should list all information that needs to be associated with the observation. This may include information about physical/digital units to report, as well as ancillary information such as date, time, location identifier, participant identifier, coordinates, etc.

2.2.5 Instrument/Device Specs

Knowledge of instrument specifications is key for collecting high-quality data. Make a specifications document available to those taking the measurements, as well as to those using the data.

At a minimum, the document should include:

a) Resolution (e.g., spatial, spectral, temporal)
b) Range of values the device can measure
c) Units of the raw measurement
d) Precision
e) Accuracy
f) Whether (and how) the instrument reports a derived variable

Optionally or when relevant, additional specifications may include:

g) Error calculations
h) Quality assurance
i) Description of power source (voltage, amperage)
j) Dimension of the device (size, weight)
k) Material requirements (hardness, pliability, focal length, magnification, etc.)
l) Environmental exposure ranges
m) Operating medium
n) Safe handling procedure

2.3 Data Archival

This section provides a discussion about how to ensure the citizen science data being collected are sharable (and usable) within the scientific community. This includes recommendations for file metadata, as well as recommendations for data formats and measurement units. When appropriate, citizen science data can and should be mapped to NASA-defined Earth science data levels. For each dataset being archived for long-term preservation, a Digital Object Identifier (DOI) and its associated dataset landing page shall be created and serve as the main gateway to access data, metadata, documentation, and other related resources. The section details quality assurance, along with suggestions for planning, gathering and conveying data quality. Further detail about archival is provided in Section 4.

2.3.1 Metadata

Metadata describes and gives information about the data. For citizen science, metadata should document information about the project as well the measurement data that were collected. Metadata helps to ensure that the citizen science data are FAIR. Metadata can also describe how data were collected, as well as any limitations associated with the data. Metadata may be embedded within the citizen science data file (e.g., geospatial and photo) or as separate files (e.g., XML or README). As a general rule, metadata is concise information about the data self-contained in the data file or always provided together with the data file. More detailed descriptions about the data collection process, instrument operation, data processing goes in a user guide on the project landing page.

There are multiple standards for archiving project data, and different data archives may have different structures (e.g., Public Participation in Scientific Research (PPSR), DataONE). Different standards may reflect different project needs – citizen science focused, NASA mission focused, social sciences focused, etc. – but all look to capture the “what”, “why”, “who”, “when”, “where”, and “how” of the project and dataset (Michener et al., 1997), as well as capturing information about the measurements themselves.

A sample metadata structure used for the GLOBE citizen science program is included in Appendix B. It includes three sections: one focuses on the project level information, the second characterizes the dataset, and the third explains the specific data (e.g., at the site). Although the structure of metadata for a specific citizen science project may take another form, focusing the project metadata on the questions above will help ensure the project is adequately documented. Although some information may be provided on a project's website, which could go offline after the project ends, information about the project and data should be captured in the metadata for long-term preservation.

Information that may be included within metadata include:

- Why was the data collected? This may include scientific questions and/or other rationale.
- Who is responsible for the dataset? Relevant information should include contact information for PI, organization, archive location, project sponsor, etc.
- Who collected the data? Depending on the project, this could be individuals' names or relative demographics, such as expertise, training, age ranges, whether they are "students", etc.
- What is within the dataset? This includes a description of the variables measured and/or data recorded. Relevant information may include units of variable, valid min/max values of the variable and flags, or information regarding missing or bad data. Additional information might include the language, version number, and/or last update to the data.
- What assumptions were used to create the dataset? The variable 'reported' may not be what is directly measured. There may be internal processes, including calculations (requiring coefficients) and/or aggregation (for example, averaging over a time interval). In addition, the participant may have to do something manually (turn on a switch, close a cover, etc.) that could be reflected within the metadata.
- What is the use and distribution policy? Some issues to consider are whether there may be restrictions on whom should use the data (terms of use), and if used, whether there should be citation and/or acknowledgement requirements (e.g., when using the data for a publication).
- What problems exist in the dataset? Are there known biases or known malfunctions (for example, collected under poor operating conditions).
- Where was the data collected and with what resolution? Useful information might include site name(s), latitude/longitude/altitude, and any other details that may be useful (e.g., "on the southwest corner of the roof").
- When were the data collected? Were the data collected at one time only or over a period of time? How frequently were they collected over the time period?
- Were there any changes to data collection methods or processing algorithms during the program?
- How was each parameter measured and/or data processed? This may include information about how each parameter was measured, instrumentation used, and/or any algorithms used to process the data.
- How was the measurement quality ensured? This information might include a description of quality measures used or enforced, including instrumentation, range checks, data analysis, etc.
- Were there events that might potentially impact the measurement? For example, for projects that measure air quality, events such as wildfires, construction work, etc., may need to be recorded to enable data interpretation and quality assurance.
- How reliable are the data (uncertainties)? This may include estimates of accuracy and/or precision. This may also include quality flags and/or other indicators that help a user evaluate the fitness for use of a given measurement. Good indicators allow for a user to extract only the portion of data products meeting their data quality needs (ESDS-RFC-

033, 2019). Section 2.3.5 provides additional information about standard measurements of data quality.

- How can someone get a copy of the dataset? This can include information about how to access the dataset archive and may include additional contact information, such as a helpdesk point of contact and/or publication references.

These bullets provide guidance to the citizen science data 'provider' in creating a FAIR dataset. Following established standards can help facilitate incorporating these metadata elements into data files. For example, the conventions for CF (Climate and Forecast) metadata provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. CF includes a table of community-sourced standard names for typical climate- and environmental-related variables. Additionally, for pictures and photos, consider using the Exif and IPTC photo metadata standards (Version 2017.1 Revision 3IPTC, 2019), which can describe and provide information about attributes, producing equipment, rights, and administration of an image.

Even though the CF metadata conventions were created with a focus on data stored in the NetCDF format, many of the metadata elements and standard names are applicable for data in other formats. Thus, CF conventions are increasingly gaining acceptance in communities beyond climate. Citizen science research pertaining to data in those fields can benefit from the CF metadata or similar conventions.

2.3.2 Data Format and Structure

2.3.2.1 File Format

Citizen science data can be large in volume and complex in structure. However, a unique aspect is that they typically consist of many small pieces of data contributed by volunteers. Efficiently managing large numbers of small pieces of data requires proper data formats, data structure, and interlinking of different data components.

Open and lightweight data formats that can facilitate remote data transmission over networks are preferred. Examples of recommended text-based data formats include ASCII, JSON, and GeoJSON. Examples of recommended image (and video) formats include JPEG, PNG, and TIFF. These are all "open source" image formats and can be chosen based on needs, such as lossy or lossless compression. Data files can be made self-descriptive by embedding metadata elements that follow community conventions. Good metadata practices help integrate data contributed by different volunteers, promote data usage, and facilitate long-term data preservation and reuse.

If using ASCII format for citizen science data, guidelines have already been established for Earth Science Data (Evans et al., 2016). This document provides recommendations for a minimum and necessary set of information to be contained in an ASCII file. These recommendations were originally developed by the NASA ASCII Earth Science Data Systems Working Group and reviewed and adopted by the Earth Science Data and Information System (ESDIS) Standards Office. ASCII files should be self-describing to the extent that a future data user will be able to decode and use the data in the file while needing to consult few, if any, external sources. These guidelines provide recommendations on general file format specification and structure, header section, data representation for different types of data (e.g., point/time series and profile/gridded data), spatial and temporal information, missing data and limits of detection, and file names.

NetCDF is an open data format that can store many different types of data, including scalar as well as multi-dimensional data. Combined with CF metadata conventions, data in the NetCDF format can be information-rich, self-descriptive, and suitable for long-term preservation.

A single data package for one observation can consist of multiple pieces of data, such as descriptive metadata, images, and videos as well as the data themselves. Therefore, properly interlinking those different data components is important as well as the assignment of consistent identifiers and filenames. Identifiers are also critical to integrate different observations by space, time, objects, themes, etc.

There is no “one way” to ensure citizen science data meets FAIR data standards. Citizen science projects and their associated long-term data archives can choose the format and structures most appropriate for their users. However, to better address different user needs, citizen science projects and data archives may choose to archive data in multiple different formats. An alternative is to consider a static archive, but leverage data transformation services to deliver citizen science data in additional formats.

2.3.2.2 Data Content

For others to use citizen science data, the contents of the dataset should be fully understandable. For example, there should be documentation introducing the parameter names, units of measure, formats, and definitions of coded values. It is best practice to be consistent throughout a dataset (ORNL DAAC Data Management Best Practices).

It is strongly recommended to standardize each parameter across files, datasets, and the project, using commonly accepted parameter names and abbreviations. To help the data user, the citizen science dataset should include a data dictionary that defines each attribute, variable, and parameter in the data. Standards for parameters currently in use include the CF Conventions and Metadata (https://earthdata.nasa.gov/esdis/eso/standards-and-references/climate-and-forecast-cf-metadata-conventions). These standards are applicable for NetCDF, as well as other data formats.

Specific recommendations to enable FAIR data include:

- Define spatial reference systems (type, datum, and spheroid). Use European Petroleum Survey Group (EPSG) code, if available. Define a bounding box if necessary. Provide a separate projection file (e.g., ProJect (PRJ) or Well-Known Text (WKT) formats).
- Report latitude and longitude coordinates in decimal degrees (up to four decimal places, ~10-meter resolution). Record south latitude and west longitude as negative values.
- Define and document all special values, including valid ranges, scale factor, and offset of the data values. This should be contained as metadata within the file and documented outside.
- Use consistent missing value notations throughout the dataset. For numeric fields, represent missing data with a specified extreme value (e.g., -9999). Section 2.3.5 provides missing data options, but the key is that representation of missing values must be documented and consistent.
- Define and standardize any coded fields in the data. A separate field may be used for quality considerations, reasons for missing values, or indicating replicated samples. Codes and flags should be consistent across parameters and data files. Definitions of flag codes should be included in the dataset documentation.
- Use appropriate encodings. Citizen science data, especially those collected in studies involving international participants, need appropriate encodings to support character sets used by different languages. The American Standard Code for Information Interchange (ASCII) is an encoding for representing English characters with 127 numbers. The American National Standards Institute (ANSI) codes extend ASCII with an additional 128-character codes. Unicode Transformation Format (UTF) encodings, including UTF-8/16/32, go beyond 8-bit and support almost every language (or script) in the world.
- Assign descriptive and consistent file names. File names should reflect the contents of the file and uniquely identify the data file. File names may contain information such as

project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type (ORNL DAAC Data Management Best Practices). To ensure easy management by various data systems and to decrease software and platform dependency, file names should contain only lower-case letters, numbers, and underscores – no spaces or special characters. Similar logic is useful when designing directory structures and names.

Additional resources:

- ORNL DAAC Data Management Best Practices, ORNL DAAC
- ORNL DAAC CSV Standards, ORNL DAAC
- NetCDF Data Requirements, ORNL DAAC
- Data Product Development Guide for Data Producers, NASA ESDIS (in review)
- Dataset Interoperability Recommendations for Earth Science, ESDIS Standards Office

2.3.3 Measurement Units

Using consistent measurement units is of key importance, especially since many projects are international in scope. All data should be reported with relevant units of measurement and conform to the International System of Units (SI) system. While actual data collection may take place in any relevant unit, all data submitted should be converted to SI units. If a variable is dimensionless, it should be noted as such in the unit field. It is also recommended that the participants specify the reference condition at which the data is collected. For example, in the atmospheric field, air volume may be specified/measured at standard conditions or actual conditions of temperature and pressure, which will impact derived concentration units such as mass/volume.

Standard and normal conditions are defined as:

- Standard condition: 273.15 K (0 °C, 32 °F) and 10_5 Pa
- Normal condition: 20 °C and 1 atm

In the event a variable does not have an accepted notation, the project producing the data may use the unit conventionally used by the relevant field. In this case, definition of the unit should be included as part of the metadata submission.

