An Ecosystem of Citizen Observatories for Environmental Monitoring

WeObserve D.4.2. Terms of reference of the standards relevant for Citizen Science, gaps and improvements

Work package	WP4: Integrate: Facilitate adoption into EO initiatives
Task	Task 4.2: Support standards, data management and interoperability via the OGC citizen science domain working group
Deliverable Lead	**CREAF**
Authors	Joan Masó (CREAF), Ester Prat (CREAF)
Dissemination level	Public (PU)
Status	Final
Due date	31/05/2019
Document date	15/06/2019
Version number	1.0
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 776740.

Revision and history chart

Version	Date	Main author	Summary of changes
0.1	10/05/2019	Joan Masó	Draft outline
0.2	24/05/2019	Ester Prat	Minor revisions
0.3	30/05/2019	Joan Masó	Drat content
0.4	30/05/2019	Joan Masó, Ester Prat	Final revisions
0.5	13/06/2019	Ester Prat	Comments removal and adding references
0.6	14/06/2019	Ian Mac Callum	Final revision
1.0	14/06/2019	IIASA	Submission

Table of contents

Glossary of terms iv

List of abbreviations and acronyms v

Executive Summary vii

1 Introduction 1

2 List of standards bodies 2

2.1 Open Geospatial Consortium (OGC) 2

2.2 ISO TC 211 3

2.3 W3C 4

2.4 IETF 5

3 The difference between an API and the standards 6

4 Standardization targets 7

5 Standardization for data collected in CS project 8

5.1.1 DMP-1: Discovery 8

5.1.2 DMP-2: Online Access 9

5.1.3 DMP-3: Data Encoding 10

5.1.4 DMP-4: Data Documentation 12

5.1.5 DMP-5: Data Traceability 13

5.1.6 DMP-6: Data Quality-Control 13

5.1.7 DMP-7: Data Preservation 14

5.1.8 DMP-8: Data and Metadata Verification 14

5.1.9 DMP-9: Data Review and Reprocessing 15

5.1.10 DMP-10: Persistent and Resolvable Identifiers 15

6 Discussion on gaps and improvements 16

7 Conclusion and Summary 19

8 References 21

Index of tables

Table 1. Summary of gaps and improvements..17

Table 2. Summary of the recommended standards 18

Glossary of terms

Term	Description
Citizen Science	Scientific research conducted, in whole or in part, by amateur (or nonprofessional) scientists.
Open standard	Standard that is publicly available and has various rights to use associated with it, and may also have various properties of how it was designed (e.g. open process).
Interoperability	Characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, at present or in the future, in either implementation or access, without any restrictions.
API	Set of subroutine definitions, communication protocols, and tools for building software. In general terms, it is a set of clearly defined methods of communication among various components. A good API makes it easier to develop a computer program by providing all the building blocks, which are then put together by the programmer.
GEOSS Data Management Principles	The GEOSS Data Management Principles build on the GEOSS Data Sharing Principles in the sense that they adumbrate what is required in terms of data management to allow data to be shared as Open Data, promptly and at minimum cost.
Metadata	Data [information] that provides information about other data.

List of abbreviations and acronyms

Abbreviation	Meaning
API	Application programming interface
BPEL	Business Process Execution Language
BPML	Business Process Modelling Language
BPMN	Business Process Model and Notation
CEO	chief executive officer
CEOS	Committee on Earth Observation Satellites
CS	Citizen Science
CSW	Catalogue Service For the Web
D&I	Defence & Intelligence
DAB	Discovery and Access Broker
DIY	Do It Yourself
DMP	Data Management Principles
DOI	Digital Object Identifiers
DWG	Domain Working Group
ECSA	European Citizen Science Association
ERCIM	European Research Consortium for Informatics and Mathematics
GEO	Group on Earth Observations
GeoJSON	Geographical JavaScript Object Notation
GEOSS	Global Earth Observation System of Systems
GML	Geographical Markup Language
GUF	Geospatial user feedback
GUM	Guide to the Expression of Uncertainty in Measurement
IETF	Internet Engineering Task Force
ISO	International Standards Organization
JSON	JavaScript Object Notation
KVP	Key and Value Pair
MIT	Massachusetts Institute of Technology
NetCDF	(Network Common Data Form
NGO	Non-governmental organization

O&M	Observations and Measurements
OAIS	Open Archival Information System
OGC	Open Geospatial Consortium
OMG	Object Management Group
OPeNDAP	Open-source Project for a Network Data Access Protocol
Orcid	Open Researcher and Contributor ID
PROV	Provenance
Pub/Sub	Publish–subscribe
RF	Royalty-Free
RSS	Really Simple Syndication
SAML 2.0	Security Assertion Markup Language 2.0
SensorML	Sensor Model Language
SOS	Sensor Observation Service
SQL	Structured Query Language
SWE	Sensor Web Enablement
SWE4CS	Sensor Web Enablement for Citizen Science
TC	Technical Committee
UML	Unified Modeling Language
UncertML	Uncertainty Markup Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USB	Universal Serial Bus
W3C	World Wide Web Consortium
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service
WMTS	Web Map Tile Service
WPS	Web Processing Service
XML	Extensible Markup Language

Executive Summary

The most recognized standardization bodies producing standards used by the community are Open Geospatial Consortium (OGC), ISO TC 211, W3C and IETF.

