Standardized Information Models to Optimize Exchange, Reusability and Comparability of Citizen Science Data. A Specialization Approach*

Ingo Simonis

Open Geospatial Consortium, isimonis@opengeospatial.org

Abstract

The number of citizen science projects is constantly growing. Local, national, and international platforms feature new projects almost every month, resulting in an endless number of new observations that are constantly gathered and stored in databases. Often, these data sets are only used for the sampling campaign's objectives, thus leaving a huge potential unused: its reusability in other contexts and its comparability with other data sets. Reusability and comparability require a number of aspects to be fulfilled. This paper describes those aspects and focuses on the citizen science application profile as a standardized information model to ensure syntactic and semantic understanding of citizen science data. Data compliant with this information model can be discovered and accessed through standardized Web interfaces and therefore easily integrated into any data processing environment or compared to any other data set. It is emphasized that the application profile described in this paper is one of two possible solutions that are currently being explored. The second one is briefly addressed and will be documented in detail in future publications.

Keywords: Citizen science, crowd sourcing, standards, information modelling

1. INTRODUCTION

Citizen science, the involvement of volunteers in research, has increased the scale of studies in many fields (Crain, Cooper, and Dickinson, 2014; Dickinson,

DOI: 10.2902/1725-0463.2018.13.art5

Zuckerberg, and Bonter, 2010) and the number of citizen science projects is constantly growing. Citizen science data, though often generated in a very purpose-oriented process, holds an enormous potential if it can be made available to the public for comparison or integration and further processing in other campaigns or research efforts. As an example, air- or noise pollution results sampled in one campaign could provide valuable insights for species sampling campaigns or public health campaigns. In order to exploit this potential, a number of requirements have to be fulfilled. First, the data needs to be made discoverable, meaning that it has to be annotated with sufficient metadata to allow others to find it. Second, the data needs to be made available at interfaces that are accessible to others, ideally in a standardized way to allow efficient access to the data of subsets thereof. Third, the data needs to be comprehensible to others, which includes aspects such as available formats, the semantics used, or descriptions of the sampling process. It needs to be annotated with details on privacy and security settings and applied processing steps, such as quality assurance or anonymization.

This paper briefly discusses the typical citizen science sampling campaign formalization and execution process to allow understanding of the full set of requirements that need to be fulfilled by citizen science data in order to maximize reuse and comparability. It then focuses on the citizen science application profile, an information model that allows the integration of citizen science data in new contexts without risking semantically incorrect use or interpretation. The model itself can be implemented and serialized in various ways, featuring XML or JSON based object oriented models, or Semantic Web approaches based on triples and links. Eventually, it is briefly shown how the citizen science application profile can be integrated into spatial data infrastructures and processing chains. The paper ends with an outlook on future work that follows a slightly different modelling approach.

2. THE CITIZEN SCIENCE PROCESS

The citizen science process commonly consists of five steps. These steps are usually executed sequentially, though they may include loops and feedback cycles:

1. Definition of the sampling campaign itself. The type and nature of the sampling protocol and corresponding observed properties are normally motivated by the survey objectives, but may take additional aspects into consideration such as the quality assurance process or publication requirements. Once defined, the survey app is released and used by the citizens during the data collection process.
2. Observation data is produced by citizen scientists and logically stored in raw format. Raw data quality is dependent on a number of factors, most

importantly, motivation and expertise of the individual citizen scientist. Individual raw data quality is therefore hard to assess.

3. Raw data from all campaign participants is collected and permanently stored. Aggregation in a logically centralized repository enables quality assessment processes by comparing multiple (raw) data.
4. Quality assurance processes work on the raw data and on previously quality-assured data, potentially taking external data sets into account. This multi-loop process may result in any number of data sets of various quality or aggregation levels, or any number of versions of raw, quality-assured, aggregated, or newly derived data sets.
5. Data of various levels of processing, ideally covering multiple well described levels from raw data up to highly processed and quality-assured data, is published.