Resources for unit specification for variables commonly encountered in the atmospheric field include:

- UDUNITS-2 is a computer package/utility that provides recognition and conversion of wide variety of measurement units. The CF metadata convention requires unit values to be compatible with and recognized by UDUNITS-2. UDUNITS-2 is based on and extends SI. The three main UDUNITS-2 components are: 1) (udunits2lib) a C library for units of physical quantities; 2) (udunits2prog) a utility for obtaining the definition of a unit and for converting numeric values between compatible units; and 3) an extensive database of units. Additional information is available at https://www.unidata.ucar.edu/software/udunits/udunits-current/doc/udunits/udunits2.html.
- Date, time, and duration formats should conform to ISO 8601 (https://www.iso.org/iso-8601-date-and-time-format.html). ISO 8601 is a standardized way of presenting dates and times, which helps to eliminate uncertainty and confusion when communicating internationally. The CF metadata convention and NASA's ASCII File Format Guidelines for Earth Science Data (Evans et al., 2016) suggest using ISO 8601 to represent date and time information. The full standard, including ISO 8601-1:2019 and ISO 8601-2:2019, provides instructions for reporting date and time in appropriate formats compared to reference times. These include 'Date', 'Time of day', 'Coordinated Universal

Time (UTC)', 'Local time with offset to UTC', 'Date and time', 'Time intervals', 'Repeating time intervals', and other useful notations.

2.3.4 Mapping to NASA Data Levels

NASA and other agencies have defined 'data processing levels,' ranging from unprocessed instrument level measurements to statistics of global geophysical parameters on defined spatial and temporal grids. Data processing levels of NASA's Earth Observing System Data and Information System (EOSDIS) data products range from Level 0 to Level 4. Level 0 products are raw data at full instrument resolution. At higher levels, the data have undergone additional processing (see full definitions). The following examples refer to data archived at the National Snow and Ice Data Center (NSIDC).

Level	Definition
0	Unprocessed instrument data
1A	Unprocessed instrument data alongside ancillary information
1B	Data processed to sensor units (e.g., brightness temperatures)
2	Derived geophysical variables (e.g., sea ice concentration)
3	Variables that are mapped on a grid (e.g., data using EASE-Grid)
4	Modeled output or variables derived from multiple measurements

Section 4.3.1.2 provides additional detail and examples of mapping of citizen science data to equivalent NASA EOSDIS data processing levels.

2.3.5 Standard Measurements of Quality

It is recommended that citizen science projects follow the best practices for data quality control and assurance outlined by DataONE.org (DataONE, 2013):

- Develop a quality assurance and quality control plan
- Communicate data quality
- Mark data quality control flags
- Identify values that are estimated (i.e., not directly measured)
- Identify missing values and define missing value codes
- Ensure basic quality control
- Double check data entry
- Publicize quality issues

Develop a quality assurance and quality control plan

Document procedures taken to assure and control the quality of a project's data. Such checks may include training the project requires before participants can contribute data, range or logic checks applied to the data upon ingestion into the project database, post processing quality checks, etc.

Communicate data quality

The Data Quality Indicators from the Handbook for Citizen Science Quality Assurance & Documentation (EPA, 2019) are a good guideline for communicating the quality of citizen science data. Data Quality Indicators are attributes of the data being collected, specifically related to minimizing the uncertainty for each measurement or set of measurements. The information below is adopted from EPA (2019). Many of these terms have mathematical definitions as well as qualitative ones.

Precision is the ability of a measurement to be reproduced consistently. Taking many measurements over a small temporal and spatial domain helps to evaluate precision. How 'closely' those measurements agree (compared to expectation of variability of the measured parameter) determines the precision. Precision is often reported as the relative percent difference or the relative standard deviation.

Bias is the ability of the measurement to 'on average' report the true or expected value. Bias is increased by any influence that might sway or skew the data in a particular direction. Bias can result from a non-representative sampling design, calibration errors, unaccounted-for interferences, and chronic sample contamination. For example, taking samples from one location where a problem is known to exist, instead of taking samples evenly distributed over a wide area, can lead to bias. Bias can also arise from human influence, including poor measurement technique. Bias can be calculated as the 'average' error (compared to a reference) within a sample.

Accuracy refers to the combination of precision and bias and indicates the degree of confidence in a new measurement. An accurate measurement is one with minimum bias and greatest repeatability. Accuracy can be determined by repeatedly taking the same measurement compared to the known or expected truth.

Representativeness is how well the collected data depict the true system. For example, a single measurement at one time may not sufficiently represent the geophysical parameter the project is trying to observe.

Comparability is the extent to which data from one dataset can be compared directly to another dataset. The datasets should have enough common ground, equivalence, or similarity to permit a meaningful analysis.

Completeness is a measure of the amount of data that must be collected to achieve the goals and objectives stated for the project.

Sensitivity is essentially the lowest detection limit of a method, instrument, or process for each of the measurement parameters of interest.

Measurement range is the range of reliable readings of an instrument or measuring device, or a laboratory method, as specified by the manufacturer or the laboratory.

Mark data quality control flags

ISO 19157: Geographic information -- Data quality provides an internationally standardized guide that can be used to measure and/or flag the quality of geospatial citizen science data. Following the implementation of ISO 19157 to citizen science data from Foody et al. (2017), the following measurements of geographic data quality should be checked and flagged, if necessary:

- Positional accuracy. Validity of a measurement's reported position. Flag example: reported position is suspect because it is not physically meaningful (e.g., elevation of -11,000 m) or outside the expected geographic boundaries (e.g., location of a land-based measurement is reported to be over the open ocean).
- Temporal quality. Validity of parameters like date of collection, update, etc. Flag example: measurement intended to be made during daytime is reported at midnight.
- Completeness. Presence or absence of required features of an observation.
- Logical consistency. Indication of whether the observation makes physical sense.

- Thematic accuracy. Accuracy of the classification. Flag example: participant reported cloud type as cumulus, but upon checking the cloud type, it is determined to be stratus.

Data should also be checked for outliers and detected outliers flagged. Statistical and visual methods for detecting outliers should be used and, if used, should be documented in the data user guide (Section 2.1.2).

Identify estimated values

This topic is covered in Section 2.1.2.

Identify missing values and define missing value codes

It is recommended that Recommendation 6: Missing Data and Limits of Detection from NASA (2016) Evans et al. (2016) be adopted for identifying and defining missing value codes.

- Indicate cases where measurements cannot be made due to instrument or other related issues. If commas or other visible characters are used as the delimiter, the field in the data record that would normally include the missing measurement can be left absent. However, if using space or tab delimiters, the field in the data record must not be left empty or blank, and instead must include some designated value for the missing data.
- Describe the value(s) used to designate missing data in the header. Represent missing data using numbers of enough magnitude to never be construed as actual data (e.g., -99999).
- Data below or above a limit of detection (LOD) are not actually “missing” but do convey useful information when used to compute descriptive statistics. These conditions should be indicated by additional missing data flags substituted for the missing data values. If used, these flags and the values of the upper and lower LOD must be described in the header. For example, the flag sometimes used for data values GREATER THAN some UPPER LOD (ULOD) is -7777 (or -77777, etc.), and the flag for data values LESS THAN some LOWER LOD (LLOD) is -8888 (or -88888, etc.).
- If LLOD or ULOD values vary from point to point, they should be given in a separate column of data.
- However, use of Not a Number (NaN) to represent missing values is a matter of debate within the ESDS community and the ESDSWG Data Interoperability Working Group recommends against using NaN (Jelenak et al., 2019, Section 3.7). In particular, NaN is a specific floating point value in many computer systems, so it cannot technically be used for integer variables. It is also not universally recognized and can create compatibility issues.
- Describe in the header any other flag values used in the data section.

Disclose quality issues

This section is based on Data Quality Working Group's Comprehensive Recommendations for Data Producers and Distributors. It is critical to expose quality issues associated with data products to the broad community of data users in a timely and efficient manner. This includes recommendations on possible approaches to capture and publicize known limitations, quality issues, and updates of data products.

- Data producers should ensure all known issues discovered by the science teams and data users are reported to the data archive in a timely manner.
- Data producers should establish a well agreed upon definition of outlier (extreme values) for each product based on science understanding of the distribution of values for the parameters of interest.
- Data producers should identify outliers, as well as produce guidance, e.g., via documentation or online alert/flag, providing users useful data quality information such as 1) quantity and location of outliers, 2) magnitude of each outlier, and its ratio relative

to the expected max/min of the data or some other well-defined statistical measure, and 3) origin of the problem.
- Data archives should host a prominent web page that captures known quality issues.
- Data archives should provide enough publicly available information with self-describing metadata and documentation such that the need for users to contact the data archives is minimized.
- Data archives should inform users, as soon as possible, when data are compromised and provide status updates when readily available. Alert PIs and/or data producers to issues that arise and/or are reported by data users.

Additional resources
- Best Practices for Data Quality Assurance, DataONE.org
- Quality Assurance for Citizen Science Projects, US Environmental Protection Agency
- Quality Assurance Handbook and Guidance Documents for Citizen Science Projects, US Environmental Protection Agency
- Resources for Data Quality in Citizen Science, Wilson Center
- Useful and evolving material relating to Citizen Science data quality is available at https://github.com/CitSciAssoc/DMWG-PPSR-Core/wiki/Approach-to-addressing-Data-Quality-and-Quality-Assurance-Processes-through-the-PPSR-Core-Standard

2.4 Attribution

Proper attribution of scientific work undergirds scientific integrity, as discussed in McNutt et al. (2018). The NASA Science Mission Directorate (SMD) specifically affirms in policy SPD-33 that citizen scientists should be acknowledged, noting “SMD citizen science projects shall acknowledge the citizen scientists they work with as a collective in publications or include them as named co-authors on these publications when their contributions warrant.” This section provides guidance as to what forms of acknowledgement may be appropriate.

2.4.1 Authorship

NASA considers a ‘data’ publication to be the act of archiving data in an open data repository, such as a DAAC. Just like a journal publication, it has an author list and a digital identifier. As such, the author list for the data product should include anyone who contributed substantially to the data collection, processing, and/or analysis that validated that data. It may not necessarily be the same list as that in a related journal publication that used the data or described the methodology used to create the data. For example, a person who used a published dataset in their research may not need to be on the author list for that data publication. Likewise, persons responsible for gathering funds for the project, paying salaries, providing a conducive environment, or being the spokesperson do not necessarily warrant authorship, unless they have made a significant contribution to the intellectual and/or scientific content of the data. An acknowledgement could be more appropriate in such cases.

For citizen science efforts, judging whether someone “contributed substantially” to data collection can be a grey area. A single contributor, out of hundreds or thousands of citizen scientists contributing to a project likely does not rise to the level of a substantial contribution. However, a participant who coordinated the recruitment of other citizen scientists, or who was directly involved in quality control, even as a volunteer, could be considered as contributing significantly to the processing and analysis of the data. The guidelines given in (McNutt et al., 2018) should be used in assessing contribution and as the basis for dialog with individuals about authorship versus acknowledgement.

The ESIP Data Citation Guidelines for Earth Science Data (ESIP DPSC, 2019) also provide useful guidance for roles meriting inclusion in the Author listing for a data product citation.

2.4.2 Acknowledgement

A key element for a data publication is the dataset landing page, which provides access to the data itself and documentation about the data. Generally, the digital object identifier (DOI) for a data product will resolve to this dataset landing page. Several guidelines for dataset landing pages exist, including the DataCite DOI Landing Page Guidelines. This landing page, as well as the accompanying documentation, are appropriate places to provide acknowledgement for those individuals and organizations who have contributed to the work but who are not authors.

However, citizen science data collection often happens near an individual's home (or at least places that they regularly frequent). Given that confounding of identity with location, there is often an expectation in citizen science projects that identification of individual data collectors is an opt-in process. In other words, for some projects, their policy is that a data collector may not be identified without that individual's explicit permission.

For further reading, see McNutt et al. (2018); https://earthdata.nasa.gov/earth-observation-data/data-citations-acknowledgements covers ESDIS and DAAC statements and guidelines. The NASA Science Mission Directorate Policy Document 33 also provides commentary about citizen scientist acknowledgement.

