Standardized Information Models to Optimize Exchange, Reusability and Comparability of Citizen Science Data

Table of Contents

Publication Date: 2016-mm-dd

Approval Date: 2016-mm-dd

Posted Date: 2016-01-01

Reference number of this document: OGC 16-NNNrN

Reference URL for this document: http://www.opengis.net/doc/PER/XXXX-xxx

Category: Public Engineering Report

Editor: Ingo Simonis, Rob Atkinson

Title: Standardized Information Models to Optimize Exchange, Reusability and Comparability of Citizen Science Data

---

OGC Engineering Report

COPYRIGHT

WARNING

This document is not an OGC Standard. This document is an OGC Public Engineering Report created as a deliverable in an OGC Interoperability Initiative and is not an official position of the OGC membership. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an OGC Standard. Further, any OGC Engineering Report should not be referenced as required or mandatory technology in procurements. However, the discussions in this document could very well lead to the definition of an OGC Standard.

LICENSE AGREEMENT

Permission is hereby granted by the Open Geospatial Consortium, ("Licensor"), free of charge and subject to the terms set forth below, to any person obtaining a copy of this Intellectual Property and any associated documentation, to deal in the Intellectual Property without restriction (except as set forth below), including without limitation the rights to implement, use, copy, modify, merge, publish, distribute, and/or sublicense copies of the Intellectual Property, and to permit persons to whom the Intellectual Property is furnished to do so, provided that all copyright notices on the intellectual property are retained intact and that each person to whom the Intellectual Property is furnished agrees to the terms of this Agreement.

If you modify the Intellectual Property, all copies of the modified Intellectual Property must include, in addition to the above copyright notice, a notice that the Intellectual Property includes modifications that have not been approved or adopted by LICENSOR.

THIS LICENSE IS A COPYRIGHT LICENSE ONLY, AND DOES NOT CONVEY ANY RIGHTS UNDER ANY PATENTS THAT MAY BE IN FORCE ANYWHERE IN THE WORLD. THE INTELLECTUAL PROPERTY IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE DO NOT WARRANT THAT THE FUNCTIONS CONTAINED IN THE INTELLECTUAL PROPERTY WILL MEET YOUR REQUIREMENTS OR THAT THE OPERATION OF THE INTELLECTUAL PROPERTY WILL BE UNINTERRUPTED OR ERROR FREE. ANY USE OF THE INTELLECTUAL PROPERTY SHALL BE MADE ENTIRELY AT THE USER’S OWN RISK. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR ANY CONTRIBUTOR OF INTELLECTUAL PROPERTY RIGHTS TO THE INTELLECTUAL PROPERTY BE LIABLE FOR ANY CLAIM, OR ANY DIRECT, SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM ANY ALLEGED INFRINGEMENT OR ANY LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR UNDER ANY OTHER LEGAL THEORY, ARISING OUT OF OR IN CONNECTION WITH THE IMPLEMENTATION, USE, COMMERCIALIZATION OR PERFORMANCE OF THIS INTELLECTUAL PROPERTY.

This license is effective until terminated. You may terminate it at any time by

destroying the Intellectual Property together with all copies in any form. The license will also terminate if you fail to comply with any term or condition of this Agreement. Except as provided in the following sentence, no such termination of this license shall require the termination of any third party end-user sublicense to the Intellectual Property which is in force as of the date of notice of such termination. In addition, should the Intellectual Property, or the operation of the Intellectual Property, infringe, or in LICENSOR's sole opinion be likely to infringe, any patent, copyright, trademark or other right of a third party, you agree that LICENSOR, in its sole discretion, may terminate this license without any compensation or liability to you, your licensees or any other party. You agree upon termination of any kind to destroy or cause to be destroyed the Intellectual Property together with all copies in any form, whether held by you or by any third party.

Except as contained in this notice, the name of LICENSOR or of any other holder of a copyright in all or part of the Intellectual Property shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Intellectual Property without prior written authorization of LICENSOR or such copyright holder. LICENSOR is and shall at all times be the sole entity that may authorize you or any third party to use certification marks, trademarks or other special designations to indicate compliance with any LICENSOR standards or specifications.

This Agreement is governed by the laws of the Commonwealth of Massachusetts. The application to this Agreement of the United Nations Convention on Contracts for the International Sale of Goods is hereby expressly excluded. In the event any provision of this Agreement shall be deemed unenforceable, void or invalid, such provision shall be modified so as to make it valid and enforceable, and as so modified the entire Agreement shall remain in full force and effect. No decision, action or inaction by LICENSOR shall be construed to be a waiver of any rights or remedies available to it.

None of the Intellectual Property or underlying information or technology may be downloaded or otherwise exported or reexported in violation of U.S. export laws and regulations. In addition, you are responsible for complying with any local laws in your jurisdiction which may impact your right to import, export or use the Intellectual Property, and you represent that you have complied with any regulations or registration procedures required by applicable law to make this license enforceable.

Abstract

This discussion paper is a result of the research project Citizen Observatory Web (COBWEB). COBWEB is supported by the European Commission through grant agreement 308513. The discussion paper describes a data model for the standardized exchange of citizen science sampling data.

Business Value

This discussion paper outlines a data model for citizen science data and outlines a future interoperability approach based on semantic web concepts. A standardized model allows for efficient exchange of citizen science data and supports re-use of data in other than originally planned contexts.

What does this ER mean for the Working Group and OGC in general

This discussion paper should be considered by the Citizen Science Domain Working Group as a baseline for future citizen science data exchange research and standardization.

How does this ER relates to the work of the Working Group

This report provides a solid base for future discussions on data models for citizen science in particular and interoperability in general.

Keywords

ogcdocs, citizen science, data models, interoperability

Proposed OGC Working Group for Review and Approval

Citizen Science Domain Working Group

Chapter 1. Introduction

The number of citizen science projects is constantly growing. Local, national, and international platforms feature new projects almost every month, resulting in huge number of observations. If these are to contribute to knowledge beyond short term project outputs these observations need to be gathered and potentially stored in Web accessible databases. For this data to fulfil that potential it needs to fulfil various aspects of comparability with other similar data, and this to be transparent as an aid to data discovery. Any automated integration of such data will require machine-readable metadata about the level of compatibility of such data. This paper describes those aspects of 'data interoperability'. It focuses on the concept of a scalable approach to "citizen science application profiles" using a standardized information model that ensures syntactic and semantic understanding of citizen science data collected by arbitrary sampling campaigns. Data compliant to 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 approach described in this paper is one out of two possible solutions. The second is briefly addressed and will be documented in detail in future publications.

1.1. Scope

This document focuses on three aspects. First, it describes common design decisions as they currently occur in many citizen science sampling campaigns. Second, it provides a data model that addresses most of the current interoperability issues by applying available models and technologies. Third, it outlines a path ahead that describes future solutions that further improve the level of interoperability.

1.2. Document contributor contact points

All questions regarding this document should be directed to the editor or the contributors:

Table 1. Contacts

Name	Organization
Ingo Simonis	OGC
Rob Atkinson	OGC

1.3. Future Work

The inherent challenges in defining a "one size fits all" approach have led to the identification of a need for, and a possible standards-based solution to, a flexible and scalable methdology for describing the semantic interoperability of data and how this is then realised in standardised data structures. This document is thus a first step towards a new approach to interoperability in heterogenous environments that supports semantic interoperability managed by the communities of practice. The concept needs further research and should be explored in other domains to evaluate its universal applicability.

1.4. Forward

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

Chapter 2. References

The following documents are referenced in this document. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. For undated references, the latest edition of the normative document referred to applies.

- OGC Abstract Specification Topic 20 Observation and Measurements (O&M)
- ISO 19109 (Geographic information - Rules for application schema)
- ISO 19136
- OGC Geography Markup Language (GML)
- SensorML
- SWECommon
- TimeseriesML
- GML coverages

Chapter 3. Terms and definitions

For the purposes of this report, the definitions specified in Clause 4 of the OWS Common Implementation Standard [OGC 06-121r9] shall apply. In addition, the following terms and definitions apply.

3.1. Citizen Science

```
tbd
```

Chapter 4. Conventions

4.1. Abbreviated terms

- APIApplication Program Interface
- COTSCommercial Off The Shelf

4.2. UML notation

Most diagrams that appear in this standard are presented using the Unified Modeling Language (UML) static structure diagram, as described in Subclause 5.2 of [OGC 06-121r9].

Chapter 5. Standardized Information Models to Optimize Exchange, Reusability and Comparability of Citizen Science Data

After a short introduction, this engineering report discusses the typical citizen science sampling campaign formalization and execution process to allow understanding the full set of requirements that needs 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 makes use of several other OGC standards and can be implemented and serialized in various ways, such as XML or JSON, or Semantic Web approaches based on triples and links. Here, focus is on XML exclusively.

The paper describes how to encode citizen science data in an efficient way. It does not define an application schema with value or other types of constraints. Instead, it should be considered as a guide that helps survey managers to share their data in an efficient way. For that reason, this report describes the usage of O&M and the derived citizen science profile for single and aggregated observations and addresses a number of semantic interoperability aspects.

The paper concludes with an outlook on future interoperability profile concepts that shall help simplifying the usage of specialized data models, support the automatic setup of data collection campaigns, and ensure shared semantics in heterogeneous environments.

Chapter 6. Citizen Science Data Collection and Exchange

Citizen science projects often fail to collect and document their data in standardized forms. Some projects design their own protocols, others use proprietary or disciplinary protocols. This report is based on experiences made in five citizen science projects co-funded by the European Commission. Based on these experiences, the goal was to develop a common model for citizen science data for different sampling campaigns based on existing standards. This engineering report outlines the fundamental structure of such a Citizen Science data profile. It uses existing standards such as OGC Sensor Web Enablement O&M, SensorML, and SweCommon together with TimeseriesML and GML-coverage.