In the following, the various steps and their possible implications are described using a survey example from the research project Citizen Observatory Web (COBWEB, https://cobwebproject.eu). Occurrences of the invasive species Japanese Knotweed are monitored in many places around the world. Japanese Knotweed, native to East Asia, is a large, herbaceous perennial plant of the Polygonaceae family. It is known for its invasive root system and strong growth that can damage concrete foundations, buildings, flood defences, roads, or sidewalks. It is monitored using a mobile application that helps to capture key parameters of Japanese Knotweed occurrences.

One of the primary goals when defining a sampling campaign is to avoid any ambiguities in the understanding of the properties observed, sensing techniques and hardware used, and sampling protocols between survey designers and citizens. If all the observed properties were using definitions in the form of simple names, the semantics would be limited to the understanding of the survey designer and – given that sufficient descriptive data is provided – to the citizens participating in the survey. If, instead, fully qualified names in the form of resolvable URLs were to be used, then the raw data becomes meaningful even to external users who have not used the mobile app but only received the raw data. Labels can still be used for display purposes. The citizen does not see any links to the definitions directly, as the style sheets render all elements able to fit the screen of a mobile phone, but the links reveal further information if followed by tapping on the labels displayed on the screen.

The mobile applications usually transfer all observations either in burst mode or continuously to a local or cloud-based centralized storage system. Currently, almost all applications make use of unique data models to transfer the observation data, and thus have a strong link between the application and the storage system. In terms of flexibility, it is desirable to remove that strong link.

This requires a standardized information model and serialization format. If applied, the mobile application could be freed from a dedicated storage and processing system, because all the transferred observation data would implement a standardized schema and data transfer would follow a predefined pattern. The user (i.e. citizen scientist) or application developer could choose the storage and processing system of their choice, rather than being required to use or build their own. This approach would yield new market opportunities, where participants could concentrate on either the application or storage and processing system design. It requires the standardization of a citizen science information model and transfer format, together with a communication model that connects mobile applications and data storage systems. Both are described in sections 3 (citizen science application profile) and 4 (integration into spatial data infrastructures and processing chains) below. Section 5 briefly compares the approach documented herein with a second approach that is currently under investigation.

In summary, using observed property identifiers in the form of resolvable URLs ensures unique semantics. The underlying technique does not have to be made visible to the citizen scientist, who only sees the necessary information as rendered by style sheets. Still, in the event of ambiguity, links can be followed to obtain further information. Further on, using a standardized information and communication model lessens the strong link between the mobile applications and storage and processing systems.

3. CITIZEN SCIENCE INFORMATION MODEL

The goal of the citizen science information model is to allow standards-compliant encoding of all types of citizen science observation data. The model should support strong semantics and support all types of observations including raw observations made in the field as well as derived or generated observations as part of the quality assurance or signal analysis processes. The model should be supported by existing Web service interfaces to allow for easy access and discovery.

3.1. Available standards

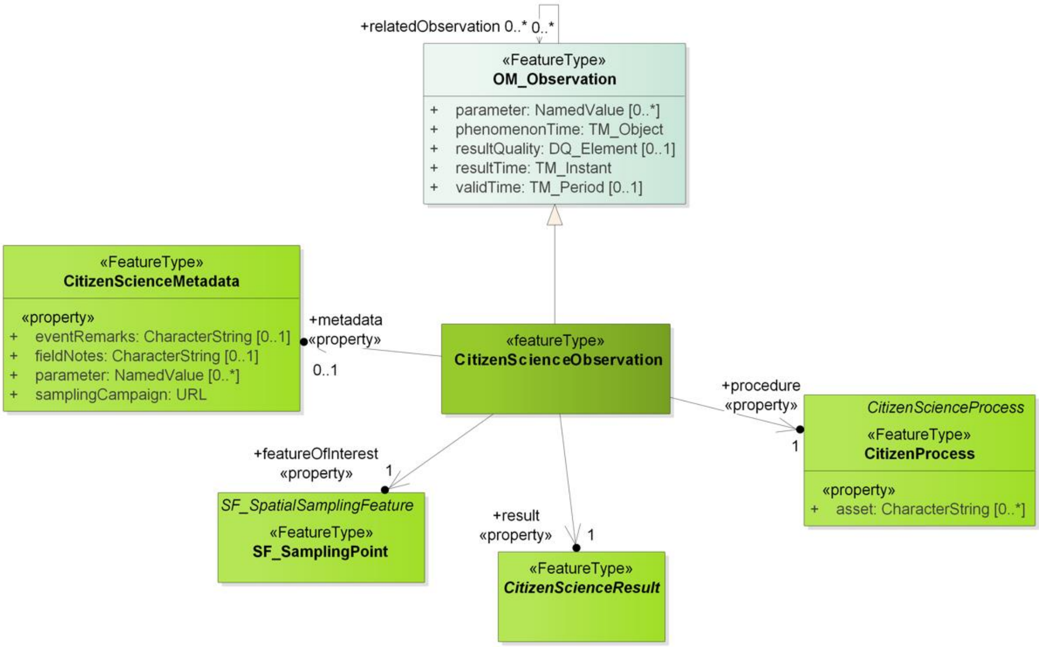
One of the key challenges to efficiently discover and access spatial data on the Web is interoperability. The Open Geospatial Consortium (OGC) has made significant contributions to this effort with a series of standards commonly referred to as SWE, Sensor Web Enablement. Originally designed to allow a network of sensors, the scope has been broadened to include humans as sensors and sensor data processing capacities. There are three information models and a number of Web service specifications, with the Sensor Observation Service being the most important one in the context of citizen science. The information models include:

- Observations and Measurements information model known as O&M. Version 2 of O&M is published in two parts: ISO 19156:2011 *Geographic Information - Observations and Measurements* (ISO, 2011) with an XML serialization published as OGC standard 10-025r1 (Cox, 2011) and a provisional OWL ontology (Cox, 2013). O&M provides a domain-neutral vocabulary and model for an observation and its associated properties.
- Sensor Model Language known as SensorML. Version 2 of SensorML has been published as OGC standard 12-000 (Botts and Robin, 2014). SensorML provides a robust and semantically-tied means of defining processes and processing components associated with the measurement and post-measurement transformation of observations.
- SweCommon, which defines low-level data models for exchanging sensor-related data between nodes of the OGC Sensor Web Enablement (SWE) framework. These models allow applications and/or servers to structure, encode and transmit sensor datasets in a self-describing and semantically enabled way. SweCommon has been published as OGC standard 08-094r1 (Robin, 2011).

3.2. Citizen Science Application Profile

The citizen science application profile is implemented as an application profile using O&M, SensorML, and SweCommon. It specializes the existing O&M observation model with citizen science aspects as illustrated in Figure 1 and introduces the concept of an observation collection to allow the promotion of properties that are shared among all members of an observation collection to the collection level. Application profile classes are shaded in green, imported classes in grey.