3 Usability and Notifications

3.1 Usability

An important NASA Earth Science program objective is to facilitate the use of data products and services derived from funded citizen science projects. Usability is the extent to which a system, product, or service can be used to achieve goals with effectiveness and efficiency in a specified context (ISO 2018). Users include scientists with specialized expertise as well as the broader community, such as students, professionals, and decision-makers. Projects that emphasize the usability of their data and services for a broad range of stakeholders are well poised to achieve long-term impact in scientific discovery and acquisition of knowledge. Additionally, usability and trust go hand-in-hand; the level of use of a dataset by the broader community is a measure of the value of its information content.

3.1.1 Usability Checklist

The following is a checklist of usability considerations for data products and services generated by NASA citizen science projects. This checklist provides high-level guidance for proposal development, project execution, and final dissemination of data and services.

Metadata

- Data description should follow a metadata standard, with proper ontology.
- Use a metadata standard to document how the data properties (why, who, what, where, when, and how) are codified when data are collected.
- Metadata recommendations are provided in Section 2.3.1.

Identifier

- To maintain persistence and enable machine readability, data and code should have a Digital Object Identifier (DOI).
- DOIs can be assigned to datasets, video, audio, streaming media, 3D objects, journal articles, supplemental material, technical reports, and visualizations.
- GitHub can be used to assign a DOI to code (e.g., with Zenodo).
- A data identifier can also be a URL that links to data at a specific time or data version (e.g., set used for a particular published analysis).
- Additional information about DOI assignment is provided in Section 4.3.5.
- Data and code should be developed with a versioning system. This is particularly important for dynamic applications, with rapidly changing data that may appear in analyses or visualizations.

Data and code archival

- Identify where data and code will be stored for long-term archival (Section 4.4).
- Establish the process and timeframe of archival prior to project initiation (Section 4.5).
- Ensure data and code are accessible. In particular, NASA has an open policy for data and code (Section 4.3.2). Additional considerations for accessibility apply when data/code are stored in NASA data centers or in other locations (Section 4.6).
- Archival services may be required to collect information about data and code users, which in turn may be used by the data providers to assess usability and value, to support the infrastructure that hosts the data.

User requirements

- Consider the users of data and code, including the scientific, citizen science, decision-maker, and other stakeholder communities. A user could also be a machine.

- An interview of a user base can identify the most appropriate types of information, data structure, metadata and tools.

Data quality

- Data quality should be regularly assessed and maintained.
- Recommendations for documenting data quality are provided in Section 2.3.5.

Documentation

- In addition to metadata (Section 2.3.1), data and code should include documentation (e.g., data dictionary, algorithm description, data collection procedures, processing methods). Recommendations for documenting data are provided in Section 2.2.1. Peer-reviewed publications can also be used to document data collection, data use, and code.
- Recommendations for documentation of instrumentation used to collect data are provided in Section 2.2.2.
- User guides and training materials should consider the scientific community as well as other communities of users, such as citizen scientists. Guides and materials may need to be tailored to these different audiences.

Policy considerations

- Consider national and international policies when distributing data and code from citizen science projects (e.g., regulatory and legal frameworks, privacy agreements and/or considerations, and ethical issues when distributing data to external audiences).
- Clearly indicate how data may be re-used and who retains ownership and licensing (Section 1.4). Data use policies must also conform with NASA open data requirements.
- Machine-readable data services may be required to conform to security standards, such as FedRAMP.

Outreach

- A project should budget resources for actively promoting data usability and use cases through scientific and non-scientific channels.
- The success of citizen science projects depends largely on effective outreach and engagement with the citizen science community to collect the data. The project should include (and budget for) an effort to provide the community with example results and news about the impact of the science that they contribute to. Example forms of outreach include publications, newsletters, brochures, and online/social media.

3.1.2 Use Cases for NASA

The working group recommends that NASA maintain a repository of use cases or demonstration applications that use Earth science data generated by citizen scientists. These use cases would demonstrate “best practices”, highlight project successes and impact, and serve as a means for outreach. For the public, these use cases would provide information on a given project and an introduction to data available for potential users. Further, use cases could be a template for scientists to follow when developing a NASA proposal. A citizen science project can volunteer their use case (e.g., example data, tutorial, software or web application) for demonstration for NASA’s Earthdata website (earthdata.nasa.gov/).

Currently, NASA maintains a list of Implementation Phase funded projects in the Citizen Science for Earth Systems Program (CSESP). The site includes a project summary, annual updates, and links to external project webpages. In the future, this site could be expanded to include the proposed use cases. For example, the citizen science project from Research Triangle Institute has a web map of air quality sensors and Soundscapes to Landscapes maintains a web-map of bird occurrence maps from species distribution models.

3.1.3 How to Provision Data to Facilitate Usability

Although the concept of “Analytics Optimized Data Stores” (AODS) arose in the context of Big Data to help address the challenges of volume and variety, the concept applies to citizen science data as well. From a citizen science data perspective, AODS can be defined as data stored in a way that 1) minimizes the need for data preprocessing, 2) uses storage forms that support fast access, and 3) uses storage structures optimized for queries relevant to specific user communities.

A primary goal of data storage is to help the end-user to acquire data in as ready-to-use format as possible. A general approach to achieve this goal is to produce “Analysis-Ready Data” (ARD), also known as “value-added data”. Producing ARDs involves tradeoffs between cost (of production), customization (to user’s current specific needs), and flexibility (for user’s other potential needs). For a given user need, there is an optimal form of ARD with respect to these tradeoffs. Generally, the aim is to provide end-use-enabled data.

AODS, ARDs, and the related concept of “data warehouse” all address the varying user data needs, mainly content, format, and service. These needs vary among scientific and other user communities as well as within individual communities. The specific tradeoffs of cost, customization, and flexibility in producing ARDs--and, thus, whether the work is done by the user or by the archive--will also vary depending on the unique needs of the user. One general strategy is to create a data warehouses based on general ontologies but with user interfaces and APIs that enable user-desired customization. The “data rods/data cubes” services provided by the GES DISC are an example of such tradeoffs (Teng, 2016). Data rods are time-series files of individual spatial points, which could be ground sampling points or grid cell “points.” Data cubes are the actual stored files from which data rods are generated on-the-fly. Data rods can be thought of as ARDs and data cubes as AODS.

Sections 3.1.3.1 and 3.1.3.2 highlight some specifics of and differences between, respectively, the scientific community and other users regarding data structure, storage, and maintenance. Topics include understanding user needs, documentation, visualization, and storage and distribution formats. Within each community, there is also a range in user groups. For the scientific community, users range from those of science mission teams to applications users (Section 3.1.3.1). Outside of the science community, users may include citizen scientists, students, professionals, decision-makers, and other stakeholders (Section 3.1.3.2).

3.1.3.1 For Scientific Community

For NASA EOSDIS-supported data, requirements for data structure, storage, and maintenance are detailed in Archiving, Distribution, and User Services Requirements Document (ADURD).

Understanding user needs. NASA user needs are reflected in ADURD. The NASA Distributed Active Archive Centers (DAACs), through their interactions with users, also continually collect user needs and respond accordingly. In addition, the DAACs participate in the annual American Customer Satisfaction Index (ACSI) survey, which includes results on federal government services. Though these existing information resources generally do not involve citizen science data, they can still provide citizen science projects with a useful general view of the needs of NASA scientific users.

Documentation. For scientific users, the main purpose of documentation is to ensure the proper use of the data as collected and processed by the mission science teams. The most common such document is the README, the purpose of which is to facilitate users to quickly open and view a file. Typically, READMEs are automatically distributed with data downloads. Consider providing README information in a markdown document, with example code, visualizations and documentation for a particular use case. Citizen science data archived at a DAAC will need to provide a README. Other user-desired documentation (examples are from

the GES DISC) include FAQs and recipes or how-to's. Those could serve as templates for citizen science projects. Also, for data archived at a DAAC, a Directory Interchange Format (DIF) document (NASA EOSDIS' "collection" or data set metadata file) will be required. See this example for the NLDAS data set at the GES DISC. For data at a non-NASA archive, if the data are to be discoverable by NASA users, projects would also need to provide a DIF document.

Visualization. The purpose of visualization for scientific users is more for browsing, quick looks, etc. than for detailed scientific analysis. However, some analysis capabilities are available in some visualization tools, and for some science users these capabilities are sufficient. NASA's Giovanni ("The Bridge Between Data and Science") is a popular example of a scientific visualization tool.

Storage and distribution format. A common data format for NASA science mission data is the Hierarchical Data Format (HDF5). What is best for citizen science data, however, will likely be project-dependent. For example, for Twitter data, Zarr is a good alternative to HDF5. Zarr has a similar structure as that of HDF5, has less overhead, and supports UTF-8 (compatible with more non-English characters, important for non-English tweets).

Whatever the format, scientific users generally want to work with the science data as collected or processed to "standard" products and not data that have been aggregated in some way or otherwise further processed. NASA mission standard products are processed to different data processing levels (Section 4.3.1). Different processing levels are preferred by different scientific users. For example, applications users generally prefer Level 3 or Level 4 (space-time gridded).

Citizen science data should be in a form that facilitates combined use with other complementary data. For projects that are linked in some way with satellite missions, this would be a requirement. For other projects, however, citizen science data should still contain standardized links, such as georeference information or other identifiers, for easy linking to other data sets.

Where feasible and appropriate, application program interfaces (APIs) should be provided to access the data (i.e., connecting to the data store, querying the data, retrieving the data). Example code and API documentation should be provided.

3.1.3.2 For Other Users

Understanding user needs. Projects should invest time and resources in making their citizen science data useful for users outside the scientific community, particularly other citizen scientists. When possible, a project should leverage existing knowledge about user needs from related organizations in a particular domain. Example user-driven and collaborative efforts in California include the Conservation Lands Network for regional land conservation and Our Coast, Our Future for assessing risks from coastal sea-level rise. If needed, a user community in a particular domain can be surveyed to better understand the types of data formats and distribution mechanisms that will best meet needs.

Documentation. A vital step is to write training materials that describe how to access and use data or code and make these accessible on project webpages or other sites that have open access. Useful approaches to this end include markdown and Jupyter notebook documents.

Visualization. Projects should consider how data will be distributed to a wide array of users. Some data can be integrated into existing, larger and well-maintained databases with a broad reach in a particular community. For example, bird observations could be stored in the Avian Knowledge Network. These larger data storage technologies usually have several data visualization and summarization tools available that can be readily used to showcase the data to a large audience. A project can also design a simple web-based interface that greatly lowers the barriers to data visualization and access. For example, an interface could permit the user to search data (e.g., with keywords or geographic extent) and display results on a map or in a

table, with an easy mechanism for direct download. Further, the interface could show simple products that use the data, such as a graph or map visualization. An example of such an interface is NASA's Giovanni (though its data are curated by the GES DISC). New open-source technologies, such as RStudio's shiny servers, are making these developments easier to accomplish. The use of large open data warehouse services, open graphic interfaces, and reporting tools such as markdown documents are in fact best practices for open science. Their use exemplifies and fosters a more transparent scientific development process.

Storage and distribution format. Minimally, data should be served in formats easily accessible to common or free software. For example, tabular data can be distributed in comma-delimited text files or Microsoft Excel. Geospatial data can be distributed in shapefiles (vector) or GeoTIFF (raster) formats, which are readily read by open-source software such as R, Python, or QGIS. These programing and software tools can also handle other vector and raster formats through the open-source GDAL translator library.

3.2 Notifications

This section includes the practice of informing the scientific and non-scientific communities about new citizen science data, applications, and derivative products. Notifications also provide a mechanism for letting parties involved in a citizen science project, such as volunteers and the science team (Section 4.2), understand how and when the data they have contributed are being used.

3.2.1 Public Outreach

3.2.1.1 By NASA

Citizen science project teams can work with staff at NASA Earth Science Data and Information System (ESDIS) to gain more recognition, attract citizen scientists, disseminate findings, and connect users to accessible data. At a higher level, NASA maintains a citizen science webpage to promote funded projects. NASA can use its social media presence (e.g., Twitter, Facebook, blogs) to engage the community of practice. For example, there is a citizen science Facebook group that provides a simple, low-cost means of sharing information on a project to a larger community of users.