The Citizen Science model developed herein reflects the current status quo in standardization of geospatial data produced in citizen science sampling campaigns. This status quo already allows the definition of a solid and semantically sound model. Nevertheless, interoperability can be further improved. Interoperability contracts and profiles allow to further simplify the sampling campaign definition process and enhance chances for common semantics across campaigns. This report describes the path ahead towards improved interoperability and enhanced ease of use for different citizen science data producers, consumers, survey designers, and quality managers.

Citizen Science data collection and exchange is not a streamlined and consistent process. Depending on the local settings and requirements, a number of aspects need to be decided upon:

- How do data models look like that support reusability of citizen science data even in other than the original context?
- How much information needs to be preserved and documented throughout citizen science data processing chains?
- How can citizen science data be shared efficiently?
- What level of semantics is required to ensure correct usage of citizen science data and how can it be realized?
- What standards need to be considered in order to maximize reusability of citizen science data?
- How can citizen science data be integrated with external data sets?
- How do citizen science data quality assurance processes look like and how can they be documented?
- How can citizen science data be made persistent and accessible (even beyond the lifetime of the original research project)?

These questions reflect the different perspectives of citizen science participants, e.g. data providers, survey designers, or data consumers.

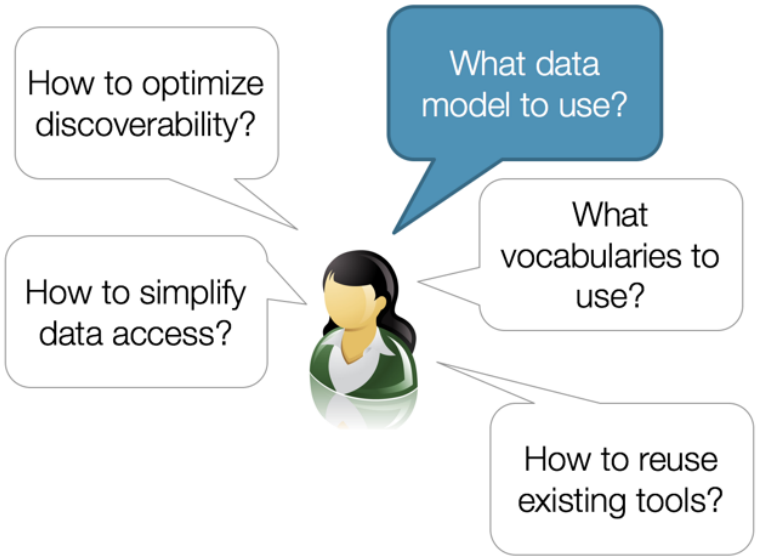

Figure 1. Particularly relevant aspects for survey designers

Survey designers usually develop their survey design and configure their survey tools and applications based on a set up local requirements. These may require survey data processing steps that include automated processes that require a certain set of metadata and the use of shared vocabularies. In the ideal case, survey designers already consider the future use of the survey data outside of the original context, which adds aspects such as discoverability or simplified access to the data to the list of requirements.

Data consumers have different motivations and requirements. They are focusing mostly on easy discovery and access, including spatio-temporal and thematic filtering together with easy-to-load data models and encodings.

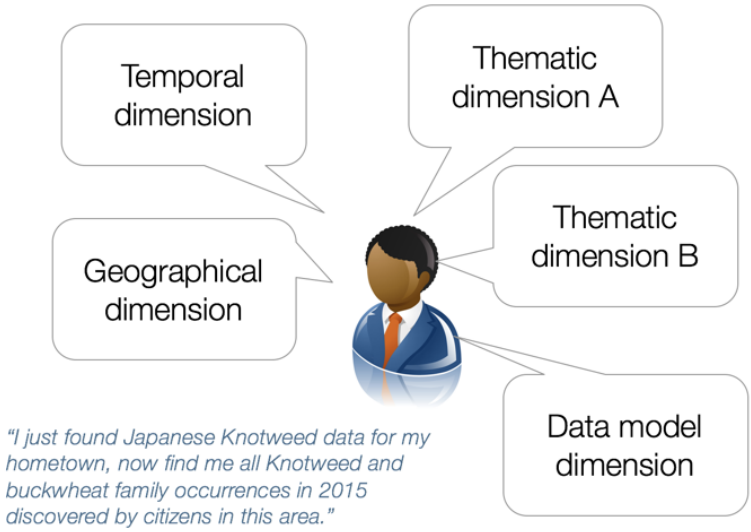

Figure 2. Particularly relevant aspects for data consumers

The goal of the Swe4CS data model was to integrate those different perspectives. In addition, the data model should support specific requirements as formulated as part of the citizen science data collection, processing, and presentation process described below.

6.1. The Citizen Science Process

The citizen science process commonly consists of five steps as illustrated in figure Citizen Science Process below. These steps are usually executed sequentially, though include loops and feedback cycles. The first step (left) includes the definition of the sampling campaign itself. The type and nature of the sampling protocol and corresponding observed properties is normally motivated by the survey objectives, but may take additional aspects into consideration such as quality assurance process or publication requirements. Once defined, an (usually) mobile app (survey app) is configured and made available to the citizens to support the data collection process. All raw data, i.e. the observation data as provided by citizens, is persistently stored and eventually published. 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.

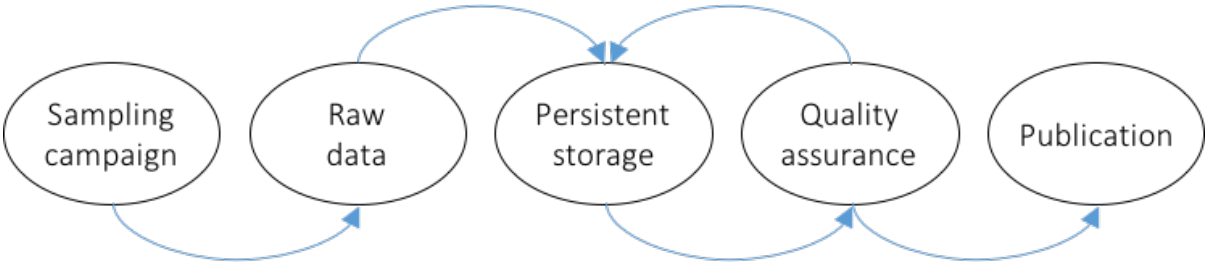

Figure 3. Citizen Science Process

The end of the citizen science process is the publication of the data, which in turn might be the beginning for another process. The publication can include all levels from raw to highly processed data. In this report, we make primarily use of an exemplary sampling campaign executed as part of the research project Citizen OBservatory WEB (COBWEB): The Japanese Knotweed survey.

6.2. Exemplary Sampling Campaign

Citizen science sampling campaigns often make use of mobile applications to support the citizen in the data collection process. Figure Japanese Knotweed Sampling Application illustrates such a mobile app that is used in Japanese Knotweed (Fallopia japonica) sampling campaigns. Occurrences of the invasive species Japanese Knotweed are monitored in many places of 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 defenses, roads, or sidewalks. It can also reduce the capacity of channels in flood defenses to carry water and accelerate river bank erosion (Bailey 2003).

One of the primary goals during sampling campaign definition is to avoid any ambiguities in the understanding of observed properties, used sensing techniques and hardware, and sampling protocols between survey designers and citizens. Both groups shall share common semantics of all terms used. Though the interaction options with a mass of citizens is limited, and little influence exists to ensure that citizens make themselves familiar with the applied semantics, it is crucial that clear instructions and definitions of all aspects, e.g. observed properties or sensors, are available. As the screen space on mobile devices is limited, it is necessary to provide resolvable links to all observed properties and clear explanations of the sampling protocol and applied sensing hardware.

The application shown in Figure 2 illustrates this aspect using the observed properties as an example. Though “Plant height” is probably easy to understand, “Evidence of management” leaves considerably more room for interpretation. There are many aspects related to user interface design that cannot discussed here. Instead, focus is laid on the definition of the semantics, which are not visible to the user.

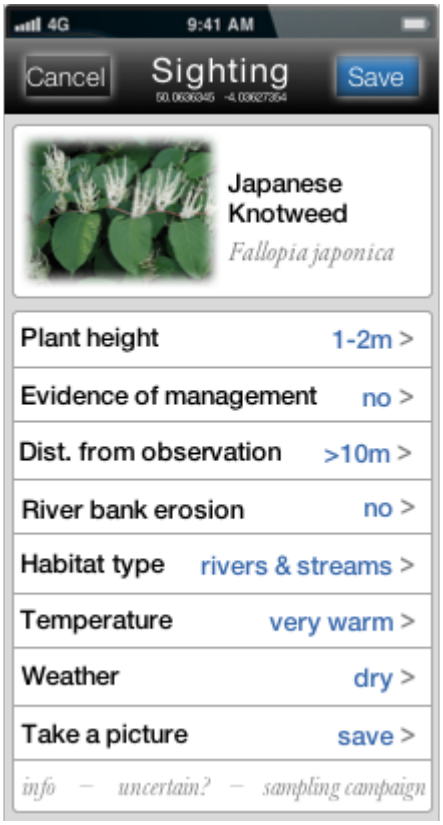

Figure 4. Japanese Knotweed Sampling Application

If all observed properties would be using definitions in the form of simple names, semantics would be limited to the understanding of the survey designer and, given that sufficient descriptive data are provided as part of the mobile app, to the citizens participating in the survey. If instead fully qualified names in the form of resolvable URLs would be used, then the raw data becomes meaningful even to external users that have not used the mobile app but only received the raw data. Labels can still be used for display purposes.

Other examples and survey types are used if necessary. This report does not cover any software architecture that could be used for fully standards-based data creation and exchange.