APIs don't provide interoperability between server systems, in the sense standards do.

Standards available for each principle in the ten GEOSS DMP are enumerated:

DMP-1: Discovery - OpenSearch provides a simple query language to query a search engine by free text. CSW is the main standard for catalogues.

DMP-2: Online Access - We have the classical OGC web services family (WFS, WCS, WMS, WMTS). SOS is the appropriate service to use. WPS is a way to expose a geospatial analytical processing tool on the web. OPeNDAP is also a popular convention.

DMP-3: Data Encoding - ISO 19109 provides what is called the General Feature Model. GeoJSON is another standard to encode features in JSON. Some observations can be also well represented in NetCDF. O&M allows fully describing sensor model. OGC has proposed a profile of O&M for CS (SWE4CS) that is ready to be used by CS projects. TimeSeriesML is a recent proposal to use time series in an easy manner.

DMP-4: Data Documentation - ISO19115 is the metadata standard that everybody in the geospatial world is using. SensorML is a standard to describe the sensor used in a set of measurements.

DMP-5: Data Traceability - ISO19115 provides a data model and XML encoding for lineage information. W3C PROV is the W3C to document provenance of web resources.

DMP-6: Data Quality-Control - The geospatial data quality is described in the ISO19157**.**

DMP-7: Data Preservation - Not much has been done to ensure preservation. The common practice is to transfer it to an archive. This practice is described in the OAIS.

DMP-8: Data and Metadata Verification - Verification of integrity and authenticity is an aspect covered by OAIS (ISO 14721) and included in ISO 19165.

DMP-9: Data Review and Reprocessing - Not aware of standards directly designed to do this. The use of WPS and provenance standards can help.

DMP-10: Persistent and Resolvable Identifiers - DOI when data is stored in open repositories. Orcid assigns people an identifier. OpenID and SAML 2.0 provide standard ways to distributed authentication.

The content of this document will be complemented by the future deliverable D4.4 were some of the concerns and recommendations expressed in this document will be addressed by new Standards, Best Practices and Engineering Reports for Citizen Science.

1 Introduction

The Open Geospatial Consortium (OGC) Citizen Science Domain Working Group (DWG) is motivated to support citizen science by providing a forum for increasing understanding and demonstration of the benefits brought by the use of open standards and best practices. This new activity has been already promoted by EC funded projects and approved by the OGC Technical Committee (TC). In WeObserve, **Task 4.2 Support standards, data management and interoperability via the OGC citizen science domain working group** is supporting the DWG by chairing the group and by building bridges with the Working Group on Data, Tools and Technology within ECSA (and other equivalent activities in other continents).

Within that Task, the current deliverable is the result of the activities conducted to identify and document relevant standards for citizen science (e.g. SWE4CS, Geospatial User Feedback), including also standards coming from other bodies (e.g. ISO) or community driven (e.g. GeoJSON), European and national regulatory standards (e.g. INSPIRE implementation guidelines), and domain-specific standards (e.g. Darwin CORE), including an identification of the gaps in standards available and the improvements needed.

When the need is detected, the development of OGC-specific profile standards, vocabularies and best practices is proposed that are useful for improving harmonization of data acquisition; data discovery, access, documentation, traceability and quality control (GEOSS DMP1, 2, 3, 4, 5, 6) in collaboration with the OGC Data Quality Working Group, and for citizen science data identification, validation, re-processing, curation and preservation (GEOSS DMP 7, 8, 9, 10), in collaboration with the OGC Data Preservation Working Group. The development of the OGC-specific standards, profiles and best practices should be supported by contributing to create consensus among the Standard Working group and the OGC membership in general.

A compendium of new best practices, standards and profiles for citizen-based data and COs will be developed in Deliverable **D4.4 Compendium of new best practices, standards and profiles** to be delivered in Month 28.

2 List of standards bodies

Many organisations are producing standards that are used by the community. The most recognized standardization bodies in the field are Open Geospatial Consortium (OGC), ISO TC 211, W3C and IETF.

2.1 Open Geospatial Consortium (OGC)

The Open Geospatial Consortium (OGC) is an international not for profit organization committed to making quality open standards for the global geospatial community. These standards are made through a consensus process and are freely available for anyone.

The OGC provides a consensus process that communities of interest use to solve problems related to the creation, communication and use of spatial information. These communities sort roughly into these ten domains: Aviation, Built Environment & 3D, Business Intelligence, Defence & Intelligence (D&I), Emergency Response & Disaster Management, Energy & Utilities, Geosciences & Environment, Government & Spatial Data Infrastructure, Mobile Internet & Location Services, Sensor Webs, University and Research.

The OGC facilitates dialog within and between these domains leading to testbeds, pilot projects and interoperability experiments that deliver candidate interface and encoding specifications and best practices. These are vetted in the OGC Standards Program. After thorough review, testing, public comment and refinement, they usually pass a vote to become adopted OGC standards packaged with associated compliance tests.