3.2.1.2 By Projects

Citizen science projects should budget time and resources for project promotion and public outreach during the funding cycle. A project webpage should be developed that includes an engaging description of the project for the non-scientific community, science objectives and questions for scientists or practitioners, example products or findings, contacts, social media connections, news, brochures, publications and links to download data. Social media is an excellent platform to notify the user community of data archives, new applications, and publications that use a project's citizen science data. Project teams are encouraged to build a robust social media presence with adequate staffing and project commitment. These activities often require frequent posts to create a presence in the constant stream of posts that users confront on a daily basis. Automation tools that schedule postings on multiple platforms (e.g., Facebook, Twitter, Instagram) can help streamline the process, organize posts in small, digestible pieces, and save time.

Project teams should also seek to collaborate with entities that promote citizen science. Besides NASA (Section 3.2.1.1), the federal government site CitizenScience.gov provides a portal to a catalog of federally-supported citizen science projects, a toolkit for designing and

maintaining citizen science projects, and a gateway to a community of citizen science practitioners and coordinators across government. Useful non-governmental entities that promote citizen science projects are SciStarter and CitSci.org. These sites and CitzenScience.org also provide additional guidance on building a community, sustaining a project, and other best practices.

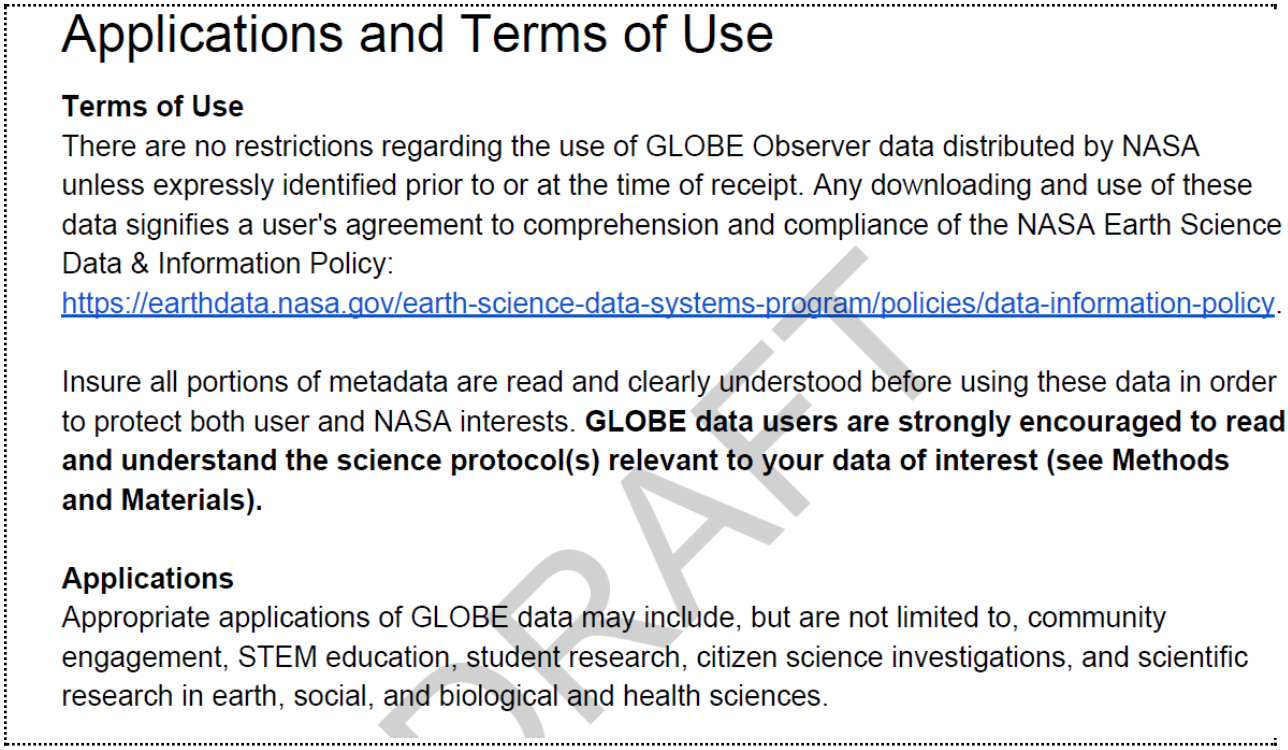
Funded projects are also encouraged to consider post-funding sustainability of their activities. This may require further fund-raising or partnerships to sustain data collection and to promote outreach and data usability. In particular, open-access, peer-reviewed publications that describe or use a project's citizen science data can maintain a long-term persistence of data notification.

3.2.2 Notifications on Use of Data

3.2.2.1 Citations of Project Data

Citing project data used in a research effort credits the project collecting the data and facilitates access to the data by other interested parties. Citations for citizen science data should account for the dynamic nature of these types of data, as they are continuously changing and growing over time. A citation should be updated accordingly to reflect a specific version. In addition, researchers may only be interested in a small subset of the data for specified values and for a select time range.

Project websites and/or user documentation should provide users with the specifications for citing the project's dataset(s), as well as any disclaimers or special instructions regarding the use of the data set. As an example, Figure3.1 contains an excerpt from the GLOBE Data User Guide that provides the Terms of Use for GLOBE Observer data; Figure 3.2 shows the prescribed GLOBE data citation format.



Applications and Terms of Use

Terms of Use

There are no restrictions regarding the use of GLOBE Observer data distributed by NASA unless expressly identified prior to or at the time of receipt. Any downloading and use of these data signifies a user's agreement to comprehension and compliance of the NASA Earth Science Data & Information Policy:
https://earthdata.nasa.gov/earth-science-data-systems-program/policies/data-information-policy.

Insure all portions of metadata are read and clearly understood before using these data in order to protect both user and NASA interests. **GLOBE data users are strongly encouraged to read and understand the science protocol(s) relevant to your data of interest (see Methods and Materials).**

Applications

Appropriate applications of GLOBE data may include, but are not limited to, community engagement, STEM education, student research, citizen science investigations, and scientific research in earth, social, and biological and health sciences.

Figure 3.1 GLOBE Observer dataset terms of use.

Citation for GLOBE Data

Global Learning and Observations to Benefit the Environment (GLOBE) Program, *Date Data was Accessed*, https://datasearch.globe.gov.

Figure 3.2 GLOBE data citation format.

3.2.2.2 Data Identifiers

Data and code should have a Digital Object Identifier (DOI) to maintain persistence of notification to the scientific and user community. DOIs can be assigned to a wide range of datasets and derived products, including video, audio, streaming media, 3D objects, journal articles, supplemental material, technical reports, and visualizations. More information about DOI assignment is provided in Section 3.3.5.

3.2.2.3 Citations for Articles Citing Project Data

In addition to crediting a project dataset, projects should also request that researchers notify the project with the citation for the published article. As an example, Figure 2. 3 shows the GLOBE User Guide request that authors of peer-reviewed articles citing GLOBE data send the citation of the published article to the GLOBE help desk.

NOTICE
If you publish a peer-reviewed article with GLOBE data, please credit the program and let us know so we can advertise your work on the GLOBE Publications page. Send the citation for your published article to help@globe.gov.

Figure 3.3 Request for notification of published article.

Projects should maintain a list of such articles on the project website, providing a clear illustration of the value and impact of the dataset, and giving site users an understanding of how the data set has already been used. Ideally, such publications lists should be searchable based on user supplied criteria.

3.2.3 Citizen Scientist Acknowledgement

Citizen science projects should acknowledge their top citizen science contributors. These citizen scientists may be volunteers who helped with data validation or those who contributed a high volume or high quality of data that was ultimately used in published papers. Project teams should consider which mechanisms in their process will report the required information needed to determine acknowledgement. For example, statistical queries on a citizen science dataset could reveal which individuals had high productivity in collecting data or superior data accuracy relative to peers.

Section 2.4 provides guidelines for determining what entities/persons should be acknowledged in publications. Other types of acknowledgements specific to citizen science may include awards, badges, prizes, or highlights in social media or webpage. Care should be given to maintaining privacy, and citizen scientists must be given the option to opt in/out of contact and acknowledgment by the project team (Section 1.2). As an alternative form of acknowledgment that maintains privacy, citizen scientists could be given a unique code to

maintain anonymity, or a project can acknowledge contributions to a larger group, such as through aggregate statistics or graphs.

4 Long-term Archival

4.1 Goals of Long-term Archival

The overall goal of long-term archival of citizen science data is to ensure availability and accessibility of data. This section is intended for the project research teams, specifically principal investigators. This section clarifies the roles and responsibilities of the parties involved and the content of the archives, identifies important factors to consider when choosing an archiving location, and offers best practices for the archive process and timeframe.

Archiving citizen science data ensures availability of valuable data sets long after the project has concluded. Archiving with a complete set of metadata ensures that the data are easily discoverable and accessible over the long term. Following standard best practices for data content (Section 2) and archives also ensures and enables the interoperability of these data with other complementary data sets generated by NASA (e.g., satellite data).

4.2 Roles and Responsibilities

In citizen science projects, there are multiple parties involved that may vary from project to project. These parties include principal investigators (PIs), research team, citizen scientists contributing to the project, funding agencies, external stakeholders, and archival organizations (e.g., data storage or archive location of an external cloud service).

Long-term archival is a shared responsibility between the data archives and the PIs; however, the requirements may vary depending on the type of archive (e.g., NASA-based or non-NASA). The other parties work with the PI to ensure that the data archived contain the necessary contents and associated metadata. Typical roles and responsibilities of the different parties for long-term archival are as follows:

Principal Investigators (PIs)

- Responsible for overall data collection and submission for archival.
- Ensure data compilation and submission as per requirements.
- Ensure compilation of necessary metadata.
- Ensure proper documentation.
- Coordinate with the archive organization in obtaining a digital object identifier (DOI) for the data set to be archived (Sections 4.3.5 and 4.3.8.2).

Research team

- Assist the PI with preparing the data for archival.
- Work with the different parties to format and compile the data and metadata for archival and to develop all necessary documentation for data and code.

Citizen scientists

- For those specifically recruited by the project, work with the PI and the research team to provide necessary metadata and documentation.
- Follow data collection requirements specified by the PI and the research team.

Funding agencies

- Provide requirements and guidance on data archival.
- Provide guidance on archival duration and, if applicable, any embargo period.

External stakeholders (designated project end users)

- Provide any feedback on data, data format, and documentation.

Archival organizations

- Have a policy in place for long-term archival, including information on, for example, duration, fees, and data restrictions/requirements.

- Provide guidelines to the PI on data format and archival requirements (e.g., required metadata) to facilitate the downstream provision of data services.

A key element for an archived data set is its landing page, which provides access to the data, documentation about the data, and data services. Landing pages are typically required regardless of the archiving entity (e.g., DAAC, Dryad). Generally, the DOI for a data set would resolve to the data set's landing page. Guidelines for data set landing pages exist (e.g., DataCite, 2020).

4.3 Archive Content

4.3.1 Data

The research team is encouraged to plan early in the project design regarding data collection, as well as what data and metadata should be archived. This will allow the team to design for appropriate data collection during the project and begin gathering data in a manner that is consistent with best practices. Section 2 provides recommended standards for data and metadata content. Section 2 holds precedence, however, another useful resource for general guidance is the NASA document, Data Management Standards and Best Practices for NASA-sponsored Citizen Science Investigations (requires Earthdata login) (Ramapriyan, 2018).

Key recommendations from this document include:

- Identify and establish contact with a NASA-designated archive early in the project.
- Prepare a data management plan that incorporates data collection and management aspects based on discussions with the archive center.
- Establish and follow standards for data and metadata.
- Develop a plan for data preservation.

Upfront planning streamlines data collection, compilation, and formatting activities and minimizes overall effort. Ramapriyan (2018) also provides additional external resources and efforts underway in the citizen science community.