Chapter 7. Citizen Science Data Models

SWE For Citizen Science (Swe4cs) is a common data model suitable for collecting and sharing citizen science data. It is based on the ISO 19156 / OGC Abstract Specification Topic 20 Observation and Measurements (O&M) standard and implemented as an ISO 19109 (Geographic information - Rules for application schema) compliant application schema, which implements key elements from other ISO standards following the conventions defined in ISO 19136 / OGC Geography Markup Language (GML). Swe4cs is based on the OGC core model, with the General Feature Model (defined in ISO 19109) as a metamodel for representing features in application schemas; therefore, Swe4cs data can easily be processed and integrated with other OGC compliant data such as coverage based on satellite data, sensor data from ground stations, or reference data sets provided by public bodies. In addition, Swe4cs uses standards from the OGC Sensor Web Enablement (SWE) family, a suite of standards initially developed to realize the Sensor Web. Originally designed to network sensors, the scope of SWE has broadened to include humans as sensors and sensor data processing capacities. Swe4cs uses the OGC standards SensorML, a robust and semantically-tied means of defining assets, to describe sensors used in sampling campaigns, and SWECommon, a low-level data model for exchanging sensor related data between nodes of the OGC Sensor Web Enablement (SWE) framework. To express time series and coverage data, Swe4cs implements OGC standards TimeseriesML and GML coverages. The package dependencies are illustrated in the figure below.

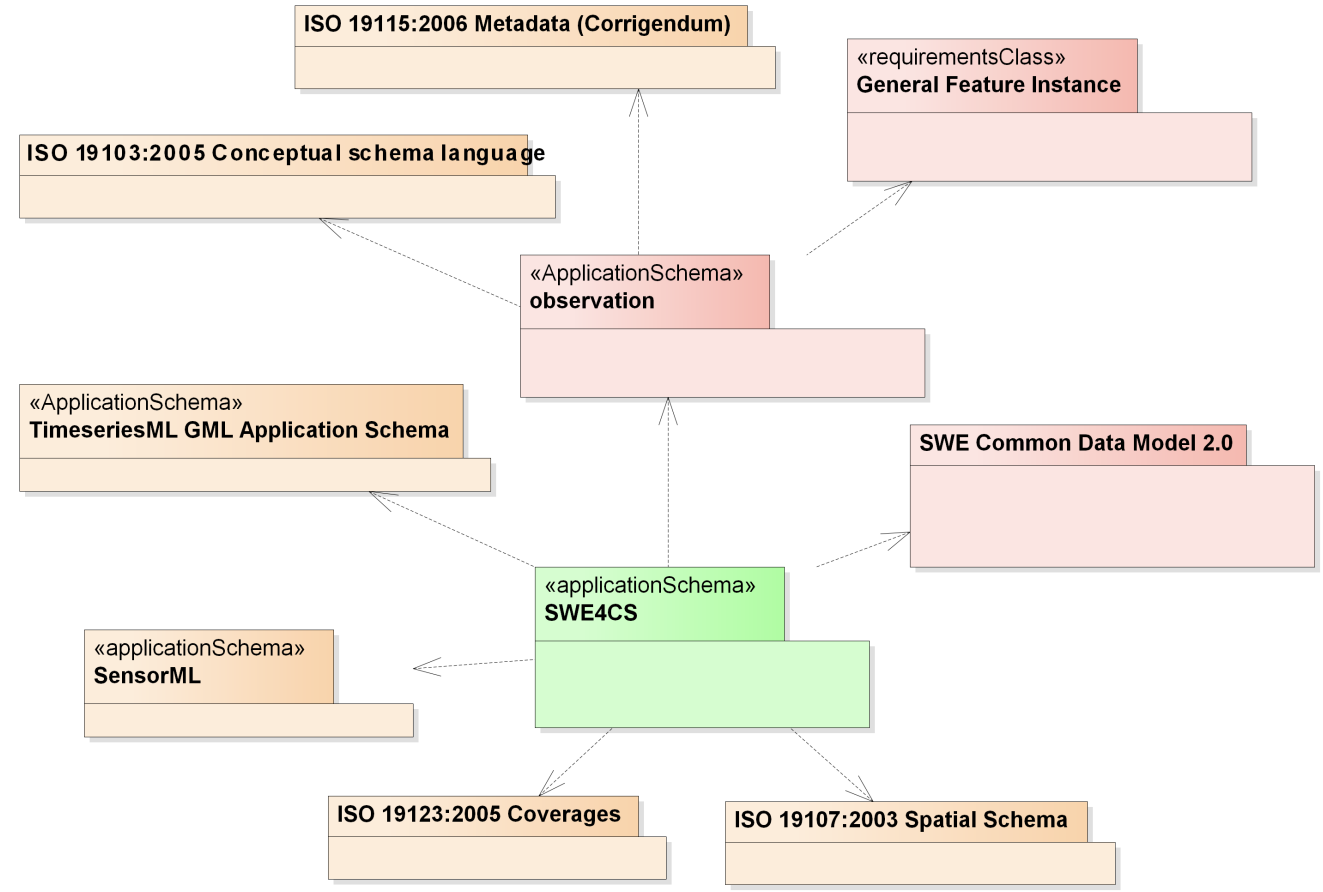

Figure 5. Package dependencies of the Swe4CS data model

The currently available Swe4cs model has been designed with a focus on reusing existing standards to the greatest degree possible. As it stands, Swe4cs is suitable for expressing the key elements of a citizen science observation, i.e. the observed property(ies), the results, temporal and spatial aspects,

potentially used hardware, and information about the volunteer herself. It is envisioned that future interoperability pilots are required to extend the model and address remaining interoperability concerns, including:

- Definition of the sampling protocol
- Semantic pointers to shared ontologies
- Definition of a Swe4cs constraint model of SensorML to overcome design issues caused by the extreme flexibility of SensorML
- Data quality and user feedback
- Further serializations such as JSON and RDF

7.1. General Design Decision

Defining a data model for a particular community commonly involves a number of design decisions. In the case of Swe4CS, it was decided after long debates to favor a data model that is fully based on existing core standards without profiling them using specializations. The reason is that specializations require adapted client applications to parse serialized data. On the other side, specializations would allow more fine-grained adaptations to particular needs - but these are commonly an issue for interoperability anyway.

7.2. O&M Base Model

Observation & Measurement (O&M) provides a general model to encode observations done by either humans, machines, or software processes and is therefore suitable for citizen science data. Citizen science data often results from human observations or is provided by citizen scientists using mobile sensor assets. O&M features a number of levels of freedom that require design decisions and constraints in order to result in a model for interoperable data exchange. These levels of freedom are illustrated in figure Levels of Freedom in the O&M model. The O&M observation model is the core of all citizen science observations. We need to understand a single observation first, before we will look how to aggregate multiple observations in the next chapter.

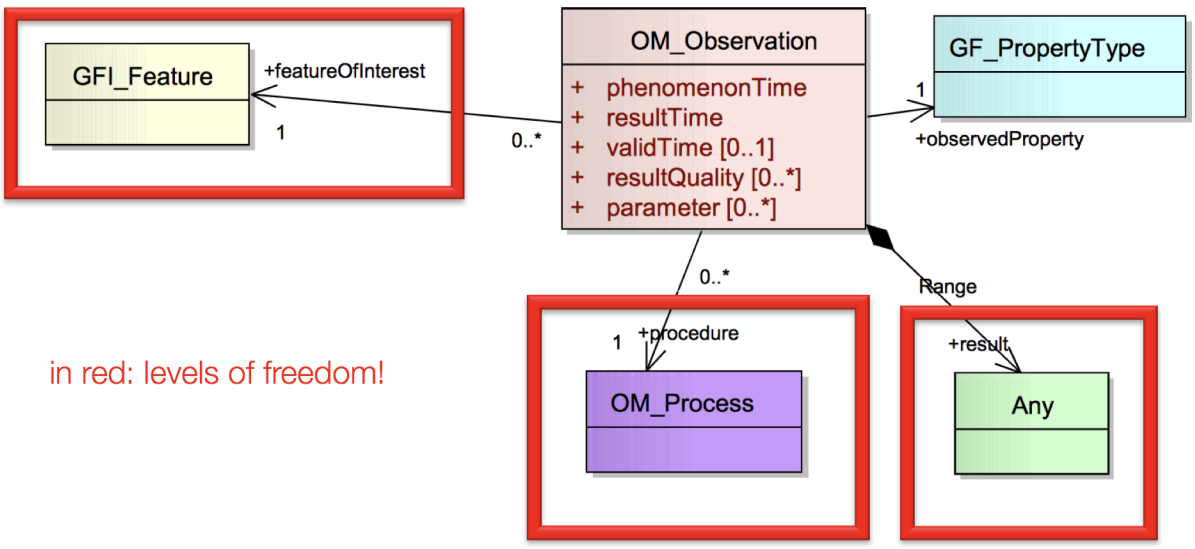

Figure 6. Levels of Freedom in the O&M model

Following the decision to use O&M as is without further specialization, design decisions had to be taken to improve interoperability of the *featureOfInterest*, the *procedure*, and the *result*. Before those are further investigated, the O&M observation properties are briefly discussed.

7.2.1. Elements Inherited From GML Super Model

O&M observation inherits a number of elements from it super model GML, the geography markup language which defines an abstract feature as one of its core elements. The *gml:AbstractFeature* and its type *gml:AbstractFeatureType* implement the ISO 19109 General Feature Model (we don't discuss the details of the internal structure of the O&M Schema with all its complex types and property types here. O&M_Observation is developed in UML and follows the UML-to-XML Schema rules as defined in ISO 19136. Interested readers are referred to the specifications ISO 19109, 19136, and 19156). Therefore, O&M Observation has a number of properties not directly visible in the O&M model.