The OGC's continual growth in membership since 1994 is evidence of the value members discover in belong to this unique, collegial and efficient networking and standards-producing organization. Their 525+ member organizations come from across government, commercial organizations, NGOs, academic, and research institutes.

The most used OGC standard is the Web Map Service WMS. The OGC also maintains a family of standards to describe sensors and to retrieve sensor information called Sensor Web Enablement that can be used to retrieve Citizen Science Observations. Another popular OGC standard is the Web Feature Service that can be useful when we assimilate each Citizen Science observation to a point, a line or a polygon. To get a panorama of the OGC standards, refer to the OGC reference model (https://www.opengeospatial.org/standards/orm)

2.2 ISO TC 211

ISO/TC 211 is the Geographic information/Geomatics division of the International Organization for Standardization (ISO) and it addresses standardization in the field of digital geographic information. Its work aims to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth.

These standards may specify, for geographic information, methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital / electronic form between different users, systems and locations. The work shall link to appropriate standards for information technology and data where possible, and provide a framework for the development of sector-specific applications using geographic data.

Standards are developed by the people who need them. Technical committees include experts from both standards and industry and these experts are put forward by ISO's national members. ISO has put together groups of experts that represent every sector imaginable. In fact there are more than 250 technical committees. ISO members can choose whether they want to be part of a particular TC, and their level of involvement. In most cases, the experts that develop ISO Standards work in the field. They may have expert knowledge but they're not isolated professors of theory. They understand and anticipate the challenges of their sector, using standardization as a tool to create a level playing field that benefits everyone.

The most used ISO/TC 211 standard is the ISO 19115:2003 Geographic information – Metadata that can be used to describe a collection of Citizen Science observations grouped as a dataset. Metadata is the bases for the data catalogues that allow for discovering geospatial datasets. Some OGC standards has been pushed into the ISO process (including OGC O&M, WMS, etc) to make it more formal in a way that is acceptable by some governmental bodies that are more favourable to accept standards ratified by OGC. Most of the ISO TC211 are not free and open. To get a panorama of the ISO TC211 standards please check the ISO TC 211 standards guide (removed from the original URL: https://www.isotc211.org/Outreach/ISO_TC%20_211_Standards_Guide.pdf but still available in https://trac.osgeo.org/geonetwork/raw-attachment/ticket/189/ISO%20TC%20211%20-%20Standards%20Guide.pdf)

2.3 W3C

The World Wide Web Consortium (W3C) is an international community where Member organizations, a full-time staff, and the public work together to develop Web standards. Led by Web inventor and Director Tim Berners-Lee and CEO Jeffrey Jaffe, W3C's mission is to lead the Web to its full potential.

W3C does not have a single physical headquarters. There are four institutions, however, that "host" W3C: MIT (in Cambridge, MA, USA), ERCIM (in Sophia-Antipolis, France), Keio University (near Tokyo, Japan), and Beihang University (in Beijing, China). The W3C staff is distributed around the world. W3C receives funds from W3C Member dues, research grants and other sources of private and public funding and individual donations of money and equipment.

W3C's primary activity is to develop protocols and guidelines that ensure long-term growth for the Web. W3C's standards define key parts of what makes the World Wide Web work. The World Wide Web Consortium achieves its mission by bringing diverse stake-holders together, under a clear and effective consensus-based process to develop high-quality standards based on contributions from the W3C Members, staff, and the community at large.

Some elements of openness that apply to W3C standards include:

- All standards are available publicly at no cost
- W3C adopted a Patent Policy in 2004 with the stated goal of assuring "that Recommendations produced under this policy can be implemented on a Royalty-Free (RF) basis."
- W3C Process requires that groups address public comments
- All technical comments are handled on their merits, whether they are made by W3C Members or public.
- W3C's process is vendor-neutral.
- W3C's persistence policy seeks to ensure that standards will be available at the same URI, unchanged, indefinitely.

W3C does provide a free service for validating Web pages against standard formats. Although validation is not mandatory on the Web, it is useful for improving the quality of pages.

Some of the basic HTTP standards are maintained by the W3C standards including the family on linked data and RDF (e.g. https://www.w3.org/TR/rdf11-mt/). Recently the W3C Spatial Data on the Web Interest Group has edited generic recommendations for spatial data on the web: Spatial Data on the Web Best Practices (https://www.w3.org/TR/sdw-bp/)

2.4 IETF

The Internet Engineering Task Force (IETF) is the premier Internet standards body, developing open standards through open processes. The IETF is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. The technical work of the IETF is done in Working Groups, which are organized by topic into several Areas. Participation in the IETF is open to individuals willing to contribute technical expertise.

The mission of the IETF is to make the Internet work better by producing high quality, relevant technical documents that influence the way people design, use, and manage the Internet. In outline, the process of creating an Internet Standard is the following: a specification undergoes a period of development and several iterations of review by the Internet community and revision based upon experience, is adopted as a Standard by the appropriate body and is published.