While this document is specific to NASA-designated archives, its guidance is generally applicable to non-NASA archives as well (Sections **Error! Reference source not found.** and 4.4). For information on designating an archive location for project data sets (Section 4.3.8). The development of a data management plan may also depend on the project needs and the specific program's requirements. It would be useful, however, for the research team to think through the contents of a data management plan and develop a draft document, because it would help plan efficient data collection, storage, processing, and other tasks. For a template of a data management plan, see Appendix A of that document (Ramapriyan, 2018).

4.3.1.1 Data Types and Formats

Research teams are strongly encouraged to plan early in the project what data types and formats will be generated and/or archived. Data types include qualitative (e.g., long text, binary variables [yes/no]), quantitative, image, audio, or some combination thereof. The type of data may dictate what data format may be useful. Appropriate data formats include comma-separated values (CSV), netCDF, HDF, etc.

Section 2.3 discusses in detail the recommended standards for data formats and for making the data FAIR. It is important that the archived data are in formats that are stable over the long term, are easily accessible using commonly available software (open-source or freely available where feasible), enable interoperability, and are independent of software version changes over time. Certain data formats, such as netCDF, incorporate metadata as part of the data file, thus

making the formats self-describing and machine-independent. In general, archived data should be “archive-stable” and “self-describing.”

4.3.1.2 Data Processing Levels

Data processing levels of NASA's Earth Observing System Data and Information System (EOSDIS) data products range from Level 0 to Level 4. Level 0 products are raw data at full instrument resolution. At higher levels, the data are converted into more useful parameters and formats (see NASA Earth Science Data Processing Levels (2019) for full definitions). Data processing levels are required for data archived at a DAAC. Typically, the DAAC determines the mapping of the data to the data processing levels. For data archived at a non-DAAC location, the mapping of data processing levels may not apply. This section provides some general guidance for PI on mapping citizen science data to the data processing levels. Table 4-1 provides abbreviated definitions from the National Snow and Ice Data Center (NSIDC, 2013).

Table 4-1. Abbreviated Data Processing Levels

Level	Definition
0	Unprocessed instrument data
1A	Unprocessed instrument data alongside ancillary information
1B	Data processed to sensor units (e.g., brightness temperatures)
2	Derived geophysical variables (e.g., sea ice concentration)
3	Variables mapped on a grid (e.g., data using EASE-Grid)
4	Modeled output or variables derived from multiple measurements

Why should data processing levels be applied to citizen science data?

The main rationale for applying data processing levels to citizen science (CS) data is to be compatible with existing NASA data archives. If a CS data set is archived at a NASA DAAC, data processing levels are required as a field in the metadata record submitted to the Common Metadata Repository (CMR) (NASA EarthData, 2020), the metadata system for EOSDIS. If a CS data set resides at a non-NASA archive, data processing levels, though not required, are still highly recommended. They help users understand the extent of processing that the raw measurements have undergone to result in the data set of interest.

How applicable are data processing levels to citizen science data?

The extent to which the existing data processing levels, as defined above, are applicable to CS data depends on whether the definitions are strictly applied or not. If strictly applied, then CS data can only be partially mapped to the existing levels. This is because of some basic differences between CS data and the typical NASA satellite data. For example, for CS, Level 0 unprocessed instrument data may not commonly be accessible or be used by citizen scientists. If the definitions are loosely applied, however, then CS data can be meaningfully mapped to some equivalent of the existing levels. It is up to the CS project PIs to apply the definitions of the EOSDIS product levels and map their particular CS data products to these levels.

The following two examples show such mappings for Twitter data. Figure 4-1 presents a general notional mapping for tweets. Level 0 Twitter data, e.g., might be keyword-filtered tweets

with time labels. Figure 4-2 is a similar mapping for the case of tweets binned to the Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) grid.

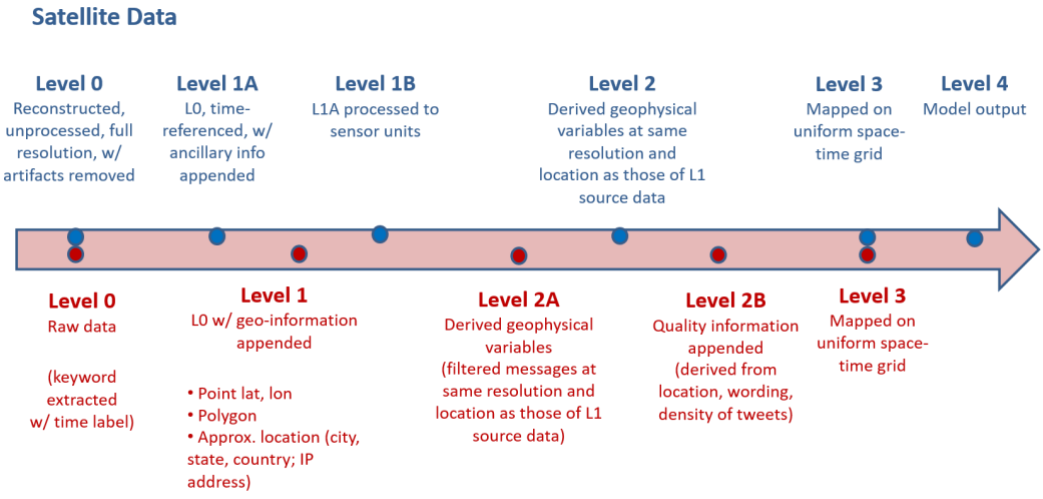


Figure 4-1. General notional mapping of citizen science data (tweets) to equivalent NASA EOSDIS data processing levels.

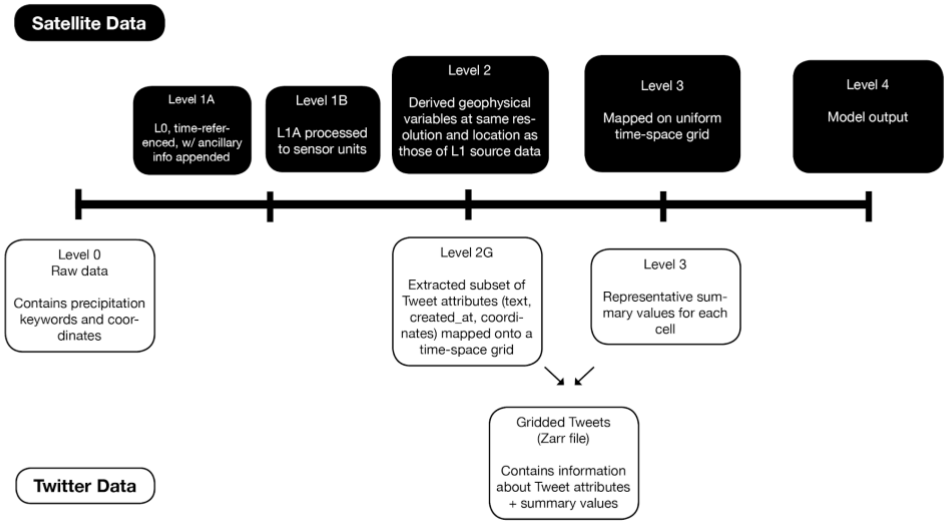


Figure 4-2. Notional mapping of citizen science data (tweets) to equivalent NASA EOSDIS data processing levels for the case of tweets binned to the Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) grid.

4.3.1.3 Data Maturity Levels

Data maturity levels of NASA's Earth Observing System Data and Information System (EOSDIS) data products provide guidance on suitability for use, as the data go through the validation process. Full definitions are available in NASA Earth Science Data Maturity Levels (2020). Table 4-2 provides a summary of the maturity level definitions.

Table 4-2. Summary of Data Maturity Levels

Level	Definition
Beta	Products for gaining familiarity
Provisional	Products for preliminary exploration
Validated	Products fully validated and quality checked, suitable for studies and publications
Validated, Stage 1	Product accuracy estimated based on small number of locations and time periods
Validated, Stage 2	Product accuracy estimated based on significant number of locations and time periods
Validated, Stage 3	Product accuracy assessed and uncertainties quantified
Validated, Stage 4	Stage 3 systematically updated

The above definitions are primarily in the context of validation of data product algorithms for NASA satellite missions and thus do not fully apply to citizen science (CS) data. The concept of data maturity, however, is still useful for CS data. To apply data maturity levels to CS data, suggested modifications to the current definitions include:

- Beta. Data (possibly synthetic) for gaining familiarity.
- Provisional. Data as collected (e.g., ground photographs, lake height readings, snow depth measurements, sensor-reported $PM_{2.5}$ concentrations, “precipitation”-filtered tweets) and processed (e.g., 1-min raw sound recordings, bird data at specific locations, and bird occurrence map, corresponding to data processing Levels 0, 1b, and 3, respectively).
- Validated, Stage 1. Data quality-checked at the original temporal resolution (if applicable) following best practices, as described in Section 2.3.5.
- Validated, Stage 2. Data adjusted based on existing reference methods (e.g., correcting $PM_{2.5}$ sensor bias and drift based on comparisons with reference monitors).
- Validated, Stage 3. Stage 2 data, systematically updated.

4.3.1.4 Metadata

All data stored in a long-term archive should be accompanied by relevant metadata. Section 2 provides information about the content of metadata and Section 4.4 includes details about how to provide the metadata to the archives. Depending on the project and the type of data, some of metadata may be embedded in the data file (e.g., spatial and temporal information, quality flags). Other types of metadata may be contained in separate files, such as documentation of

known issues with measurements over time. In such cases, linkage should be provided between the data set in question and the separate files by, for example, including URLs within the embedded metadata.

4.3.2 Code

4.3.2.1 NASA Open-source Policy

NASA requires all software developed through its funding be open-source. This requirement applies as well to citizen science projects funded by the Citizen Science for Earth Systems Program (CSESP). See NASA ESDS Open Source Software Policy (2019) for NASA's policy on open-source software. Similarly, NASA has developed a policy on sharing of data, services, and software through its funding (NASA Open Data, Services and Software Policies, 2019). Section 1 provides for guidance about open-source policy for data and any exceptions for citizen science data (e.g., if it contains sensitive information).

NASA's open-source software policy requires that software be associated with a "permissive license" that allows for free use, modification, and redistribution. The policy also requires that codes be delivered to a publicly-accessible repository such as GitHub. The NASA Earth Science Data Systems (ESDS) program recommends developing software using NASA's GitHub repository (NASA GitHub, 2020).

A good complementary reference, in the form of a review checklist, is available from the Journal of Open Source Software (JOSS, 2020). Following such a checklist should improve code reusability.

4.3.2.2 Code Format

All code developed using any software, including any proprietary software, should be viewable using a standard text editor or browser.

4.3.2.3 Code Documentation

All code should be well-documented and follow good coding practices. Good software documentation (e.g., continuous documentation in an agile environment [TechBeacon, 2020; InfoQ, 2014; Nuclino, 2018]) helps in long-term maintenance, enables reproduction by the larger scientific community, and allows further refinements and improvements to the code.

All code should, at a minimum, contain:

- A header that describes the purpose of the code, function, or module, and a running history of version changes (date modified/created, author, reason for making the change, and version number if applicable).
- In-line documentation describing what each section does and how it works.
- Specification of any assumptions.
- Definition of variables in the code, at least for key input and output variables.
- References to literature and papers for material used in the code.

4.3.3 Project Software User Guide

"Project software" as used here includes stand-alone software and online application programming interfaces (APIs). All project software should be accompanied by a user guide that describes the software installation process and guides the user on how to run the software. User guides should incorporate visuals for specific steps to illustrate actions. It is also valuable to provide example data sets (inputs and outputs) for the user to verify software functionality.

User guide should have a version history and clearly identify the software version to which the guide applies.

4.3.4 Documentation

All data submitted for long-term archival should include documentation (Section 2.2).

How data producers can submit material to be included in data documentation is provided in Section 4.4.

How documentation should be delivered or made available (e.g., auto-attach README files to data) when users access data is provided in Section 4.5.