Here, we focus on elements that are relevant for citizen science data only. All three elements are optional:

- (1) **gml:description**: General description of this observation
- (2) **gml:name**: Name of the observation that could be used as a label in a client application
- (3) **gml:boundedBy**: Only relevant if the bounding box of all locations aggregated in an observation shall be documented

XML example: Elements inherited from GML

```
<gml:description>Snowdonia National Park, Japanese Knotweed Survey
2015</gml:description> ①
<gml:name>Japanese Knotweed Observation</gml:name> ②
<gml:boundedBy> ③
    <gml:Envelope srsName="urn:x-ogc:def:crs:EPSG:6.11:4326">
        <gml:lowerCorner>-5.009766 51.266412</gml:lowerCorner>
        <gml:upperCorner>-2.927977 53.127076</gml:upperCorner>
    </gml:Envelope>
</gml:boundedBy>
```

7.2.2. phenomenonTime

The *phenomenonTime* describes the time the *observedProperty* was observed.

XML example: phenomenonTime

```
<om:phenomenonTime>
   <gml:TimeInstant gml:id="t001">
      <gml:timePosition>2015-07-07T10:32:48.460Z</gml:timePosition>
   </gml:TimeInstant>
</om:phenomenonTime>
```

7.2.3. resultTime

The *resultTime* describes the time the result value(s) was assigned to the observation. In most cases, this time is similar to the *phenomenonTime* defined above. It is different if there is a time gap between the actual observation and the assignment of the result value. This might be the case if a water sample is taken from a river (*phenomenonTime*), analyzed in a lab and the value assigned to that observation once the analysis is completed (*resultTime*).

XML example: resultTime

```
<om:resultTime xlink:href="#t001"/>
```

7.2.4. validTime, resultQuality, parameter

These three elements may be of little relevance for the original observation, because citizen scientists are often not able to determine how long the observation they did is valid (*validTime*) or the quality of the assigned result value (*resultQuality*). Nevertheless, during the quality assurance processes and control flows citizen science raw data may get annotated with additional information. Then, validity and quality parameters may be set.

The *parameter* element is a generic extension point that allows adding context specific key-value pairs to observation data without breaking the schema. Clients that cannot make sense of these additional parameters are required to ignore these.

7.2.5. procedure

The procedure property of an observation defines the process used to generate the observation. The O&M model uses the empty *OM_Process* that needs to be further defined in order to achieve interoperability. It is recommended to follow the Timeseries Profile of Observations and Measurements specification (OGC 15-042r3), which is about to be released to the public soon.

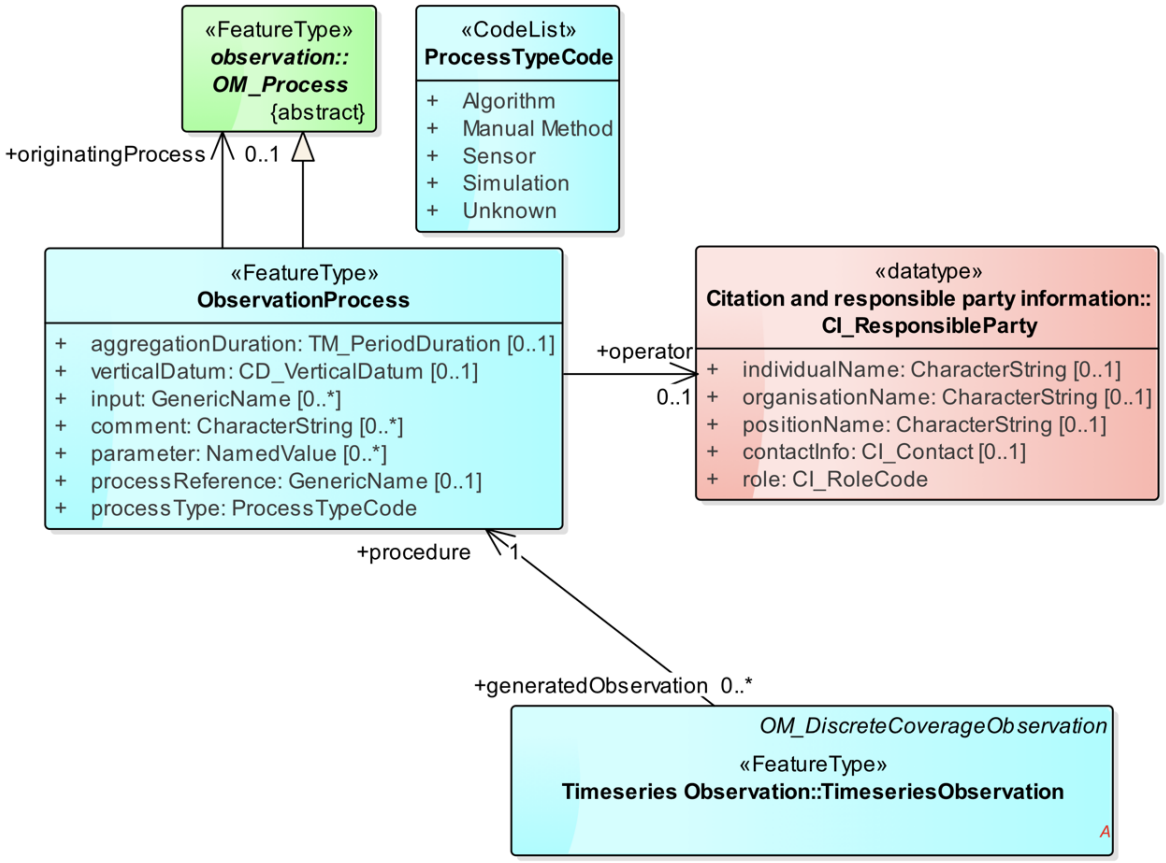

Figure 7. Observation process feature type, source OGC 15-042r3

tsml:ObservationProcess defines a number of properties that are less relevant in the context of citizen science. At the same time, *tsml:ObservationProcess* provides the same *parameter:NamedValue* extension mechanism as *om:Observation*.

We recommend to use the following properties:

- **processType** to define the type of the process. From the *ProcessTypeCode* code list, the following properties are important:
  - **Manual Method** if the observation was performed by a human **without** any additional hardware
  - **Sensor** if the observation was performed by a human **with** additional hardware
- **processReference** to link to the sampling protocol. The sampling protocol includes all rules and guidelines on how the sampling should be performed. Currently, there are no models available to define sampling protocols in a standardized way. Instead, sampling protocols are usually provided in text form, often with accompanying images and often enough several pages long.
- **parameter**: the generic extension mechanism should be used to provide information about the sensor(s) being used to generate the observation. In the ideal case, the sensor description is provided using SensorML.
- **operator** to define the citizen scientist who performed the observation. A potential issue here: *operator* can only be provided once. If more than one person has performed the observation,

the data about the additional person(s) needs to be added by alternative ways that require further discussion. The approach used herein In the case, *parameter* shall be used to identify the additional persons. Examples are given below for one, two, and anonymous.

The following examples illustrate the *procedures* definition. Annotations in the examples help understanding the various elements.

XML example: procedure; anonymous citizen scientist with cellphone sensors

```
<om:procedure>
  <tsml:ObservationProcess gml:id="op1">
    <!-- processType defines observation performed by human with sensor -->
    <tsml:processType
xlink:href="http://www.opengis.net/def/waterml/2.0/processType/Sensor"/>
    <!-- processReference defines sampling protocol -->
    <tsml:processReference
xlink:href="https://dyfi.cobwebproject.eu/skos/JapaneseKnotweedSamplingProtocol"/>
    <!-- if a sensor is used, provide the link to the sensor definition here. Use
SensorML if possible -->
    <tsml:parameter>
      <om:NamedValue>
        <om:name xlink:href="http://www.opengis.net/def/property/OGC/0/SensorType"/>
        <om:value>http://www.motorola.com/XT1068</om:value>
      </om:NamedValue>
    </tsml:parameter>
    <!-- operator defines the citizen scientist producing this observation -->
    <tsml:operator>
      <!-- anonymous observation producer from ISO 19115, roleCode "Expert" -->
      <gmd:CI_ResponsibleParty>
        <gmd:role>
          <gmd:CI_RoleCode codeList="https://dyfi.cobwebproject.eu/skos#CI_roleCodes"
codeListValue="Expert">
          </gmd:CI_RoleCode>
        </gmd:role>
      </gmd:CI_ResponsibleParty>
    </tsml:operator>
  </tsml:ObservationProcess>
</om:procedure>
```

CAUTION

This report shows a number of hyperlinks in the XML examples. Not all of these hyperlinks resolve. Reason is that this report is a discussion paper. No terms have been registered with the OGC Naming Authority so far. Existing names from the OGC Definitions Service have been used wherever available. In all other cases, the URLs start with https://dyfi.cobwebproject.eu.

The following example shows how the citizen scientist can be identified. The ISO 19115 element CI_ResponsibleParty allows providing more detail such as email, phone number etc. as illustrated in figure below.