Figure 1: Citizen Science Observation Model

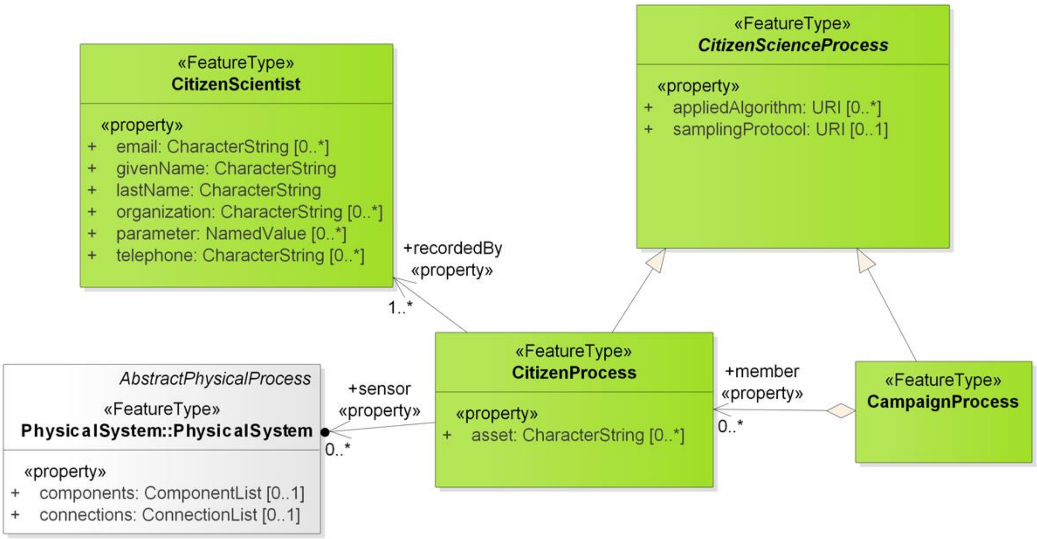


The CitizenScience model is fully documented on GitHub, thus this paper concentrates on the most important aspects and modelling characteristics. For further details, please refer to http://tinyurl.com/o6m5g44. The *CitizenScienceObservation* consists of a number of properties, including:

- all properties directly inherited from *OM_Observation*, including associations to other observations implemented as *relatedObservation* association; this includes the option to define the role of each related observation;
- *CitizenScienceMetadata*, currently loaded with the absolute minimum that was reported as required from the COBWEB field studies, but extendable in future by adding additional *NamedValues*;
- a *featureOfInterest*, implemented in the form of a domain or sampling feature, representing the feature the observation applies to or was sampled at respectively;
- a *procedure* property that allows definition of the citizen, potentially involved sensors or algorithms;
- and the *result*, implemented as SweCommon *DataRecords*, to hold the observation result values.

The *CitizenProcess* is further illustrated in Figure 2. It allows the description of *appliedAlgorithms* used during quality assurance processes. The combination with the *relatedObservation* association provides a provenance model that allows the history of each observation to be understood.