The goal of technical competence, the requirement for prior implementation and testing, and the need to allow all interested parties to comment all require significant time and effort. On the other hand, today's rapid development of networking technology demands timely development of standards. The Internet Standards Process is intended to balance these conflicting goals. The process is believed to be as short and simple as possible without sacrificing technical excellence, thorough testing before adoption of a standard, or openness and fairness.

Some of the basic HTTP standards are maintained by IETF. IERF is also responsible for hosting the GeoJSON standard that can be used to describe the Citizen Science geospatial data fragment.

3 The difference between an API and the standards

Recently, there has been some confusion between APIs and standards. In the old days, companies opted for closed systems with no documented formats or interfaces. Recently many vendors release some tailored APIs (many times paired to tailored JSON formats). The Google maps and the Twitter APIs are two well-known examples. This document recognizes that a Citizen Science project that publishes the API endpoint and the API documentation has done an important step to openness but this has nothing to do with interoperability. Open APIs allow others to build clients on top of some systems; but this position assumes that the others need to adapt their products to out personal view. In most cases, this reflects a dominant position in the market by providing a "single" web service that interacts with many clients, all of them technological lock-in to the server vendor. This approach does NOT provide interoperability between server systems, in the sense that clients and services from different vendors can communicate and be replaced if needed. It also does not reflect any aim to create the necessary consensus in the community to develop standards that many vendors can openly adopt and implement.

Nevertheless, we have to learn from the market that the approach based on web APIs is more dominant than the perspective of web services. The web APIs are an effort to adopt HTTP as the main protocol of the web and use it respecting the spirit of the original web that Roy Fielding formalized in his PhD dissertation "Architectural Styles and the Design of Network-based Software Architectures" at UC Irvine. Recently OpenAPI (formerly known as Swagger) has gained popularity as a way to document and present an API. Recently, the OGC has started a modernization of his web services to adapt them to the OpenAPI paradigm. In the next OGC API Hackathon (https://www.opengeospatial.org/OGCAPI_Hack2019) the OGC will experiment with lay the foundations of the next generation of the OGC services for features, coverages, maps, tiles and processes. He new generation of OGC service will provide requirements for services that will mandate dual discovery mechanism based on landing pages and OpenAPI documents. If the approach gets traction other services including the Sensor Observation Service could also adopt the same approach and influence the standards selected for Citizens Science data.

4 Standardization targets

Citizen science activities are composed by several components and actors. When discussing about standards it is important to have a clear idea on which is the standardization target. This section enumerates the possible ones and identifies the standards that are relevant for each target.

Citizen science projects: There are so many projects these days that it could be good to have an inventory of projects to be able to discover them. Essentially we need a data model to collect the necessary information about them (topic, responsible party, URL to the app, URL to the collected data, etc.). Associations of Citizen Science (e.g. ECSA) might want to exchange information between them in an interoperable way. These standard is out of scope of this document.

Citizen science client applications: The proliferation of applications for mobile devices to facilitate the task of data collectors for each project ends in a myriad of them. It could be good to have standard interfaces to capture data from different projects. iNaturalist is one crosscutting application. These standards are out of scope of this document.

Citizen science variables: The objective of this target is to collect data in a model that is interoperable with other systems (such as remote sensing of Earth observation in-situ research infrastructures) to ensure that the data collected will be “compatible” with other sources and potentially conflated with them. The concept of Essential Variables can help in this direction. Data capturing standards are published by CEOS; they are intended to specify how to capture in-situ data that can be useful for Remote Sensing calibration. These standards and methodologies are out of scope of this document.

Citizen science sensor communication interfaces: Many citizen science projects use cheap or DIY sensors that need to communicate with other devices to store or transmit the data. Standards related to this target could deal with USB, WiFi, 3G and other interfaces or radio communication infrastructures. These standards are out of scope of this document.

Citizen science collected data: This is subjected to the aspects considered in the Data Management Principles in GEOSS and all the standards related to them.

5 Standardization for data collected in CS project

To produce a coherent list of standards it is important to identify a classification criterion. A classical approach classifies the standards in web services, data encodings and query and filter languages. In this list we prefer to follow a classification that follows the GEOSS ten Data Management Principles (DMP) produced by GEO that can be found here: https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf

Taking this route, we favour a better interoperability with GEOSS if the listed standards are implemented by the projects.

For each principle in the ten DMP, this section will enumerate the standards available and will provide a very short justification of its usability in this context.

5.1.1 DMP-1: Discovery

Data and all associated metadata will be discoverable, through catalogues and search engines, and data access and use conditions, including licenses, will be clearly indicated.

Note for the reader. This topic is fundamentally linked with Data documentation and we recommend that you read DMP-4 in preparation for reading this one.

Discovery of information is achieved by search engines. They follow two main approaches: Metadata catalogues and Information crawlers. In metadata catalogues, metadata about resources is registered by catalogues that will index the metadata and allow for querying it. In information crawlers, text available on Internet is automatically read and indexed for direct search. Unfortunately, data is generally formed by sequences of numbers, dates or categories that cannot be directly interpreted without a description of them. This is commonly provided in form of metadata tags even if other alternatives could be possible. To standardize discovery we should deal with query languages and output formats. Google has provided detailed guidance on structure markup of dataset description pages.