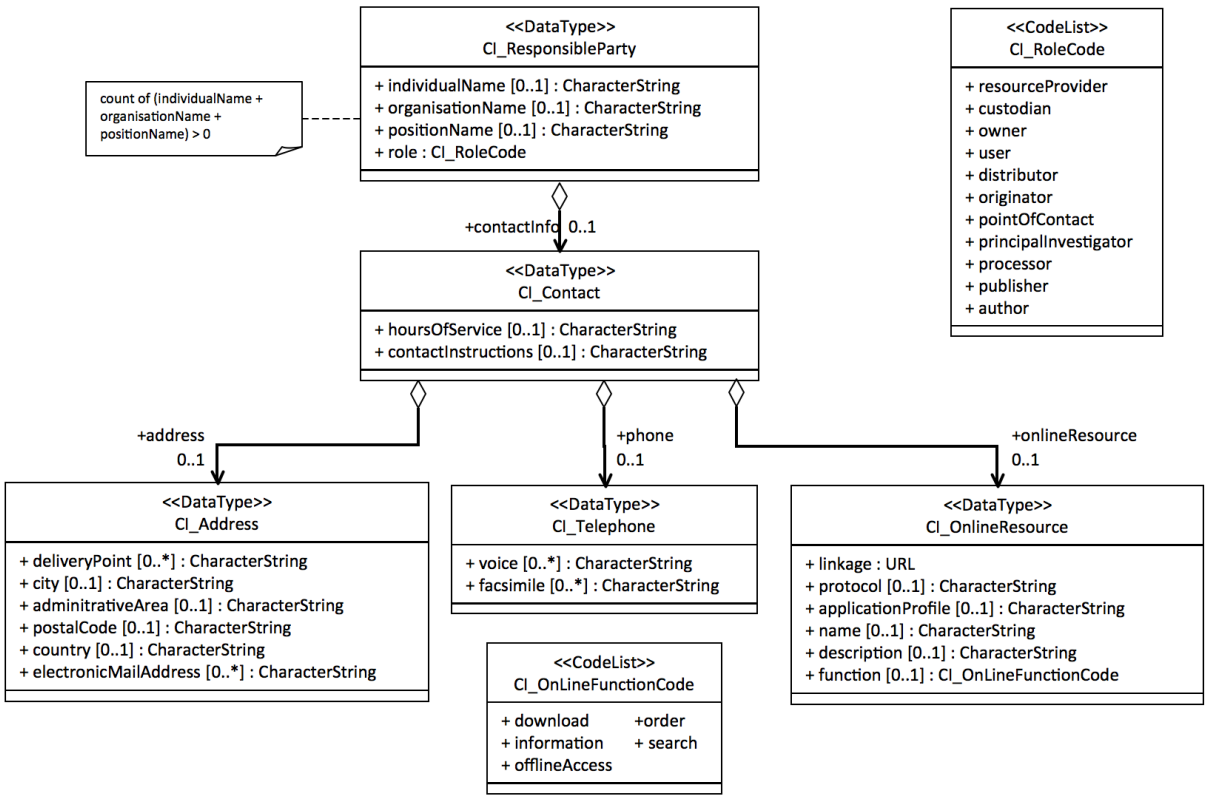

Figure 8. CI_ResponsibleParty object describes people and organizations that are related to a resource and their roles

The following XML example illustrates how the *operator* element of the example above could be modified to provide more details about the citizen scientist who provided the observation.

XML example: operator; identified citizen scientist. CI_RoleCode from ISO 19115

```
<tsml:operator>
  <gmd:CI_ResponsibleParty>
    <gmd:individualName>
      <gco:CharacterString>Ingo Simonis</gco:CharacterString>
    </gmd:individualName>
    <gmd:organisationName>
      <gco:CharacterString>OGC</gco:CharacterString>
    </gmd:organisationName>
    <gmd:role>
      <gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml"
codeListValue="resourceProvider"/>
    </gmd:role>
  </gmd:CI_ResponsibleParty>
</tsml:operator>
```

7.2.6. observedProperty

The property(ies) that is/are of interest in the citizen science sampling campaign. The *observedProperty* might be a single aspect, such as occurrences of a specific species, e.g. Japanese Knotweed, or a complex of multiple aspects. In all cases, the *observedProperty*'s details can be retrieved from the result section described further below. Following the link shall lead to a detailed description of the observedProperty. If available, existing vocabularies shall be used.

XML example: observedProperty

```
<om:observedProperty xlink:href="https://dyfi.cobwebproject.eu/skos/bogs"/>
```

7.2.7. featureOfInterest

The *featureOfInterest* is a tricky element. Following the rules and guidelines provided in Observations and Measurements – Part 2 - Sampling Features, the *featureOfInterest* can describe the ultimate feature of interest, called *domain feature*, or a (spatial) sampling feature. Sampling features are used if the ultimate feature of interest only allows observations being made on a subset of the complete feature, with the intention that the sample represents the whole. This is for example the case if we sample Japanese Knotweed in Snowdonia National Park. We cannot assume that we sample the whole national park, but only walk randomly around and spot knotweed here and there. The random walk is a sampling feature, and if the citizen walks around long enough, we can assume that the observations made represent the whole park. In this case, the *featureOfInterest* is the sampling feature that in this case represents the whole ultimate feature of interest, the domain feature Snowdownia National Park.

If the citizen scientist tracks his path, i.e. can provide the full trajectory, then the *featureOfInterest* is a *spatial sampling feature* in the form of a *SF_SamplingCurve* that reveals all locations of the citizen scientist during his walk in Snowdonia National Park. This is highly valuable information, as it allows estimating the coverage of the sampling campaign and helps understanding if areas without any knotweed occurrences have not been explored or de facto have no knotweed growing. In this case, the *featureOfInterest* would be defined in more detail: It contains the path itself in the form of a shape definition, and the ultimate feature of interest in the form of the *sampledFeature*.

The following examples illustrate this concept.

XML example: featureOfInterest defines a domain feature

```
<om:featureOfInterest
xlink:href="https://dyfi.cobwebproject.eu/skos/Snowdonia_National_Park"/>
```

The link to the feature of interest can be a call to a Web Feature Service also.

XML example: featureOfInterest defines a domain feature as accessible at a Web Feature Service (WFS) instance

```
<om:featureOfInterest
xlink:href="http://example.com/wfs?service=WFS&request=GetFeature&version=2.0.0&featur
eID=SnowdoniaNationalPark"/>
```

The following example illustrates a survey with existing trajectory data.

XML example: featureOfInterest defines a spatial sampling feature (1) with sampled feature (2) and shape information (3)

```
<om:featureOfInterest>
  <sams:SF_SpatialSamplingFeature gml:id="sf001"> ①
    <sf:type xlink:href="http://www.opengis.net/def/samplingFeatureType/OGC-
OM/2.0/SF_SamplingCurve"/>
    <sf:sampledFeature
xlink:href="https://dyfi.cobwebproject.eu/skos/Snowdonia_National_Park"/> ②
    <sams:shape> ③
      <gml:Curve gml:id="sc1" srsName="urn:ogc:def:crs:EPSG:6.8:3857">
        <gml:segments>
          <gml:LineStringSegment>
            <gml:posList>52.4096027 -4.0782345 52.4095827 -4.0782352 52.409551
-4.0782377 52.4094811 -4.0782878 52.4095147 -4.0789545 52.409452 -4.0787875 52.409124
-4.0785565 52.4091245 -4.0782447 52.4097877 -4.0782454 52.4097797 -
4.0781024</gml:posList>
          </gml:LineStringSegment>
        </gml:segments>
      </gml:Curve>
    </sams:shape>
  </sams:SF_SpatialSamplingFeature>
</om:featureOfInterest>
```

7.2.8. result

The *result* property provides the actual observation result data. *om:result* points to a generic placeholder *Any*, which has been further specialized in the citizen science profile to *SweCommon DataRecord* to encode all result data. For a single observation, this is straight forward and illustrated in the XML example below.

XML example: result property defining topography type, a photo, and three plant species

```
<om:result>
  <swe:DataRecord>
    <swe:field name="topographyType">
      <swe:Text definition="https://dyfi.cobwebproject.eu/skos/topographyType">
        <swe:value>Mountain</swe:value>
      </swe:Text>
    </swe:field>
    <swe:field name="photo">
      <swe:Text definition="https://dyfi.cobwebproject.eu/skos/photo">
        <swe:value>https://dyfi.cobwebproject.eu/1.3/pcapi/records/local/2338e388-
f34e-25d9-945c-54cffd9c46c2/ob (11)/1434891560330.jpg</swe:value>
      </swe:Text>
    </swe:field>
    <swe:field name="plants">
      <swe:DataRecord> ①
        <swe:field name="plant">
          <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
            <swe:value>Bog cotton</swe:value>
          </swe:Text>
        </swe:field>
        <swe:field name="plant">
          <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
            <swe:value>Common rush</swe:value>
          </swe:Text>
        </swe:field>
        <swe:field name="plant">
          <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
            <swe:value>Other</swe:value>
          </swe:Text>
        </swe:field>
      </swe:DataRecord>
    </swe:field>
  </swe:DataRecord>
</om:result>
```

The *DataRecord* contains any number of fields with semantics coded in the *definition* attribute. It is recommended to link to common vocabularies to improve reusability. Alternatively, if survey managers set up their own vocabulary as illustrated here, it is recommended to use the SKOS Simple Knowledge Organization System with links to external, commonly used vocabularies, as availability of the vocabulary can be assured and shared semantics achieved through links to external ontologies.

As illustrated in the example above, *SweCommon* allows nesting further *DataRecord_s. This allows adding more than one value for a single type, here done with three different plants that have been observed. Instead of (1) _swe:DataRecord* a *swe:DataArray* could be used, but it does not add any value here, as the number of entries fixed.

NOTE

In principle, O&M supports associations to other observations. This is - in principle - a very powerful linking mechanism that allows describing associations to other observations. In practice, it turned out that users prefer to have all relevant data in a single file with minimum links embedded. Therefore, we don’t make use of this mechanism in this engineering report.

Chapter 8. Observation Collections and Aggregations

The previous chapter introduced the citizen science model for simple and complex observations, i.e. individual observations done by citizen scientists. This chapter introduces observation collections and aggregation patterns. These patterns will make use of the citizen science model as described, but extend it to allow different types of aggregations, e.g. all observations by a specific citizen scientist, or all observations in a given area, etc. These collections and aggregations are either produced directly by the (mobile) applications used by citizen scientists to optimize data transfer, or result from data analysis and filtering processes.

The following figure illustrates different aspects that need to be taken into consideration. Here, a citizen scientist walks from left to right and detects four occurrences of Japanese Knotweed (1). These occurrences can be modeled in different ways, depending on the sampling protocol and other potentially available information. One option is *DiscretePointCoverage* (2). Here, the extent of the coverage describes the projected path of the scientist to the plane. Occurrences of Japanese Knotweed are modeled as *MultiPoint* elements in a *MultiPointCoverage*. The second option, if the exact path of the scientist is not available or does not matter, is a simple *TimeSeries*. The third option and fourth option illustrated in the figure below provide the trajectory and encode all occurrences together with the sampling path (4), or, if the sampling protocol defines to watch out for occurrences 10m left and right of the path, as a sampling polygon (5).