4.3.5 Digital Object Identifier

DOIs are unique, persistent identifiers that do not change over time. If the location of the data set associated with a DOI changes, the pointer to the data set in the DOI metadata would be updated accordingly. Obtaining a DOI requires working with an organization that has a contractual arrangement with one of the DOI registrars. In almost all cases, an organization accepting the long-term responsibility for the data would assign the DOI. Options for obtaining a DOI depend on whether the data are archived at a NASA Distributed Active Archive Center (DAAC) or elsewhere and are described in Sections 4.3.5 and 4.3.8. If the data are archived at a DAAC, the DAAC would assign a DOI. If the data are archived somewhere other than a DAAC, a DOI should be assigned by the archiving organization.

4.3.6 Privacy Considerations

Depending on the nature of the project, citizen science data may include personally identifiable information (PII), sensitive PII (SPII), and/or sensitive content. Section 1 details best practices for managing such citizen science data.

If a citizen science project contains proprietary data, refer to Section 4.4.2.1 for guidance on embargo period before the data have to be archived and made public.

4.3.7 Archive Location Goals

The specific goals for storing citizen science project data in a designated quality archive location are to ensure the data are 1) secure, 2) discoverable, 3) long-term accessible, and 4) open and freely available. The following subsections provide additional guidelines. Project data management plans should include designation of an archive location, description of data to be archived, and timeline for interacting with the designated archive.

4.3.8 Archive Location Selection

As indicated in the CSESP Call for Proposals (ROSES 2016, A.47), “data from projects selected for full implementation will be archived at a NASA designated data center, following a successful peer review of data quality.” On behalf of NASA, the CSESP Manager designates the organization responsible for long-term archiving of data from a given CSESP project. Project PIs are strongly encouraged to discuss specific preferences on the location for long-term archival with the CSESP Manager well in advance of initiating the archival process. The following sections detail considerations that assist in making such designations.

4.3.8.1 NASA Archives

Distributed Active Archive Centers (DAACs) are one approach NASA uses to archive Earth science data collected during NASA-funded projects, as well as data from selected other projects essential to NASA's Earth science mission.

Storing NASA-funded citizen science project data within an existing NASA DAAC achieves the stated goals and helps ensure the citizen science data are discoverable by users of NASA data, providing greater visibility to the citizen science data. Each DAAC has a process for proposing a data set for archival at that DAAC, and NASA is currently developing a process to allow projects to propose archival without specifying a particular DAAC. Projects interested in archiving data at a DAAC should contact that DAAC as early in the project as possible to determine what approvals, if any, are required.

NASA also offers non-DAAC archives for selected projects. For example, the NASA citizen science data repository, GLOBE (2020), has been operating since 1994, collecting and archiving data from observations by teachers and students around the globe. Section 4.4 outlines the process and timeframe for archiving data within a NASA data archive and provides recommendations on a timeline for non-NASA archive locations.

4.3.8.2 Non-NASA Archives

If approved by NASA CSESP, non-NASA archives may be an alternative to NASA DAACs, provided they achieve the goals stated in Section 4.3.7 and meet the guidelines detailed in Section 4.3.9. There may be compelling reasons to store data in a non-NASA archive, including privacy concerns, ongoing data collection, special archive requirements, etc. In these cases, the first preference is for an archive process that, once set up, does not depend on actions by the PI. The possibility and appropriateness of non-DAAC archival options may vary from project to project. PIs should discuss with their organizations and with the CSESP Manager regarding non-DAAC archival prior to initiating the archival process.

Some examples of non-NASA archives or distributed network of archives include Zenodo, Harvard Dataverse, Dryad, DataONE, institutional repositories, and discipline-specific repositories.

- Refereed journal: The final project data set could be published in a refereed scientific journal in a manuscript that describes the project and the data. New journals are emerging to support the publishing of scientific data sets (e.g., Data in Brief, Earth and Space Science, Scientific Data, Earth System Science Data, and Geoscience Data Journal). The journal would assign a DOI for the data and manuscript at the time of publication. Data set size limits may apply.
- Dryad: "Dryad is a nonprofit organization that provides long-term access to its contents at no cost to researchers, educators or students, irrespective of nationality or institutional affiliation. Dryad is able to provide free access to data due to financial support from members and data submitter." Dryad does require that the data being archived be associated with a published peer-reviewed article. The advantage of Dryad is that it gives PIs more freedom as to where they might publish their manuscripts. Dryad is journal agnostic. Dryad charges extra fees on the order of $100 for data sets exceeding 50 GB. Dryad assigns a DOI to data sets upon archival.
- Zenodo: Zenodo is part of the OpenAIRE Project, which is "in the vanguard of the open access and open data movements in Europe [and] was commissioned by the EC (European Commission) to support their nascent Open Data policy by providing a catch-

all repository for EC funded research." Zenodo, which is free and open source, provides DOIs on request for scientific data sets.

- Harvard Dataverse: Harvard Dataverse is a free, open repository for long-term archival of scientific data sets. It assigns a DOI to data sets upon archival. PIs can archive a data set without an associated peer-reviewed publication. Harvard Dataverse supports data set versioning, should the PI wish to archive new or modified versions of the original data set.
- DataONE: DataONE is a distributed network of member repositories but does not actually archive submitted data. Citizen science project PIs would need to get project data stored in one of the DataONE member nodes. Or, if data are stored in the PI's institution, the institution would need to become a member of DataONE.

Factors to consider when selecting a non-NASA archival location include:

- First preference is a DOI-issuing archive center specializing in scientific data.
- Second preference is for a long-term archive at a university or for some other institutional collection designed and run by data archive specialists.
- Third option, a long-term archive on a dedicated project server, may be viable but only if the project is well-resourced and the archive is designed to be ongoing.
- Discipline-specific archive locations could also be options to increase the visibility/discoverability of the data (e.g., the Avian Knowledge Network for avian data).

In general, project-specific repositories would not be considered adequate for long-term archival, because they may not meet the stated archival goals (Section 4.3.7) and guidelines (Section 4.3.9). Specific concerns relate to the long-term viability of the project repository, i.e., long-term operational support, adequate processes for maintaining storage system backups, maintenance and security, staff to provide operational support, and ability to provide adequate data discovery.

4.3.9 Guidelines for Long-term Archive Location (Non-NASA Archives)

Long-term archives should support archival best practices, including the following capabilities and features:

1. Storage of data in formats conducive to long-term accessibility (i.e., non-proprietary formats).
2. Clear path and support for long-term operations.
3. Free and open access to data.
4. Adequate capabilities to support storage and search of project- and data-specific metadata.
5. Adequate capabilities to support versioning if revised data sets become available.
6. Adequate processes, backups, and offsite/secondary storage facilities to limit the possibility of data loss or hacking.
7. Adequate best practice system security protections, including regular patching, auditing, and review. Security includes network, physical, and server-level security.
8. Staff/support to answer operational questions and a process to direct technical questions to the relevant knowledgeable source (e.g., maintain an up-to-date list of technical contacts).
9. Interaction with other search engines and repositories to maximize discoverability.
10. Capability to issue DOIs (Section 4.3.5). A DOI associated with a data set should be specified in the metadata for that data set (Section 2.3.1).
11. Succession planning for data migration in the event of system closure.

4.4 Archive Process and Timeframe

4.4.1 What is the overall process for archiving your data and code?

The process of archiving data depends on the selection of a NASA vs. non-NASA archive, as summarized in Section **Error! Reference source not found.**. In the case of non-NASA archives, the process depends on the archiving organization and is not described here. Minimum criteria for non-NASA archives are described in Section 4.3.9.

In the case of archiving at a NASA DAAC, data producers should contact the appropriate DAAC as early in the project as possible. They should provide information on the characteristics of the data set, relationships to existing data sets archived at NASA or other organizations, desired services from the DAAC (e.g., search/download, subsetting, conversion, visualization), data access and privacy considerations, and desired public release date. Data producers should also provide appropriate metadata for all submitted data sets, at the granule (file) and data set level, along with associated README files. The standards for these metadata may differ from those for other NASA data (Section 2). There is no DAAC that is specifically tasked with handling citizen science data. Data producers should contact the DAAC most suitable for the subject matter of their projects (e.g., ORNL DAAC for biogeochemistry and PO.DAAC for physical oceanography).

Because many citizen science projects collect data on an ongoing basis, their data sets often require updates. DAACs may, in some cases, be able to accommodate near-real-time updates to data sets. In other cases, it may be best to provide updates on a fixed schedule or at project milestones. In any case, it would be important for citizen science data producers to coordinate with their respective designated DAACs to ensure appropriate archive of updated, augmented, or reprocessed data sets.

4.4.2 Guides for the Data Production Process

The details of producing citizen science data will depend on the exact nature of the project (e.g., field data collection vs. products derived from remote sensing). However, existing guides for NASA data producers can provide some context that may be useful for citizen science projects.

Examples include:

- GES DISC, "GES DISC Data and Metadata Recommendations to Data Providers"
- ORNL DAAC, "ORNL DAAC Detailed Submission Guidelines"
- ESDIS, Data Product Development Guide for Data Producers (in preparation), ESDS-RFC-041, URL (TBD)

4.4.2.1 Archive Timeframe

Short-term archive and distribution of data is largely at the discretion of individual citizen science projects, so long as all data are preserved and data are made available within a reasonable timeframe. For projects where it is feasible to do so, data should be made available in near real time. For projects where distributable data require additional processing, it should generally be made available as soon as the processing is complete. There may be exceptions to these guidelines in some cases. There should be no embargo period during which the distributable data are withheld from the scientific community and the general public, to be consistent with NASA's data and information policy (2019).

Data should be fully archived at their final archive location (whether a DAAC or elsewhere) no later than the due date for the final deliverable(s) for the project. This would generally require appropriate planning early in the project development phase and beginning the archive process

well in advance of project completion. This planning should include communication with the archive location on general timeframe and any requirements on the project.

4.4.2.2 Archive Length

All data should ultimately be archived permanently in an accessible and searchable location. Exceptions to this guideline include raw data that might generate privacy concerns (e.g., phone numbers of participants, voice recordings) and intermediate processing steps that do not need to be retained permanently.

In some cases, reprocessing of data may be required, generating several different versions. The nature of this reprocessing may vary from project to project. What is important is to be clear about choices. In general, minor changes (e.g., corrections of minor data errors) should be documented but do not require the generation of an entirely new version of the data set. Major revisions, additions, or modifications should generally produce a new version, with at least one prior version retained in the archive for a defined overlap period.

4.5 Providing Data Access and Distribution Services

Requirements for data access and distribution services provided by NASA archives are well established across all the DAACs (NASA ESDIS ADURD, 2020). Project data archived at a DAAC should be compliant with these requirements. Data archived at non-NASA locations should conform, as much as possible, to these requirements. Regardless of how the data are stored, accessed, and distributed, the goal is to satisfy the end user desire for data in as ready to use a format as possible. Towards that end, the related concepts of Analytics Optimized Data Stores (AODS) (NASA, 2018), Analysis-Ready Data (ARD), and data warehouse apply to NASA and non-NASA archives, as well as to NASA mission data and citizen science data (Section 3.1.3).

The following recommendations and issues regarding data access and distribution services are grouped into those services provided by NASA archives and those by non-NASA archives.

4.5.1 Data Access and Distribution Services for NASA Archives

The means of data access and related distribution services would be those of the designated NASA archive, either one of the DAACs or the Global Learning and Observation to Benefit the Environment Data and Information System (GLOBE, 2020). Section 2.3.3 (Data Distribution) of the NASA ESDIS ADURD (2020) document lists the distribution requirements for NASA DAACs. A potential issue regarding citizen science data at some NASA archives is the integration of these non-standard/irregular data into the archives.

ADURD provides common requirements for EOSDIS-supported data, i.e., primarily NASA satellite mission data archived at DAACs. As such, not every requirement in ADURD would apply or fully apply to citizen science data. For example, item 8 of Section 2.3.3 (NASA ESDIS ADURD, 2020), “The XDS shall distribute data to various data processing systems, instrument teams’ science computing facilities, SIPS, and other DAACs to support product generation and quality assurance in a timely manner to support production schedules,” would not currently apply to citizen science data.