OpenSearch: This standard provides a simple query language to query a search engine by free text. It provides a Key and Value Pair (KVP) syntax (An example of KVP is: www.google.com?q=Ground+Truth+2.0. For more details: https://en.wikipedia.org/wiki/Query_string) that modern web search engines support and favours a response in an Atom feed. There is an extension of it called *OpenSearch Geo*(http://www.opengeospatial.org/standards/opensearchgeo) that includes geospatial and temporal queries.

Metadata catalogues: The **Catalogue Service For the Web (CSW)** (http://www.opengeospatial.org/standards/cat) is the main standard for catalogues. There are two profiles, one for ISO 19115 records and another for ebRIM implementations that supports a more flexible model based on objects and relations.

Most of the CS projects act as sensors, collecting data. These sensors are not commonly described as ISO19115 datasets but as Observations and Measurements (O&M). Current implementations of the translations of sensor descriptions into ISO 19115 doesn't work well (as experimented in the ConnectinGEO project).

Another topic to consider is that sensors can be able to register themselves without human intervention. The equivalent could be CS activities that register themselves in catalogues. Even if there are solutions in the mass market arena, we are not aware of any standards to allow this.

Finally it's worth to mention GEO-DAB (Discovery and Access Broker) that, in the end, to be discoverable in GEOSS, the CS datasets need to be registered in the DAB. Register them one by one seems impractical so it could be more beneficial that we figure out a protocol where there is a central catalogue where all the relevant CS initiatives register and this catalogue is regularly harvested by the GEO-DAB.

5.1.2 DMP-2: Online Access

Data will be accessible via online services, including, at a minimum, direct download but preferably user-customizable services for access, visualization and analysis.

For this we have the classical OGC web services family.

- **Web Feature Service (WFS)** (http://www.opengeospatial.org/standards/wfs) can be used to serve feature data (point, line or polygon based). The output format is usually GML.
- **Web Coverage Service (WCS)** (http://www.opengeospatial.org/standards/wcs) can be used to serve coverages (imagery)
- **Web Map Service (WMS)** (http://www.opengeospatial.org/standards/wms) can be used to server map renderization of data.
- **Web Map Tile Service (WMTS)** (http://www.opengeospatial.org/standards/wmts) can be used to server map renderization of data in a more efficient way.

The natural way to proceed is to consider the CS Observatories as sensors. Then the **Sensor Observation Service (SOS)** (http://www.opengeospatial.org/standards/sos) is the appropriate service to use. It better handles long time series of data capture in different stations that regularly measure one or more parameters. These parameters can be numerical values but also more complex data formats like pictures and videos. SOS is based on O&M that is specified in XML. Recently some vendors have stated to make implementations based on JSON encodings that are fasters and easier to use in web browsers. The proliferation of this new encodings will favour the Citizen Science implementations.

To subscribe to a service and be notified when something has been observed that is significant for us and receive an alert, the recently published **Pub/Sub** (http://www.opengeospatial.org/standards/pubsub) standard can be very useful.

Please note that data access might require also to consider security and licensing issues. We have learnt that standard licensing is one of the topics that LandSense project wants to capture.

Finally **Web Processing Service (WPS)** (http://www.opengeospatial.org/standards/wps) is a way to expose a geospatial analytical processing tool on the web.

OPeNDAP is also a popular convention for online access to data.

5.1.3 DMP-3: Data Encoding

Data should be structured using encodings that are widely accepted in the target user community and aligned with organizational needs and observing methods, with preference given to non-proprietary international standards.

Several standards deal with data modelling that could be useful in CS.

The ISO 19109 provides what is called the General Feature Model. This is an abstract specification (not implementable directly) defining the concept of Feature and Feature Type.

An implementation of it is the **Geographical Markup Language (GML)** (http://www.opengeospatial.org/standards/gml) that allows for describing points, lines, polygons and more complex features.

Recently **GeoJSON** (http://geojson.org/) is another standard to encode features in JSON that is now an IETF standard (https://tools.ietf.org/html/rfc7946).

Some observations can be also well represented in **NetCDF**. Initially used for the climate community is being used more broadly and has been brought to the OGC process.

The natural way to proceed is to consider the CS Observatories as sensors. **Observations and Measurements (O&M)** (http://www.opengeospatial.org/standards/om) allows fully describing sensor model that could be used for CS (SOS, O&M and SensorML for a family of standards called Sensor Web Enablement (SWE)). SWE is not a standard in itself. However, there is a SWE Common standard. Actually, OGC has proposed a profile of O&M for CS (**SWE4CS**) (https://github.com/opengeospatial/swe4citizenscience) in the COBWeb project that is ready to be used by other CS projects. To be able to do this the CS activity needs to map their Earth observation activities into the O&M concepts. O&M was specified in XML but recently some vendors have stated to make implementations based on JSON encodings that are fasters and easier to use in web browsers. The proliferation of this new encodings will favour the Citizen Science implementations.