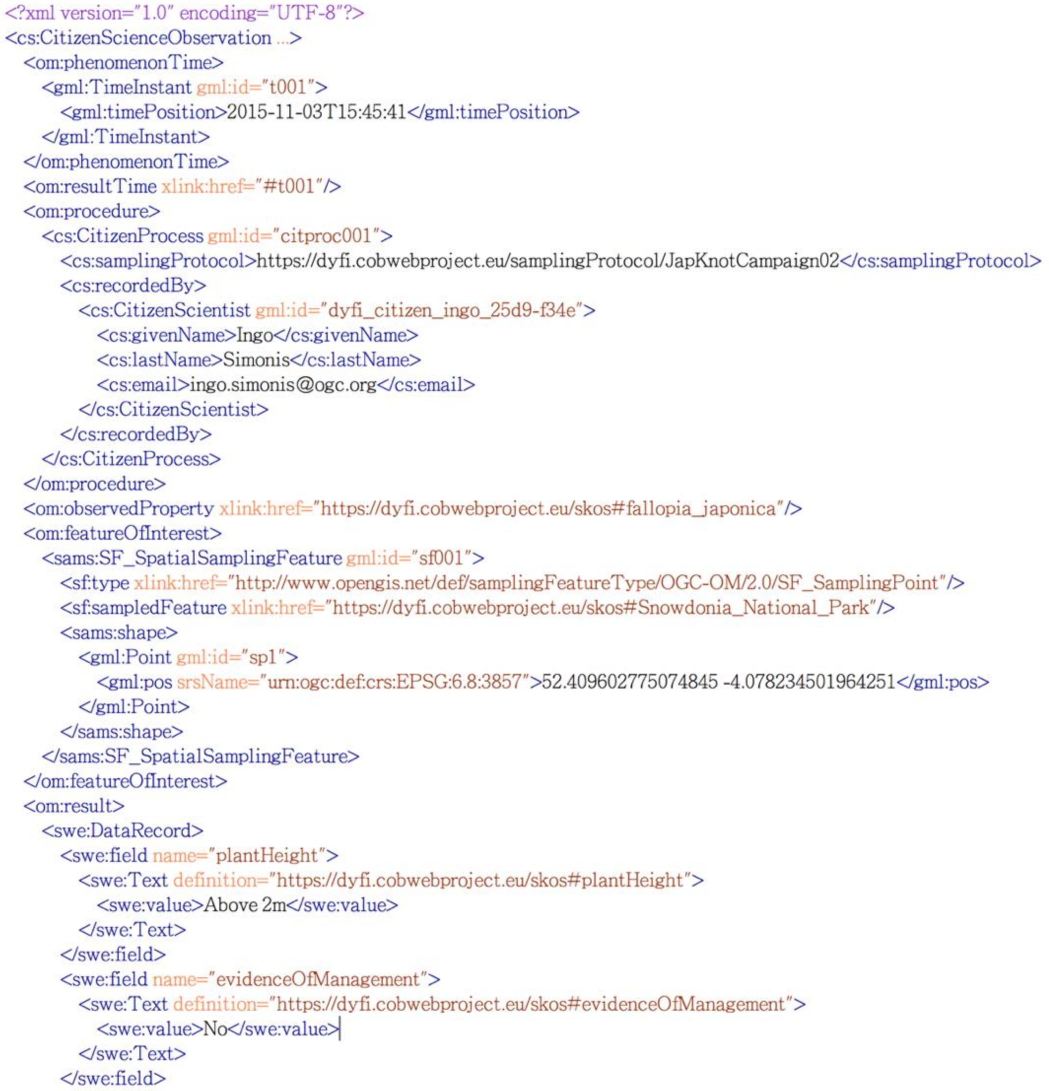
Figure 2: *CitizenScienceProcess* Model


«FeatureType»
CitizenScientist
«property»
+ email: CharacterString [0..*]
+ givenName: CharacterString
+ lastName: CharacterString
+ organization: CharacterString [0..*]
+ parameter: NamedValue [0..*]
+ telephone: CharacterString [0..*]
«FeatureType»
CitizenScienceProcess
«property»
+ appliedAlgorithm: URI [0..*]
+ samplingProtocol: URI [0..1]
+recordedBy
«property»
1..*
AbstractPhysicalProcess
«FeatureType»
PhysicalSystem::PhysicalSystem
«property»
+ components: ComponentList [0..1]
+ connections: ConnectionList [0..1]
+sensor
«property»
0..*
«FeatureType»
CitizenProcess
«property»
+ asset: CharacterString [0..*]
+member
«property»
0..*
«FeatureType»
CampaignProcess

The *CitizenScienceProcess* is further generalized by the *CitizenProcess* for individual observations, and the *CampaignProcess* that allows the aggregation of homogeneous observations to a campaign collection. The *CitizenProcess* allows the addition of *assets* used during the observation process, e.g. colour cards to determine the sea colour, *sensors* such as thermometers or complex sensing devices, and details about the *CitizenScientist*.

The following example shows an excerpt of the citizen science application profile, serialized using XML, based on an XML Schema that is auto-generated from the UML model. The semantics of each observed property are provided in the form of links.

Figure 3: ***CitizenScienceObservation* Example**



```
<?xml version="1.0" encoding="UTF-8"?>
<cs:CitizenScienceObservation ...>
  <om:phenomenonTime>
    <gml:TimeInstant gml:id="t001">
      <gml:timePosition>2015-11-03T15:45:41</gml:timePosition>
    </gml:TimeInstant>
  </om:phenomenonTime>
  <om:resultTime xlink:href="#t001"/>
  <om:procedure>
    <cs:CitizenProcess gml:id="citproc001">
      <cs:samplingProtocol>https://dyfi.cobwebproject.eu/samplingProtocol/JapKnotCampaign02</cs:samplingProtocol>
      <cs:recordedBy>
        <cs:CitizenScientist gml:id="dyfi_citizen_ingo_25d9-f34e">
          <cs:givenName>Ingo</cs:givenName>
          <cs:lastName>Simonis</cs:lastName>
          <cs:email>ingo.simonis@ogc.org</cs:email>
        </cs:CitizenScientist>
      </cs:recordedBy>
    </cs:CitizenProcess>
  </om:procedure>
  <om:observedProperty xlink:href="https://dyfi.cobwebproject.eu/skos#fallopia_japonica"/>
  <om:featureOfInterest>
    <sams:SF_SpatialSamplingFeature gml:id="sf001">
      <sf:type xlink:href="http://www.opengis.net/def/samplingFeatureType/OGC-OM/2.0/SF_SamplingPoint"/>
      <sf:sampledFeature xlink:href="https://dyfi.cobwebproject.eu/skos#Snowdonia_National_Park"/>
      <sams:shape>
        <gml:Point gml:id="sp1">
          <gml:pos srsName="urn:ogc:def:crs:EPSG:6.8:3857">52.409602775074845 -4.078234501964251</gml:pos>
        </gml:Point>
      </sams:shape>
    </sams:SF_SpatialSamplingFeature>
  </om:featureOfInterest>
  <om:result>
    <swe:DataRecord>
      <swe:field name="plantHeight">
        <swe:Text definition="https://dyfi.cobwebproject.eu/skos#plantHeight">
          <swe:value>Above 2m</swe:value>
        </swe:Text>
      </swe:field>
      <swe:field name="evidenceOfManagement">
        <swe:Text definition="https://dyfi.cobwebproject.eu/skos#evidenceOfManagement">
          <swe:value>No</swe:value>
        </swe:Text>
      </swe:field>
```