A typical suite of access and distribution services at NASA DAACs includes HTTPS online access directly from the archive, Earthdata Search (to access data across multiple DAACs), visualization/analysis/download (e.g., Giovanni at GES DISC, AppEEARS at LPDAAC), and other subsetting/analysis services (e.g., OPeNDAP, GDS).

4.5.2 Data Access and Distribution Services for Non-NASA Archives

For project data stored at non-NASA archives, the goal is to conform, as much as possible, with the requirements of data access and distribution services of NASA archives. Existing NASA user communities are familiar with these services; thus, conforming with these standards would make the project data more readily usable by NASA data users.

The recommended minimal set of services non-NASA archives should provide to enable/facilitate access and distribution (Section 4.3.9) includes:

- Online access directly from the archive.
- Search and subset by space, time, and variable.
- Documentation
    - README to accompany data download.
    - Directory Interchange Format (DIF) documents (EOSDIS "collection" or data set metadata) published to Common Metadata Repository (CMR) (See example from GES DISC).
- Visualization (for browsing and quick looks).
- Data download in recommended and user-desired formats.

Conclusion

This document provides guidelines about the standards for Earth Science citizen science data. Included guidelines address: legal and policy issues, standards, usability, and long-term archival. The guidelines presented reflect best practices assembled by practitioners of NASA-funded Earth Science citizen science programs/projects and members of the NASA ESDS community. The guidelines are intended for pre-proposal and post-award data producers/providers, the former to inform programmatic expectations while proposals are prepared and the latter to assist in the conduct of the awarded projects.

Managing citizen science data, as opposed to satellite or typical field campaign data, can come with its own set of legal and policy considerations. NASA promotes full and open sharing of all data with the research and applications communities, private industry, academia, and general public. NASA-funded data producers/providers, including citizen scientists, are expected to comply with NASA's Open Data Policy. Some citizen science projects may collect personal information (e.g., email address). Submission of such information by a participant is strictly voluntary, and projects are encouraged to grant participants the ability to opt out. It is strongly recommended that citizen science projects do not request or store Sensitive Personal Identifiable Information (SPII), such as social security numbers or biometric identifiers. Citizen science projects are strongly encouraged to post a Terms of Use statement, including a liability clause, on their website and/or mobile application (Section 1.3). Even when a project does not directly collect PII, data collection itself, such as around a person's residence, may provide enough information to defeat anonymization attempts. It is strongly encouraged that the Terms of Use clearly state the project's data ownership policy.

The Standards section establishes a basic set of guidelines such that NASA Earth Science citizen science data can be "findable, accessible, interoperable and reusable (FAIR)", as described by Force11.org. Projects can achieve this goal by providing participants with standardized measurement protocols, also known as Standard Operating Procedures (SOPs), and documentation clearly stating instrument operation, data collection methods, data processing, and data contents. Projects are encouraged to include sufficient metadata and use standard measurement units, data formats, and data structures for the field of study. Citizen science data producers/providers are encouraged to document and communicate their data

quality and quality assurance procedures. DataONE provides a useful and commonly used framework.

Citizen science data, particularly data relevant to the Earth and environmental sciences, can have tremendous scientific and societal value, particularly with longer periods of record, time since data collection, and higher degrees of FAIRness. Achieving that value, however, requires planning, particularly up-front planning, and effort. Far less effort is needed to realize the full value of data, including citizen science data, when the target archive, archival processes, relevant standards, and best practices are understood at the start of the project, and efforts are made to ensure that data, code, and documentation will be preserved after the project's completion. Unfortunately, at the time of this writing, it is often the case that the appropriate destination archive and relevant archival processes will not be clear to a project's leaders. However, a consideration of usability and best practices can be used, as described in Chapters 3 and 4, and Principal Investigators can work with the cognizant Program Officers to determine the appropriate archival destination. Particularly for projects which intend to distribute their data themselves during the project's lifetime, an understanding of the likely destination is important for managing digital identifiers (particularly DOIs), data formats, and discovery metadata.

References

Bowser, Anne, Elizabeth Tyson 2015, “Crowdsourcing, citizen science, and the law: legal issues affecting federal agencies”, https://www.wilsoncenter.org/sites/default/files/executive_summary_gellman.pdf.

Bowser, Anne, Peter Brenton, Rob Stevenson, Greg Newman, Sven Schade, Lucy Bastin, Alison Parker, and Jessie Oliver 2017, “Citizen Science Association Data & Metadata Working Group: Report from CSA 2017 and Future Outlook” Available at: https://www.wilsoncenter.org/sites/default/files/wilson_171204_meta_data_f2.pdf.

Children’s Online Privacy Protection (COPPA), COPPA, accessed October 25, 2019, http://www.coppa.org/comply.htm.

Climate and Forecast (CF) Metadata Conventions, NASA ESDIS, accessed January 28, 2020, https://earthdata.nasa.gov/esdis/eso/standards-and-references/climate-and-forecast-cf-metadata-conventions.

Daines, Gary, Media Usage Guidelines, NASA, accessed January 28, 2020, http://www.nasa.gov/multimedia/guidelines/index.html.

Data and Information Policy, NASA, January 28, 2020, https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy.

Data Product Development Guide for Data Producers, NASA ESDS Data Product Developers Guide Working Group, (2020) in review.

DataONE, Best Practices, accessed January 28, 2020, https://www.dataone.org/best-practices .

DataCite, 2020. “Best Practices for DOI Landing Pages”, https://support.datacite.org/docs/landing-pages (accessed January 30, 2020).

DataONE, Data Management Guide for Public Participation in Scientific Research, DataONE Public Participation in Scientific Research Working Group (February 2013). https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf.

Duley, J., 2016: NASA Guidelines for Quality of Information. NASA,. http://www.nasa.gov/content/nasa-guidelines-for-quality-of-information (Accessed October 25, 2019).

ESIP Data Preservation and Stewardship Committee. 2019. “Data Citation Guidelines for Earth Science Data. Ver. 2.” Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816.

ESO, Data Quality Working Group’s Comprehensive Recommendations for Data Producers and Distributors, ESDS-RF-033, accessed January 30, 2020, https://earthdata.nasa.gov/esdis/eso/standards-and-references/recommendations-from-the-data-quality-working-group.

Evans et al., ASCII File Format Guidelines for Earth Science Data, NASA, ESDS-RFC-027v1.1 (May 2016). https://cdn.earthdata.nasa.gov/conduit/upload/4827/ESDS-RFC-027v1.1.pdf.

General Data Protection Regulation (GDPR), Everything you need to know about the “Right to be forgotten, 2018.” GDPR.eu,. https://gdpr.eu/right-to-be-forgotten/ (Accessed October 25, 2019).

Global Learning and Observations to Benefit the Environment (GLOBE) Program, GLOBE Data User Guide v1, (July 2019). https://www.globe.gov/globe-data/globe-data-user-guide .

Guidelines for Quality of Information, NASA, January 28, 2020. https://www.nasa.gov/content/nasa-guidelines-for-quality-of-information.

Foody, G, See, L, Fritz, S, Mooney, P, Olteanu-Raimond, A-M, Fonte, C C and Antoniou, V (eds.), Mapping and the Citizen Sensor. London: Ubiquity Press (2017). https://doi.org/10.5334/bbf.

Fox, S., NASA Web Privacy Policy and Important Notices, NASA, accessed January 28, 2020. http://www.nasa.gov/about/highlights/HP_Privacy.html.

Freitag, A., Meyer, R. and Whiteman, L., Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data. Citizen Science: Theory and Practice, (2016):1(1), p.2. http://doi.org/10.5334/cstp.6.

GLOBE, 2020. “The GLOBE Program”, Available at: https://www.globe.gov/en (accessed January 30, 2020).

Department of Homeland Security, 14 Dec 2018, Privacy Policy for DHS Mobile Applications, (December 2018): 047-01-003, Revision 1. https://www.dhs.gov/sites/default/files/publications/047-01-003.pdf.

Department of Homeland Security, 4 Dec 2017, Handbook for Safeguarding Sensitive PII, (December 2017):047-01-007, Revision 3. https://www.dhs.gov/sites/default/files/publications/dhs%20policy%20directive%20047-01-007%20handbook%20for%20safeguarding%20sensitive%20PII%2012-4-2017.pdf.

Goddard Earth Sciences Division Data and Information Services Center (GES DISC), NASA, README Document for North American Land Data Assimilation System Phase 2 (NLDAS-2) Products, (October 2019). https://hydro1.gesdisc.eosdis.nasa.gov/data/NLDAS/README.NLDAS2.pdf.

Goddard Earth Sciences Division Data and Information Services Center (GES DISC), Data and Metadata Recommendations to Data Providers, (November 2017). https://docserver.gesdisc.eosdis.nasa.gov/public/project/DataPub/GES_DISC_metadata_and_data_formats.pdf.

InfoQ, 24 July 2014, “A Roadmap to Agile Documentation”, https://www.infoq.com/articles/roadmap-agile-documentation, (Accessed 16 Jan 2020).

International Organization for Standards (ISO), ISO 19157:2013 Geographic Information - Data Quality, (2013). https://www.iso.org/standard/32575.html.

International Organization for Standards (ISO), ISO 8601-1:2019 Date and Time Format, (2019). https://www.iso.org/iso-8601-date-and-time-format.html.

International Organization for Standards (ISO), ISO 9241-11:2018 Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts, (2018).

https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

IPTC, Photo Metadata Standard 2019.1, January 28, 2020, https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata.

Jelenak, A., P.J.T. Leondard et al., Dataset Interoperability Recommendations for Earth Science: Part 2, NASA, ESDS-RFC-036 (April 2019). https://cdn.earthdata.nasa.gov/conduit/upload/11261/ESDS-RFC-036.pdf.

JOSS (2020). “Review Checklist”, Available at: https://joss.readthedocs.io/en/latest/review_checklist.html, (accessed, January 30, 2020).

Landslide Reporter, NASA, accessed January 28, 2020,. https://pmm.nasa.gov/landslides/index.html.

Lewandowski, E. and Specht, H., Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology, (2015): 29: 713-723. doi:10.1111/cobi.12481.

McNutt, M.K, et al., Transparency in authors’ contributions and responsibilities to promote integrity in scientific publication. Proceedings of the National Academy of Sciences. (2018): 115(11): 2557-2560. https://doi.org/10.1073/pnas.1715374115.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B. and Stafford, S.G., Nongeospatial metadata for the ecological sciences. Ecological Applications, (1997) 7: 330-342. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

NASA, 2018. “Enabling Analytics in the Cloud for Earth Science Data”, Workshop Report 21-23 February 2018. Available at: https://github-production-repository-file-5c1aeb.s3.amazonaws.com/101820659/1903771?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20200131%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20200131T172116Z&X-Amz-Expires=300&X-Amz-Signature=3d7b9f58ec6577b7373d058195c4c62a281a9e69d6aed49a47c22aafee5a965e&X-Amz-SignedHeaders=host&actor_id=0&response-content-disposition=attachment%3Bfilename%3DCloud.Analytics.Workshop.Report.pdf&response-content-type=application%2Fpdf (accessed January 31 2020).

NASA EarthData, 2020 “Common Metadata Repository”, Available at: https://earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr”, (Accessed January 30, 2020).

NASA Earth Science Data Processing Levels, accessed 25 October 2019, https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels.

NASA Earth Science Data Maturity Levels, Available at: https://science.nasa.gov/earth-science/earth-science-data/data-maturity-levels/, (accessed, January 30, 2020).

NASA ESDS Open Source SoftwarePrivacy Policy, 2019. Available at: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/esds-open-source-policy (accessed, January 30, 2020).

NASA ESDS Data Product Development Guide for Data Producers, ESDS-RFC-041 (in preparation).

NASA Open Data, Services and Software Policies, 2019. Available at: https://earthdata.nasa.gov/collaborate/open-data-services-and-software (accessed, January 30, 2020).