One of the advantages of using O&M is the capacity of using time series in an easy manner and **TimeSeriesML** (http://www.opengeospatial.org/standards/tsml) is a recent proposal to do that (it was extracted from the WaterML standard).

There are many other data standards that could eventually be used for more specific purposes, such as raster data formats (e.g. GeoTIFF) or the recently proposed **GeoPackage** (http://www.opengeospatial.org/standards/geopackage) (a Geospatial extension of MySQL) or the Geodatabase formats (some of them using **Simple Features for SQL** (http://www.opengeospatial.org/standards/sfs) as a query language).

5.1.4 DMP-4: Data Documentation

Data will be comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata based on international or community-approved standards. To the possible extent, data will be described in peer-reviewed publications referenced in metadata records.

Data documentation is what makes the data catalogues work. In essence it requires a metadata of some sort.

ISO19115: The metadata standard that everybody in the geospatial world is using. Even it is the **core metadata** it is limited to ~20 entries, the full standard specifies more than a hundred entries. NB: the concept of "Core Metadata" was removed from 19115-1:2014.

Sensor description standards: **Sensor Model Language (SensorML)** (http://www.opengeospatial.org/standards/sensorml) is a standard to describe the sensor used in a set of measurements. Can be used to describe a DIY sensor or a measurement done by a human sensor. **O&M** includes the semantic description of the measurements that can be considered metadata about the meaning of the measurements.

Apart from the more common metadata there are other annotation standards that we can consider. They provide a more light and flexible schema for metadata.

- **Schema.org** (http://schema.org): This W3C standard to annotate web content that is used by web crawlers to better interpret and index the web content.
- **RSS** (http://cyber.harvard.edu/rss/rss.html) and **Atom** (https://tools.ietf.org/html/rfc4287) feeds: These standards were initially designed to publicize news but are now used to notify additions and changes in web resources. They consist in a dynamic xml file that is regularly updated by the server and subscribed by client applications. See (OGC OpenSearch Geo and Time Extensions).

- The **web annotation model** (http://www.w3.org/TR/annotation-model/ (not approved yet.)): This W3C standard offers a flexible model to annotate web resources by creating links between bodies (the note) and targets (the resource annotated).
- **Geospatial user feedback (GUF)** (http://www.opengeospatial.org/standards/guf): This standard provides a data model to document user feedback for geospatial objects in XML.

5.1.5 DMP-5: Data Traceability

Data will include provenance metadata indicating the origin and processing history of raw observations and derived products, to ensure full traceability of the product chain.

Traceability is achieved by documenting details about the process done to a resource, also mentioning data sources used and actors involved in the processing.

The **ISO19115** (extended in the **ISO19115-2**) provides a data model and XML encoding for lineage information.

The **W3C PROV** (http://www.w3.org/TR/prov-overview/) is the W3C to document provenance of web resources.

It is worth noticing that by mentioning the actors involved in the data collection of individual observations, we can incur into privacy issues and personal data protection issues that needs to be considered.

It is worth noticing that the **Business Process Modelling Language (BPML)** is an OMG (Object Management Group) standard (formerly known as BPEL) to document processing chains, based on **Business Process Model and Notation (BPMN)** (http://www.bpmn.org) (that is an extension UML activity diagrams).

5.1.6 DMP-6: Data Quality-Control

Data will be quality-controlled and the results of quality control shall be indicated in metadata; data made available in advance of quality control will be flagged in metadata as unchecked.

The geospatial data quality is described in the **ISO19157** (formerly ISO19138) that provides a data model for providing quantitative and conformance quality and also a vocabulary of quality measurements. An important component of data quality is the uncertainty defined in the **Guide to the Expression of Uncertainty in Measurement (GUM)** (http://www.bipm.org/en/publications/guides/gum.html) and the **Uncertainty Markup Language (UncertML)** (http://www.uncertml.org/), that is extended in the QualityML (A vocabulary and an encoding, broader than the proposed in ISO19157, were developed during the EC FP7 GeoViQua project and recently updated in the OGC testbed 12. The results of the update have been published as an OGC Public Engineering Report (http://www.opengeospatial.org/). It is however not an OGC standard but providing a list of quality statistics and a way to encode them.

5.1.7 DMP-7: Data Preservation

Data will be protected from loss and preserved for future use; preservation planning will be for the long term and include guidelines for loss prevention, retention schedules, and disposal or transfer procedures.

Not much has been done to ensure preservation of the CS Observatories data when the project is no longer able to maintain. The common practice in the geoinformation is to transfer it to an archive. This practice is described in the **Open Archival Information System OAIS** (also known as ISO 14721). The particularities of the geospatial information are being captured in the draft candidate **ISO 19165** data and metadata preservation (based in both OAIS and ISO 19115).

5.1.8 DMP-8: Data and Metadata Verification

Data and associated metadata held in data management systems will be periodically verified to ensure integrity, authenticity and readability.