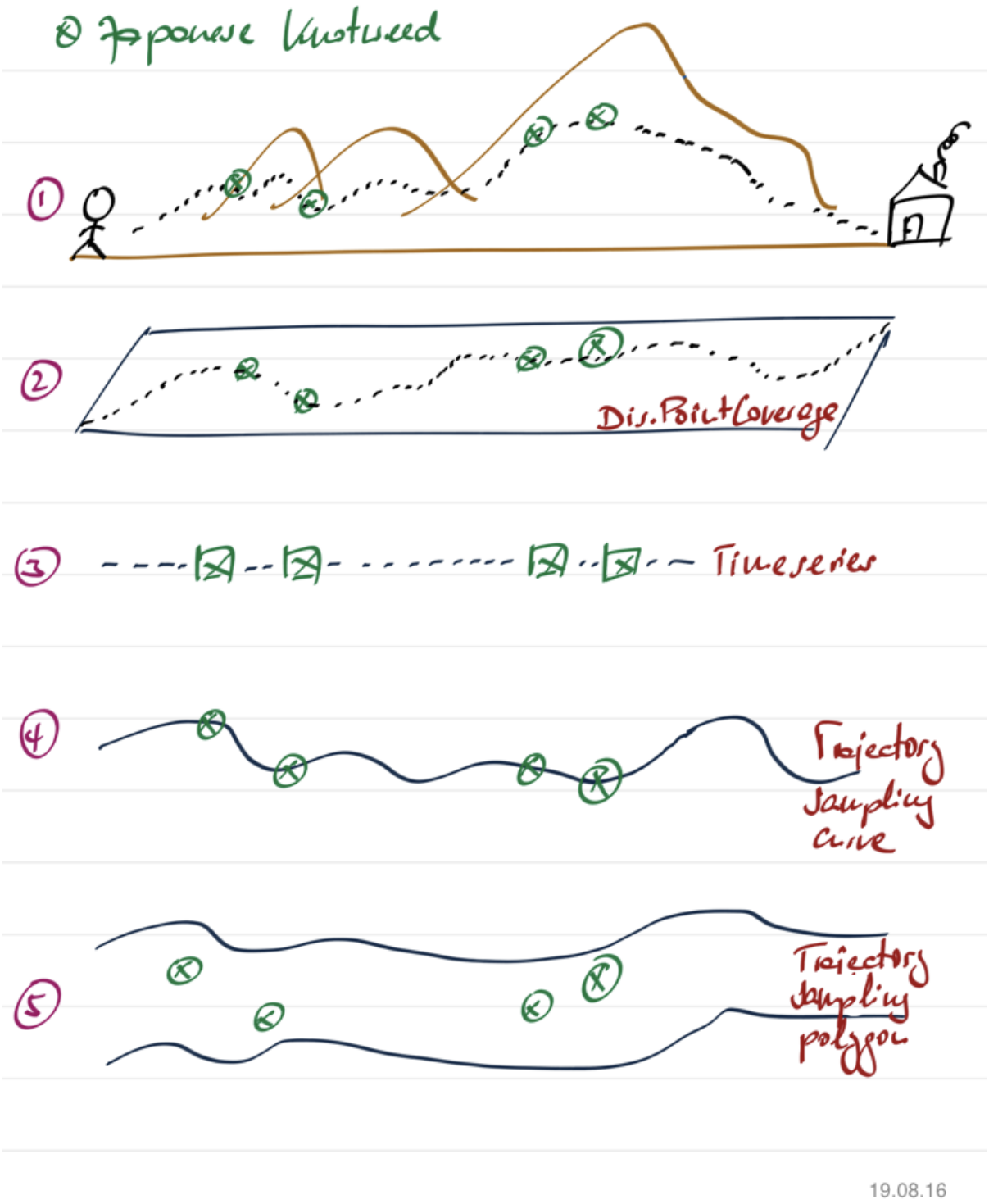

Figure 9. Observation aggregation patterns

The following XML examples show the encoding of the various options described above. To improve readability, the number of Japanese Knotweed occurrences has been reduced to two and all elements except for *featureOfInterest* and *result* have been omitted.

8.1. Aggregation encoding: DiscretePointCoverage

The first examples illustrates Japanese Knotweed occurrence encoded as a *DiscretePointCoverage*. This is the recommended approach if a compact serialization is intended and the number of observed properties and corresponding result elements remains stable. Two different serializations

are possible. The first one below uses *gmlcov:MultiPointCoverage*, the second further below uses *SweCommon DataArray*. Both express the same data. It requires further discussion on which one should be preferred.

XML example: DiscretePointCoverage using gmlcov:MultiPointCoverage

```
<om:featureOfInterest
xlink:href="https://dyfi.cobwebproject.eu/skos/Snowdonia_National_Park"/>
<!-- the result uses a gmlcov:MultiPointCoverage. Each observation occurrence results
in a   MultiPoint:member gml:Point as part of the domain set. Values are provided as
part of the rangeSet,   which is described in rangeType -->
<om:result>
  <gmlcov:MultiPointCoverage gml:id="c001">
    <gml:domainSet>
      <gml:MultiPoint gml:id="mp0001_C0042"
srsName="http://www.opengis.net/def/crs/EPSG/0/4979">
        <gml:pointMember>
          <gml:Point gml:id="sp1">
            <gml:pos srsName="urn:ogc:def:crs:EPSG:6.8:3857">52.409602775074845
-4.078234501964251</gml:pos>
          </gml:Point>
        </gml:pointMember>
        <gml:pointMember>
          <gml:Point gml:id="sp2">
            <gml:pos srsName="urn:ogc:def:crs:EPSG:6.8:3857">53.1139046729
-3.78766989708</gml:pos>
          </gml:Point>
        </gml:pointMember>
      </gml:MultiPoint>
    </gml:domainSet>
    <gml:rangeSet>
      <!-- Note: Order of components within a composite rangeSet value (e.g. tuples in
tupleList) corresponds to document order of the rangeType elements (e.g. fields). -->
      <gml:DataBlock>
        <gml:rangeParameters/>
        <gml:tupleList decimal=".">                        2015-11-
03T15:45:41,https://dyfi.cobwebproject.eu/6265141986.jpg,1.0
2015-11-07T13:06:48,https://dyfi.cobwebproject.eu/433ds70609.jpg,2.0
</gml:tupleList>
      </gml:DataBlock>
    </gml:rangeSet>
    <gmlcov:rangeType>
      <swe:DataRecord>
        <swe:field name="samplingTime">
          <swe:Time
definition="http://www.opengis.net/def/property/OGC/0/SamplingTime">
            <swe:label>Sampling Time</swe:label>
            <swe:uom xlink:href="http://www.opengis.net/def/uom/ISO-
8601/0/Gregorian"/>
          </swe:Time>
        </swe:field>
```

```
        <swe:field name="photo">
          <swe:Text definition="https://dyfi.cobwebproject.eu/skos/photo" />
        </swe:field>
        <swe:field name="approxPlantHeight">
          <swe:Quantity
definition="https://dyfi.cobwebproject.eu/skos/approxPlantHeight">
            <swe:uom code="m"/>
          </swe:Quantity>
        </swe:field>
      </swe:DataRecord>
    </gmlcov:rangeType>
  </gmlcov:MultiPointCoverage>
</om:result>
```

XML example: DiscretePointCoverage using SweCommon

```
<om:featureOfInterest
xlink:href="https://dyfi.cobwebproject.eu/skos/Snowdonia_National_Park"/>
<!-- the result uses a SweDataArray. Each observation occurrence results in a value.
The value types are described in the swe:elementType -->
<om:result>
    <swe:DataArray>
        <swe:elementCount>
            <swe:Count>
                <swe:value>5</swe:value>
            </swe:Count>
        </swe:elementCount>
        <swe:elementType name="occurrence">
            <swe:DataRecord>
                <swe:field name="timestamp">
                    <swe:Time
definition="https://dyfi.cobwebproject.eu/skos/phenomenonTime">
                        <swe:uom xlink:href="http://www.opengis.net/def/iso-
8601/Gregorian+UTC"/>
                    </swe:Time>
                </swe:field>
                <swe:field name="lat">
                    <swe:Quantity
definition="http://sweet.jpl.nasa.gov/2.0/spaceCoordinates.owl#Latitude" axisID="Lat">
                        <swe:uom code="deg"/>
                    </swe:Quantity>
                </swe:field>
                <swe:field name="lon">
                    <swe:Quantity
definition="http://sweet.jpl.nasa.gov/2.0/spaceCoordinates.owl#Latitude" axisID="Lon">
                        <swe:uom code="deg"/>
                    </swe:Quantity>
                </swe:field>
                <swe:field name="photo">
                    <swe:Text definition="https://dyfi.cobwebproject.eu/skos/photo" />
                </swe:field>
```

```
                <swe:field name="approxPlantHeight">
                    <swe:Quantity
definition="https://dyfi.cobwebproject.eu/skos/approxPlantHeight">
                        <swe:uom code="m"/>
                    </swe:Quantity>
                </swe:field>
            </swe:DataRecord>
        </swe:elementType>
    </swe:DataArray>
    <swe:TextEncoding blockSeparator="&#10;" tokenSeparator="," decimalSeparator="."/>
    <swe:values>
        2015-11-03T15:45:41,52.409602775074845,-
4.078234501964251,https://dyfi.cobwebproject.eu/6265141986.jpg,1.0
        2015-11-07T13:06:48,53.1139046729,-
3.78766989708,https://dyfi.cobwebproject.eu/433ds70609.jpg,2.0
    </swe:values>
</om:result>
```

8.2. Aggregation encoding: Collection of observations

The second examples illustrates Japanese Knotweed occurrence encoded as a collection of individual observations and thus represents a simple form of time series. This option allows to capture different number of elements per observation, e.g. at the first observation lists two species, Sphagnum moss and Bog cotton, where as the next lists Bog cotton, Common rush, and Star moss. Those varying number of observed property results are best serialized using the *swe:DataRecord* approach.