NASA Data and Information Policy, 2019. Available at: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy (accessed January 31, 2020).

NASA GitHub, https://github.com/nasa, (accessed, January 30, 2020).

NASA ESDIS ADURD, 2020. “Archiving, Distribution, and User Services Requirements Document (ADURD). Available online at: https://earthdata.nasa.gov/esdis/esdis-policy/adurd#adurd-archive-dist (accessed January 31, 2020).

NASA-NSPIRES, 23 May 2016, ROSES 2016: A.47 CITIZEN SCIENCE FOR EARTH SYSTEMS PROGRAM - Solicitation https://nspires.nasaprs.com/external/viewrepositorydocument/cmdocumentid=507108/solicitationId=%7B96C8752A-37DF-B46A-C2C3-3F0EC4C599E9%7D/viewSolicitationDocument=1/A.47%20CSESP%20FAQ%20posted.pdf (Accessed 16 Jan 2020).

NASA Policy Directive 1382.17J, January 28, 2020.

NASA Section 508 Standards, accessed January 28, 2020. https://www.nasa.gov/accessibility/section508/sec508_standards.html

NASA SMD-33, NASA Science Mission Directorate Policy Document 33: Citizen Science (December 2018). https://smd-prod.s3.amazonaws.com/science-red/s3fs-public/atoms/files/SPD%2033%20Citizen%20Science.pdf.

NSIDC, 2013. “Is it 1B, 2 or 3? DefinitionsUnited States Code Homepage, Office of data processing levels”, The DRIFT News & Tips for Data Users, National Snow & Ice Data Center, Available at: https://nsidc.org/the-drift/2013/08/is-it-1b-2-or-3-definitions-of-data-processing-levels/ (the Law Revision Counsel, accessed, January 30, 2020).

Nuclino, 21 Dec 2018, “Agile Development Methodology: To Document or Not to Document?”, https://blog.nuclino.com/agile-development-methodology-to-document-or-not-to-document, (Accessed 16 Jan 2020).

OLRC Home. October 25, 2019. https://uscode.house.gov/browse.xhtml.

ORNL DAAC, Detailed Submission Guidelines, accessed January 28, 2020, https://daac.ornl.gov/submit/.

PPSR_CORE, PPSR_CORE Metadata Standard, Citizen Science Association (Fall 2013), https://www.citizenscience.org/2015/10/09/ppsr_core-metadata-standard/.

Ramapriyan, H. K., 29 Jan 2018, “Data Management Standards and Best Practices for NASA-sponsored Citizen Science Investigations”, Available at https://wiki.earthdata.nasa.gov/download/attachments/131662879/DM%20Stds%20and%20Best%20Practices%20for%20NASA%20CSESP-20180129.docx?api=v2 (Accessed January 30, 2020).

Rosenthal, Isaac S., Jarrett E. K. Byrnes, Kyle C. Cavanaugh, Tom W. Bell, Briana Harder, Alison J. Haupt, Andrew T. W. Rassweiler, et al, Floating Forests: Quantitative Validation of Citizen Science Data Generated from Consensus Classifications. (2018) ArXiv:1801.08522 [Physics, q-Bio], http://arxiv.org/abs/1801.08522.

Teng, W., H. Rui, R. Strub, and B. Vollmer, 2016. “Optimal reorganization of NASA earth science data for enhanced accessibility and usability for the hydrology community”, Journal of American Water Resources Association (JAWRA), 825-835, doi:10.1111/1752-1688.12405.

TechBeacon, “Why Agile Teams should Care about Documentation”, https://techbeacon.com/app-dev-testing/why-agile-teams-should-care-about-documentation, (Accessed 16 Jan 2020).

UDUNITS-2, University Corporation for Atmospheric Research (UCAR), accessed August 10, 2019, https://www.unidata.ucar.edu/software/udunits/udunits-current/doc/udunits/udunits2lib.html#Top

US Environmental Protection Agency (US EPA), Handbook for Citizen Science Quality Assurance & Documentation – Version 1, EPA 206-B-18-001 (2019). https://www.epa.gov/sites/production/files/2019-03/documents/508_csqapphandbook_3_5_19_mmedits.pdf.

Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, (2016): 3, 160018. https://www.nature.com/articles/sdata201618.

Glossary

ASCII	American Standard Code for Information Interchange
CF	Climate and forecast
COPPA	Children's Online Privacy and Protection Act
DataONE	Data Observation Network for Earth
DOI	Digital Object Identifier
EPA	Environmental Protection Agency
EPSG	European Petroleum Survey Group
ESDIS	Earth Science Data and Information System (Project)
ESDS	Earth Science Data System (Program)
ESIP	Earth Science Information Partners
ESO	ESDIS Standards Office
EXIF	Exchangeable Image File Format
FAIR	Findable, Accessible, Interoperable and Re-Usable
GES DISC	Goddard Earth Sciences Data and Information Services Center
GLOBE	Global Learning and Observations to Benefit the Environment
IPTC	International Press Telecommunications Council
IQA	Information Quality Act
ISO	International Organization for Standardization
IT	Information Technology
ITPO	Innovative Technology Partnership Office
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
NAS	National Academy of Sciences
NASA	National Aeronautics and Space Administration
ORNL	Oak Ridge National Laboratory
PII	Personal Identifiable Information

PNG	Portable Network Graphics
PPSR	Public Participation in Scientific Research
PRA	Paperwork Reduction Act
RRI	Responsible Research and Innovation
SMD	Science Mission Directorate
SOP	Standard Operating Procedure
SPII	Sensitive Personal Identifiable Information
TIFF	Tagged Image File Format

Authors/Editors

Alphabetical. Section leads are in bold.

Helen M. Amos
GLOBE Program
Science Systems and Applications, Inc.
NASA Goddard Space Flight Center
Greenbelt, Maryland
helen.m.amos@nasa.gov

Travis Andersen
GLOBE Program
University Corporation for Atmospheric Research
Boulder, Colorado
andersen@ucar.edu

Anthony Arendt
Community Snow Obs
University of Washington
Seattle, Washington
arendta@uw.edu

Jarrett Byrnes
Assistant Professor, University of Massachusetts
Boston, Massachusetts
jarret.byrnes@umb.edu

Matthew Clark
Professor, Sonoma State University
Rohnert Park, California
matthew.clark@sonoma.edu

Lisa Dallas
GLOBE Program
NASA Goddard Space Flight Center
Greenbelt, Maryland
lisa.m.dallas@nasa.gov

Narendra Das
NASA Jet Propulsion Laboratory
Pasadena, California
narenda.n.das@jpl.nasa.gov

Prakash Doraiswamy
Project: Can Citizen Science and Low-Cost Sensors Help Improve Earth System Data? Implications to Current
and Next Generation of Space-Based Air Quality.
RTI International
Research Triangle Park, NC
pdoraiswamy@rti.org

Robert Levy
Project: Can Citizen Science and Low-Cost Sensors Help Improve Earth System Data? Implications to Current
and Next Generation of Space-Based Air Quality.
NASA Goddard Space Flight Center
Greenbelt, Maryland
robert.c.levy@nasa.gov

Tamlin Pavelsky
Associate Professor
University of North Carolina
pavelsky@unc.edu

David Overoye
GLOBE Program
Science Systems and Applications, Inc.
Pasadena, California
david.overoye@ssaihq.com

Hampapuram Ramapriyan (Rama)
Science Systems and Applications, Inc. and
Earth Science Data and Information System (ESDIS) Project
NASA Goddard Space Flight Center
Greenbelt, Maryland
hampapuram.ramapriya@ssaihq.com
0000-0002-8425-8943

Leonardo Salas
Senior Scientist, Point Blue Conservation Science
Petaluma, California
lsalas@pointblue.org

William Teng
ADNET Systems
NASA Goddard Space Flight Center
Greenbelt, Maryland
william.l.teng@nasa.gov

John Volkens
Assistant Professor, Colorado State
Fort Collins, Colorado
john.volckens@colostate.edu

Yaxing Wei
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)
Oak Ridge, Tennessee
weiy@ornl.gov

Bruce E. Wilson
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)

Oak Ridge, Tennessee
wilsonbe@ornl.gov
ORCID: 0000-0002-1421-1728

Appendix B: GLOBE Sample Project Metadata

	Description	Example - Aerosol protocol
Project Information		
	DOI	
	Project Name	GLOBE Aerosols Protocol
	Project Data Provider	The GLOBE Program
	Project Aim	Measure aerosol data from all GLOBE Countries (https://www.globe.gov/globe-community/community-map)
	Project Description	Measure the aerosol optical thickness of the atmosphere (how much of the sun's light is scattered or absorbed by particles suspended in the air). Students point a GLOBE sun photometer at the sun and record the largest voltage reading they obtain on a digital voltmeter connected to the photometer. Students observe sky conditions near the sun, perform the Cloud, Barometric Pressure and Relative Humidity Protocols, and measure current air temperature. More information on the protocol - https://www.globe.gov/documents/348614/e9acbb7a-5e1f-444a-bdd3-acff62b50759
	Project URL	globe.gov
	Project Status	Active
	Project Sponsor(s)	NASA, NSF, NOAA, US Dept of State
	Project POC	The GLOBE Implementation Office
	Project Address	3300 Mitchell Lane
	Project City	Boulder
	Project State	CO
	Project Zip	80301
	Project Email	help@globe.gov
	Project Phone	1-800-858-9947

	Project Website	globe.gov
	Project Social Media	facebook/globe, twitter/globe, instagram/globe
	Project PI	Margaret Pippin
	Project PI Email	mpippin@email.com
	Citation	Global Learning and Observations to Benefit the Environment (GLOBE) Program, date data was accessed, https://www.globe.gov/globe-data
	Terms of Use	Data Free for use with appropriate credit given
	Science Keywords	Air Quality, Particulates, Visibility
	Educator Materials	globe.gov
	Participation Tasks	Measuring, Observing, Site Selection and Description
	Project Equipment/Instrument	Aerosols measuring device - Calitoo, other
	Field of Science	Atmosphere
	Publication Reference	https://www.globe.gov/do-globe/publications
	Metadata Update (YYYYMMDD)	20180901
	Metadata Language	English
Dataset Information		
	Dataset Version Number	1
	Dataset Version Description	Initial release of Aerosols data
	Dataset Last Updated (YYYYMMDD)	20180624
	Data Language	English
	API Help Contact	help@globe.gov

	Variables Measured	Aerosol Optical Thickness (AOT); GLOBE Clouds Measurements also required with submission
	Temporal Extent (Start Date - End Date)	20040114 - present
	Spatial Extent	GLOBE Countries (https://www.globe.gov/globe-community/community-map)
	Project Data API	https://api.globe.gov/search/swagger-ui.html Aerosols data retrieved in the form: https://api.globe.gov/search/v1/measurement/protocol/measureddate/?protocols=aerosols&startdate=2010-01-01&enddate=2011-01-01&geojson=TRUE&sample=FALSE
	Project Data API Documentation	https://www.globe.gov/globe-data/globe-api
	GLOBE Protocol ID	154
Data Information		
	Data Collection Procedure	Aerosols field guide - procedure for collecting data - https://www.globe.gov/documents/348614/a557fa2d-e4cb-429a-9bd4-e049b0ab023c Aerosols Data Entry Form - What data is collected? - https://www.globe.gov/documents/348614/96c7aa35-78ff-42b9-9a4f-fe4766ffaefd Participants are encouraged to enter data daily. Most enter data occasionally.
	Quality	The quality of the AOT measurement is consistent with the capabilities of the instrument used, and the ability of the individual to use the instrument. Participants are asked to record time to "the nearest 15 seconds" Values outside of the range 0 - xx are rejected. The participant is warned and provided the opportunity to change their submission. Three measurements are required at each bandwidth, with the averaged value returned via the API. Individual measurements may be provided by the GLOBE helpdesk if required by a researcher.
	Data Processing	Participants are required to perform 3-4 measurements. The AOT value returned via the API is an average of those measurements. The equations used to calculate AOT from the instrument voltage are available in the Aerosols field guide.