Verification of integrity and authenticity is an aspect covered by **Open Archival Information System OAIS** (ISO 14721) and included in **ISO 19165**. A strategy that can help is the application of a packaging format such as ISO 29500-2 Open Packaging Convention (A recent paper has been published in support of this concept: X. Pons, J. Masó (2016) A comprehensive open package format for preservation and distribution of geospatial data and metadata. Computers & Geosciences 97, 89‑97), an alternative of a GeoPackage specifically designed for preservation purposes. While GeoPackage requires to transform data into database tables, Open Packaging Convention does not require format change.

5.1.9 DMP-9: Data Review and Reprocessing

Data will be managed to perform corrections and updates in accordance with reviews, and to enable reprocessing as appropriate; where applicable this shall follow established and agreed procedures.

The authors of this document are not aware of standards directly designed to do this. In any case, the use of **Web Processing Service (WPS)** and provenance standards can help on having processing facilities ready and knowing how the previous version was created respectively.

5.1.10 DMP-10: Persistent and Resolvable Identifiers

Data will be assigned appropriate persistent, unique and resolvable identifiers to enable documents to cite the data on which they are based and to enable data providers to receive acknowledgement for use of their data.

In CS, there are two different types of unique identifiers: user id and record id. It is important that users are able to identify their own records and being able to request their removal. Several attempts have been done to standardize data identifiers (http://libguides.lib.msu.edu/citedata) with not much success. It seems that there is some consensus on assigning **Digital Object Identifiers (DOI)** (https://www.doi.org/) when data is stored in open repositories such as Pangaea (https://www.pangaea.de/) and Zenodo (https://zenodo.org/).

Equivalent initiatives exist to register people and assign them an identifier such as **Orcid** (https://orcid.org/). On the other side, standards like **OpenID** (http://openid.net/) and **SAML 2.0** (http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-tech-overview-2.0.html) provide standard ways to distributed authentication. As an example Google and Facebook id's can be used by third parties to authenticate user id's with OpenID avoiding the need to register in the third party website to get access and be identified.

6 Discussion on gaps and improvements

In the discovery of projects and datasets produced by them we have two complementary initiatives. On one hand we have the ISO 19115 metadata standard that was mainly designed for datasets (that can be assimilated to a Citizen Science campaign) but can be easy adapted to products (that can be assimilated to citizen science collections of sequential or parallel campaigns) or to collections of products (that can be assimilated to a Citizen Science project results). On the other hand, we have the PPSR_CORE Metadata Standard that was originally focussed in describing Citizen Science projects (https://www.citizenscience.org/2015/10/09/ppsr_core-metadata-standard/ or https://www.wilsoncenter.org/article/ppsr-core-metadata-standards) but is being extended by a mainly European team into a second version that also covers more detailed aspects (https://basecamp.com/2071195/projects/13342949/messages/78875086?enlarge=348066038#attachments_for_comment_648590664).

Recommendation 1: There is a need to experiment with combining both approaches avoiding unnecessary duplications of overlapping data structures such us the spatial and temporal extent of a dataset in ISO19115 and a Citizens Science project in the PRSR.

ISO 19115-1 provides a way to describe the data model used in a dataset, the General Feature Model has another and the Observations an Measurements proposes yet another alternative. There is a need to clarify which one is appropriate in each case and how to translate from one to the other.

Recommendation 2: There is need to clarify how to elaborate citizen science data models and how encode them in a way that the can be offered using the O&M model (in a SOS service) and the General Feature Model (in a WFS service).

The SWE relays on the SensorML to describe the sensor that is capturing observations. In Citizen Science, many times sensors (in particular low cost or DIY) are used and the data quality and trustworthy of the observations deeply relay on it. Some other time the project uses the human perception as a sensor (e.g. odour or noise perception). No much efford has been done so far on illustrating how SensorML or an alternative can be used to describe the sensors used in Citizen Science.

Recommendation 3: There is need to clarify how to use SensorML or an alternative when applied to in Citizen Science sensors

Citizen Science projects are more volatile compared with other data capturing initiatives conducted by the government or the scientific research infrastructure. New Citizen Science emerge, others conduct punctual campaigns in some critical time spans while others disappear. There are other examples that are more persistent (e.g. Open Street Map or eNaturalist). This dynamic nature of the Citizen Science projects also affects the ability to discover their datasets and to connect them to the bigger infrastructures in general and to GEOSS in particular that require some time to accept a registration request. SciStarter offers a solution for registering and discovering Citizen Science projects and their datasets but these initiatives are not directly connected to research infrastructures. In contrast, GBIF can be considered a biodiversity research infrastructure that regularly adopts filtered Citizen Science data.

Recommendation 4: There is need to facilitate the discovery of mature citizen science projects that can be of use for research infrastructures as well as to GEO based on the adoption of international standards as part of the technical solution.

In that sense, it's worth to mention GEO-DAB (Discovery and Access Broker) that, in the end, to be discoverable in GEOSS, the CS datasets need to be registered in the DAB. Register them one by one seems impractical so it could be more beneficial that we figure out a protocol where there is a central catalogue where all the relevant CS initiatives register and this catalogue is regularly harvested by the GEO-DAB.