XML example: Collection of individual observations

```
<?xml version="1.0" encoding="UTF-8"?>
<gml:FeatureCollection gml:id="JapanesKnotweedFeatureCollection_1">
  <gml:description>Collection of Japanese Knotweed observations, Dyfie,
Wales</gml:description>
  <gml:name>Observation Collection 1</gml:name>
  <gml:featureMember>
    <om:OM_Observation gml:id="_x3ebvgy65">
      <gml:description>Swe4CitizenScience example observation from the Japanese
Knotweed field sampling campaign</gml:description>
      <gml:name>Japanese Knotweed Observation, pure OM, TSML and SWECommon</gml:name>
      <om:type xlink:href="http://www.opengis.net/def/observationType/OGC-
OM/2.0/OM_ComplexObservation"/>
      <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>
        <tsml:ObservationProcess gml:id="op1-moto">
          <!-- processType defines observation performed by human with sensor -->
```

```
          <tsml:processType
xlink:href="http://www.opengis.net/def/waterml/2.0/processType/Sensor"/>
          <!-- processReference defines sampling protocol -->
          <tsml:processReference
xlink:href="https://dyfi.cobwebproject.eu/skos/JapaneseKnotweedSamplingProtocol"/>
          <!-- if a sensor is used, provide the link to the sensor definition here.
Ideayll, value points to SensorML definition -->
          <tsml:parameter>
            <om:NamedValue>
              <om:name
xlink:href="http://www.opengis.net/def/property/OGC/0/SensorType"/>
              <om:value>http://www.motorola.com/XT1068</om:value>
            </om:NamedValue>
          </tsml:parameter>
          <!-- operator defines the human producing this observation -->
          <tsml:operator>
            <!-- anonymous observation producer from ISO 19115 -->
            <gmd:CI_ResponsibleParty>
              <gmd:role>
                <gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml"
codeListValue="resourceProvider"/>
              </gmd:role>
            </gmd:CI_ResponsibleParty>
          </tsml:operator>
        </tsml:ObservationProcess>
      </om:procedure>
      <!-- namedParameter to provide the sampling campaign identifier -->
      <om:parameter>
        <om:NamedValue>
          <om:name xlink:href=
"https://dyfi.cobwebproject.eu/skos/SamplingCampaignID"/>
          <om:value>SnowdoniaNationalParkJapaneseKnotweedSurvey2015_Cleaned</om:value>
        </om:NamedValue>
      </om:parameter>
      <om:observedProperty xlink:href="https://dyfi.cobwebproject.eu/skos/BogTaxa"/>
      <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="taxon">
            <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
              <swe:value>Sphagnum moss</swe:value>
            </swe:Text>
          </swe:field>
          <swe:field name="taxon">
            <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
              <swe:value>Bog Cotton</swe:value>
            </swe:Text>
          </swe:field>
        </swe:DataRecord>
      </om:result>
    </om:OM_Observation>
  </gml:featureMember>
  <gml:featureMember>
    <om:OM_Observation gml:id="_vz9f5kbbe">
      <om:type xlink:href="http://www.opengis.net/def/observationType/OGC-
OM/2.0/OM_ComplexObservation"/>
      <om:phenomenonTime>
        <gml:TimeInstant gml:id="t002">
          <gml:timePosition>2015-11-03T16:06:48.394Z</gml:timePosition>
        </gml:TimeInstant>
      </om:phenomenonTime>
      <om:resultTime xlink:href="#t002"/>
      <!-- procedure links to procedure data from observation above -->
      <om:procedure xlink:href="#op1-moto"/>
      <!-- namedParameter to provide the sampling campaign identifier -->
      <om:parameter>
        <om:NamedValue>
          <om:name xlink:href=
"https://dyfi.cobwebproject.eu/skos/SamplingCampaignID"/>
          <om:value>SnowdoniaNationalParkJapaneseKnotweedSurvey2015_Cleaned</om:value>
        </om:NamedValue>
      </om:parameter>
      <om:observedProperty
xlink:href="https://dyfi.cobwebproject.eu/skos/fallopia_japonica"/>
      <om:featureOfInterest>
        <sams:SF_SpatialSamplingFeature gml:id="sf002">
          <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="sp2">
              <gml:pos srsName="urn:ogc:def:crs:EPSG:6.8:3857">53.1139046729
-3.78766989708</gml:pos>
            </gml:Point>
          </sams:shape>
        </sams:SF_SpatialSamplingFeature>
      </om:featureOfInterest>
```

```
      <om:result>
        <swe:DataRecord>
          <swe:field name="taxon">
            <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
              <swe:value>Bog Cotton</swe:value>
            </swe:Text>
          </swe:field>
          <swe:field name="taxon">
            <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
              <swe:value>Common moss</swe:value>
            </swe:Text>
          </swe:field>
          <swe:field name="taxon">
            <swe:Text definition="http://rs.tdwg.org/dwc/terms/index.htm#Taxon">
              <swe:value>Star moss</swe:value>
            </swe:Text>
          </swe:field>
        </swe:DataRecord>
      </om:result>
    </om:OM_Observation>
  </gml:featureMember>
</gml:FeatureCollection>
```

8.3. Aggregation encoding: Collection of observations with track information

The third example illustrates combined path and occurrences information. The applied pattern is *SpatialSamplingFeature* using a *gml:Curve* for trajectory information. If a corridor instead of the trajectory would be required, a *gml:Polygon* would be used instead of the *gml:Curve*. If the collection would not aggregate observations by the same citizen scientist using the same sensing device, a *gml:Collection* would be used that needs to repeat the various data sets. A full example of such a situation is given in annex XXX.

XML example: Sampling curve option. Here, the location of the observation is part of the result element

```
<om:OM_Observation gml:id="_x3ebvgy65">
  <gml:description>Collection of Japanese Knotweed observations, Dyfie,
Wales</gml:description>
  <gml:name>Observation Collection 2</gml:name>
  <om:type xlink:href="http://www.opengis.net/def/observationType/OGC-
OM/2.0/OM_ComplexObservation"/>
  <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>
```

```
    <tsml:ObservationProcess gml:id="op1-moto">
      <!-- processType defines observation performed by human with sensor -->
      <tsml:processType
xlink:href="http://www.opengis.net/def/waterml/2.0/processType/Sensor"/>
      <!-- processReference defines sampling protocol -->
      <tsml:processReference
xlink:href="https://dyfi.cobwebproject.eu/skos/JapaneseKnotweedSamplingProtocol"/>
      <!-- if a sensor is used, provide the link to the sensor definition here.
Ideayll, value points to SensorML definition -->
      <tsml:parameter>
        <om:NamedValue>
          <om:name xlink:href="http://www.opengis.net/def/property/OGC/0/SensorType"/>
          <om:value>http://www.motorola.com/XT1068</om:value>
        </om:NamedValue>
      </tsml:parameter>
      <!-- operator defines the human producing this observation -->
      <tsml:operator>
        <!-- anonymous observation producer from ISO 19115 -->
        <gmd:CI_ResponsibleParty>
          <gmd:role>
            <gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml"
codeListValue="resourceProvider"/>
          </gmd:role>
        </gmd:CI_ResponsibleParty>
      </tsml:operator>
    </tsml:ObservationProcess>
  </om:procedure>
  <!-- namedParameter to provide the sampling campaign identifier -->
  <om:parameter>
    <om:NamedValue>
      <om:name xlink:href="https://dyfi.cobwebproject.eu/skos/SamplingCampaignID"/>
      <om:value>SnowdoniaNationalParkJapaneseKnotweedSurvey2015_Cleaned</om:value>
    </om:NamedValue>
  </om:parameter>
  <om:observedProperty
xlink:href="https://dyfi.cobwebproject.eu/skos/fallopia_japonica"/>
  <om:featureOfInterest>
    <sams:SF_SpatialSamplingFeature gml:id="ssf1">
      <sf:type xlink:href="http://www.opengis.net/def/samplingFeatureType/OGC-
OM/2.0/SF_SamplingCurve"/>
      <sf:sampledFeature
xlink:href="https://dyfi.cobwebproject.eu/skos/Snowdonia_National_Park"/>
      <sams:shape>
        <gml:Curve gml:id="curve1">
          <gml:segments>
            <gml:LineStringSegment>
              <gml:posList srsName="urn:ogc:def:crs:EPSG:6.8:3857">
                52.0409627 -4.0732345
                52.0410527 -4.0742352
                52.0410612 -4.0752377
```

```
                52.0411411 -4.0762878
                52.0411547 -4.0779545
                52.0412423 -4.0787875
                52.0413144 -4.0795565
                52.0414145 -4.0802447
                52.0414477 -4.0812454
                52.0414797 -4.0821024
              </gml:posList>
            </gml:LineStringSegment>
          </gml:segments>
        </gml:Curve>
      </sams:shape>
    </sams:SF_SpatialSamplingFeature>
  </om:featureOfInterest>
  <om:result>
    <swe:DataArray>
      <swe:elementCount>
        <swe:Count>
          <swe:value>4</swe:value>
        </swe:Count>
      </swe:elementCount>
      <swe:elementType name="occurrence">
        <swe:DataRecord id="occurrenceRecord">
          ①
          <swe:field name="lat">
            <swe:Quantity
definition="http://sweet.jpl.nasa.gov/2.0/spaceCoordinates.owl#Latitude" axisID="Lat">
              <swe:label>Latitude</swe:label>
              <swe:uom xlink:href="deg"/>
            </swe:Quantity>
          </swe:field>
          <swe:field name="lon">
            <swe:Quantity
definition="http://sweet.jpl.nasa.gov/2.0/spaceCoordinates.owl#Longitude" axisID=
"Lon">
              <swe:label>Longitude</swe:label>
              <swe:uom xlink:href="deg"/>
            </swe:Quantity>
          </swe:field>
          <swe:field name="image">
            <swe:Text definition="https://dyfi.cobwebproject.eu/skos/image"/>
          </swe:field>
          <swe:field name="approxPlantHeight">
            <swe:Quantity
definition="https://dyfi.cobwebproject.eu/skos/approxPlantHeight">
              <swe:uom code="m"/>
            </swe:Quantity>
          </swe:field>
        </swe:DataRecord>
      </swe:elementType>
      <swe:encoding>
```

```
        <swe:TextEncoding blockSeparator="&#10;" tokenSeparator=" "
decimalSeparator="."/>
      </swe:encoding>
      <swe:values>
        52.0411411 -4.0762878 https://dyfi.cobwebproject.eu/5141986.jpg 1.5
        52.0414145 -4.0802447 https://dyfi.cobwebproject.eu/12144d1.jpg 2.0
        52.0414797 -4.0821024 https://dyfi.cobwebproject.eu/65dfe43.jpg 1.0
      </swe:values>
    </swe:DataArray>
  </om:result>
</om:OM_Observation>
```