4. INTEROPERABLE INFRASTRUCTURE

Data encoded according to the citizen science application profile can be made available at a variety of interfaces, including OGC Web service interfaces such as WFS or SOS or RESTful services. Offering the data as part of a service environment such as the OGC Web services has many advantages, for example, that the data becomes part of a seamless service infrastructure. Citizen science data provided by a WFS can be processed using Web Processing Services and stored using another WFS again, while catalogue services such as CSW provide options to describe and discover the data in its various stages of processing and analysis. WFS can further serve as an intermediate layer between geographic linked open data and traditional spatial data infrastructures (Jones et al., 2014) and as a transaction endpoint that receives raw observations from field applications.

5. SPECIALIZATION VS. BEST PRACTICE

The approach documented in this paper is one of two approaches that are currently under investigation. It makes heavy use of specializations of existing standardized information models. The second approach uses standardized information models "as is" and limits the high number of degrees of freedom by either defining constraints or providing examples and best practices. Both approaches have several advantages and disadvantages. The second approach together with a detailed comparison of both will be the subject of future publications.

6. CONCLUSIONS

This paper has provided some insights into interoperability aspects when serving data from citizen science projects. The citizen science application model has been suggested as a standardized way of serving sampling campaign data. The model is compliant with the ISO 19100 series of standards, implementing ISO 19156 as well as SensorML and SweCommon to ensure maximal interoperability with existing systems. As next steps, a security model needs to be added to ensure that all data security, privacy and other legal concerns are met. Then, the model needs to be compared to linked data approaches to further evaluate its usability in an emerging Semantic Web.

ACKNOWLEDGEMENTS

This research and the development of this paper have been supported by the COBWEB project. The project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308513.

REFERENCES

Botts, M., and A. Robin (Co-editors) (2014). "SensorML: Model and XML Encoding Standard. OGC 12-000." https://portal.opengeospatial.org/files/?artifact_id=55939.

Cox, S. (2011). “Observations and Measurements-XML Implementation.” *OGC 10-025r1*. doi:http://www.opengeospatial.org/.

Cox, S. (2013). “An Explicit OWL Representation of ISO/OGC Observations and Measurements.” In *SSN@ ISWC*, 1–18.

Crain, R., C. Cooper, and J. L. Dickinson (2014). Citizen Science: A Tool for Integrating Studies of Human and Natural Systems, *Annual Review of Environment and Resources,* 39 (1). Annual Reviews: 641–65. doi:10.1146/annurev-environ-030713-154609.

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter (2010). Citizen Science as an Ecological Research Tool: Challenges and Benefits, *Annual Review of Ecology, Evolution, and Systematics,* 41: 149–72. http://kbsgk12project.kbs.msu.edu/wp-content/uploads/2011/02/annurev-ecolsys-102209-144636.pdf.

ISO (2011). “ISO 19156: 2011-Geographic Information: Observations and Measurements.”

Jones, J., W. Kuhn, C. Keßler, and S. Scheider (2014). “Connecting a Digital Europe Through Location and Place”, in Huerta, J., S. Schade, and C. Granell (Eds) *Connecting a Digital Europe Through Location and Place*, pp. 341–61. doi:10.1007/978-3-319-03611-3.

Robin, A. (2011). “OGC SWE Common Data Model Encoding Standard.” *OGC 08-094r1*. http://portal.opengeospatial.org/files/?artifact_id=41157.