It is worth noticing that by mentioning the actors involved in the data collection of individual observations, we can incur into privacy issues and personal data protection issues that needs to be considered. On the other hand, knowing the reliability of the users providing data is one of the techniques used to improve data quality in Citizen Science projects where data can come from many sources. In addition, some uses will be willing to give up privacy in exchange for recognition that can stimulate participation.

Recommendation 5: There is a risk that a too restrictive interpretation of the GDPR European lay can affect that data quality assessment process as well as some of the common strategies for recognition and rewards. Standards for Single Sign On and their interpretation need to allow for both privacy and public recognition of public participation.

Not much has been done to ensure preservation of the CS Observatories data. An exception is the Joint Research Center (JRC) to catalogue present and part projects. Standards currently under development or deployment for geospatial data such a **Open Archival Information System OAIS** (ISO 14721) and **ISO 19165** can help. The application of a packaging format such as ISO 29500-2 Open Packaging Convention as well as the OGC GeoPackage could be tested.

Recommendation 6: Standards for data and metadata preservation such as ISO 19165, OAIS etc need to be tested in Citizen Science project to determine if the generation of Information Packages is the right solution for preserving their results for future generations.

In addition to that, the authors of this document are not aware of standards directly designed to deal with data review and reprocessing. In any case, the use of **Web Processing Service (WPS)** and provenance standards can help on having processing facilities ready and knowing how the previous version was created respectively.

Data preservation is also linked to the need to provide data identifiers for datasets. Some data repositories such as Zenodo provide ways add data identifiers to some static datasets. Some data repositories assign data identifiers to any data query that can be latter reproduced (https://www.gbif.org/en/document/80575/a-beginners-guide-to-persistent-identifiers).

Recommendation 7: Standards for data identification could help to disseminate and preserve the data. More effort need to be done on creating best practices on how to use persistent identifiers for citizen science data.

Table 1. Summary of gaps and improvements.

DMP	Description
DMP-1: Discovery	Translations of sensor descriptions (O&M) into ISO 19115 doesn't work well. Sensors to be able to register themselves Protocol with a central catalogue to register to GEO-DAB
DMP-5: Data Traceability	Privacy issues and personal data protection issues
DMP-7: Data Preservation	Is it enough with ISO 19165?
DMP-8: Data and Metadata Verification	Application of a packaging format such as ISO 29500-2 Open Packaging Convention
DMP-9: Data Review and Reprocessing	No specific standards available
DMP10: Persistent and Resolvable Identifiers	Need for a consensus solution

7 Conclusion and Summary

Most of the Data Management Principles are reasonably well covered by Citizen Science or other similar disciplines standards. This is the case of Online Access, Encoding, Documentation and Data Quality. The following table provides a generic list of the useful standards detected.

There are still some things to improve in terms of usability, applicability and simplification of the use of standards that has been discussed throughout this document.

Table 2. Summary of the recommended standards.

DMP	Description
DMP-1: Discovery	OpenSearch
	Catalogue Service For the Web (CSW)
DMP-2: Online Access	Web Feature Service (WFS)
	Web Coverage Service (WCS)
	Web Map Service (WMS)
	Web Map Tile Service (WMTS)
	Sensor Observation Service (SOS)
	Pub/Sub
	Web Processing Service (WPS)
	OPeNDAP
DMP-3: Data Encoding	Geographical Markup Language (GML)
	GeoJSON
	NetCDF
	Observations and Measurements (O&M)
	SWE4CS
	TimeSeriesML
DMP-4: Data Documentation	ISO19115
	Sensor Model Language (SensorML)
DMP-5: Data Traceability	ISO19115-2
	W3C PROV
	Business Process Modelling Language (BPML)
DMP-6: Data Quality-Control	ISO19157
	Uncertainty Markup Language (UncertML)

DMP-7: Data Preservation	Open Archival Information System OAIS ISO 19165
DMP-8: Data and Metadata Verification	Open Archival Information System OAIS ISO 19165
DMP-9: Data Review and Reprocessing	Web Processing Service (WPS)
DMP10: Persistent and Resolvable Identifiers	Digital Object Identifiers (DOI) Orcid OpenID SAML 2.0

8 References

1. OGC Reference Model. Open Geospatial Consortium Inc. Date: 2011-12-19. Reference number: OGC 08-062r7. Version: 2.1. Available in HTML, PDF and DOC formats: https://www.opengeospatial.org/standards/orm
2. Standards Guide: ISO/TC 211 Geographic Information /Geomatics. 1 June 2019.
3. Spatial Data on the Web Best Practices. W3C Working Group Note 28 September 2017: https://www.w3.org/TR/sdw-bp/
4. Roy Thomas Fielding. Architectural styles and the design of network-based software architectures. Doctoral Dissertation. University of California, Irvine ©2000. ISBN:0-599-87118-0. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf
5. GEOSS Data Management Principles. Prepared by the GEO Data Management Principles Task Force. Approved by Data Management Principles Task Force on April 28, 2015. https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf
6. X. Pons, J. Masó (2016) A comprehensive open package format for preservation and distribution of geospatial data and metadata. Computers & Geosciences 97, 89-97.

An Ecosystem of Citizen Observatories for Environmental Monitoring