Chapter 9. From Schemas to Data Interoperability Contracts

The use of a known schema, whether it be encoded as XML, JSON or even the classes and properties in an RDF graph, provides for some of the semantics of the data to be understood by the recipient. In general, however, it provides little or no information about a range of key issues that determine if that data can be understood, combined, or even discovered in the context of very large collections.

Standard practice to date has been the provision of documents defining "Application Profiles" - for example the NetCDF [http://www.unidata.ucar.edu/software/netcdf/conventions.html]

The OGC has been using the ISO 19100 approach to defining Application Schema, using an UML idiom that allows data structures to be defined, but offers no standardised way to allow data providers to further refine the rules about the data content contained. The OGC Modular Specification Policy however supports the inheritance of specifications, and the identification of specification elements (conformance classes and requirements) using URI identifiers.

These processes form a backbone for defining interoperability contracts around data structures and service interfaces. Defining interoperability requirements of data within these constraints needs to be supported by additional mechanisms.

We can summarise the design goals for such mechanisms:

- reduced transaction costs
- improve interoperability between components
- improve ease of use for non-experts
- allow more auto-configuration

reduced transaction costs : Implementing interface standards requires development of software. Implementing schema standards requires mapping data structures into the internal data models of consuming systems. These have high transaction costs, borne by the experts building systems, to reduce the transaction cost of users of that software accessing such data.

Interpretation of the data, and semantic translations, remain the problem of the end user, who relies on metadata or prior knowledge of data to discover, formulate appropriate requests, transform and exploit the data. With data descriptions available in large documents, at best, or relatively terse dataset metadata, identify a basis for data integration remains a high transaction cost. Declarative machine readable statements about the conformance of data to one or more interoperability contracts reduce the transaction costs compared to interpretation of unstructured, relatively ad-hoc documents or descriptive metadata/

The goal is therefore to allow data publishers to make statements about data content that can reduce the burden on the end user to discover and exploit data, through both unambiguous identification of conformance with interoperability contracts, and making as much of the aspects of those contracts machine-readable.

improve interoperability between components : It is recognised that similar data may be structured and packaged in many different ways, yet share common elements. For example, a record of Sea Surface Temperature may be held as a gridded coverage over an area, or may be present in discrete samples combined with Salinity readings. A set of readings at different locations may be available for a time period, or the same type of data may be available as a timeseries at a given location. In the case of Citizen Science activities, it will be important to be able to distinguish different methodologies used to collect the same type of data, including "official data". The requirement is thus to be able to make statements about aspects of the data, and where that aspect is expressed in data encodings. Each aspect will this need its own identifier, and be related to the broader data set description as well as further parameters about how that aspect is implemented. The useful implication is that this will allow partial description of datasets, with key aspects being documented with declariative semantics, whilst allowing less important aspects, or harder to describe, to be documented in an ad-hoc fashion.

improve ease of use for non-experts Non experts need explicit statements regarding the semantic compatibility of data for a given purpose, including simple comparability of data. The alternative is locating, accessing, reading and comparing, and ultimately citing, potentially detailed and inconsistently structured descriptive documentation. The design goal is for interoperability contracts to be constructed of components that can be immediately tested for comparability, using identifiers that can be de-referenced to immediately access relevant sections of documentation.

allow more auto-configuration Data interoperability is enhanced by enforced compliance during data collection, therefore the goal is to provide sufficient guidance to data collection software configuration to automate much of that compliance and user-assistance. Auto-configuration of data integration processes needs to be supported, partially or wholly, by unambiguous machine-readable metadata of data, preferably carried from the data collection process. Conformance to structural and semantic contracts then allows auto-configuration of data utlisation through re-use of configurations for display or model assimilation.

9.1. Current situation in SDIs

It is noted that current Spatial Data Infrastructures based on "Publish-Find-Bind" paradigm using catalogues of static metadata records seem to lose performace as the number and heterogeneity of data sets grows. Furthermore, with increased numbers and backgrounds of end-users (such as a broad Citizen Science community) there must be expected a lower level of familiarity of users with the descriptive conventions. Scientists working in a narrow field may be expected to know the code name of a particular instrument (e.g. MODIS Version 6 products to AρρEEARS: MCD43A1, MCD43A3, and MCD43A4 [https://lpdaac.usgs.gov/about/news_archive/release_appeears_version_12]).

We note the following areas where current practices are fragile in multi-stakeholder contexts:

- keywords for discovery
- identification of observed property and methodology
- use of naming conventions in dataset/layer etc names to convey semantic meaning
- data models - identification of , declaration of, descriptions of, relationships between
- consistency of profile and data product specification documents

- discovery of datasets containing information about a particular feature
- lack of easily discoverable links between related data elements

9.2. Current situation in "domain standards"

Currently "communities of practice" (COP) emerge through various fora and try to address their interoperability requirements. OGC has formalised such a process whereby "Domain Working Groups" can be established, and then work within the OGC framework to generate specifications, which are then vetted for consistency with similar approaches by other domains.

Domains with stakeholders willing and able to take the "long view" may thus standardise data models and service interfaces for interoperability. Applying such standards in the wider community is done by a much broader community, on shorter timescales. Such short term demands mean the payoff for developing standards is ahrd to realise, and the value of conforming to a given standard/COP requirement must be easily understood and realised.

COPs also emerge out of technical sub-groups from within existing cooperations with the domain. such groups develop "fit-for-purpose" but idiosyncratics APIs and data models. (e.g. GBIF)

Some COPs are created by design, through projects and programmes targetting cooperation, such as the GEOSS system, or the COBWEB project. They may be infrastructure oriented, or "network building" attempts. Participation requires conformance to a specification provided by a controlling interest. Typically the aim is that such COP may grow into "opt-in" models embracing a wider audience than the initial participants.

Finally, many COP emerge through common experiences applying common tools to a problem space. User groups for particular toolsets may simply share experiences and resources, and de facto standards emerge.

In the case of complex subject domains, such as Citizen Science, Earth Observation, Urban Design, it is likely that all these models of COP will co-exist. What is missing however is a well-known means for each COP to share its particular concerns in ways which can be combined, compared or even discovered.

Chapter 10. Improving the status quo

We must recognise that effective COP and standards are not going to "go away" - and that leveraging multiple heterogenous approaches has advantages for both legacy system integration, and flexibility to optimise future system design.

Secondly, we must recognise that for each system (or COP) some aspects will be unique but many will be common between COPs. Thus, *granularity* of requirements specification must be a driving principle. In fact, this is the main shortcoming of the status quo for both SDIs and standards development.

Thirdly, recognising that the same data can be packaged, transferred and access using different technologies, but still conform to an underlying semantics suggests that technical standards need to be applied to data standards, rather than the converse - where each technical standard (schema or interface) needs multiple independent specifications of the data content.

At this point we can note that the trend to separating the "conceptual model" from schema encodings in the OGC standards process is addressing this concern. In addition there is an emerging supporting infrastructure of the OGC Modular Specifications Policy - and publishing components of specifications (conformance classes and requirements) as indivual Web addressible components.

If we then examine, for example, the use of SWE schemas for Citizen Science, we can see that the OGC process works well to a point, at which we start to need to tie data specifications into specific schema elements, and we find ourselves with multiple possible schemas, and no standard way to define the commonality of data elements between these.

The question then is whether an approach to defininining data-centric requirements can be "bound" to multiple alternative technical standards, working in a lightweight process suitable for the data design lifecycle, not the software and technical standards lifecycles.

10.1. Data Cube Approach

Interoperability dimensions:

- geographical dimension
- time
- thematic
- data model
- ...?

10.2. Future SDI Situation

aspects such as:

- read RDF QB dimensions to understand what vocabularies to query
- query catalog to get the URI template structures for a given vocabulary (or linked data entries?)

- interact with vocabulary to get relationships of query terms and other resources
- data access with content negotiation

Appendix A: Annex A - XML Examples

A.1Introduction

This annex provides complete XML examples covering the following sampling situations:

- Bogs in Wales: Citizen scientists observe bogs in Western Wales.
  - Individual citizen scientist
  - Collection of citizen scientist surveys
    - uncompressed version with annotations (1.6MB)
    - compressed version (43kb)
- Trees Survey: Citizen scientists observe trees in

Appendix B: Revision History

Table 2. Revision History

Date	Release	Editor	Primary clauses modified	Descriptions
Aug 10, 2016	0.1	I. Simonis	all	initial version
Aug 20, 2016	0.2	I. Simonis	chap 7-9	complemented

Appendix C: Bibliography

- tbd