Jira GBIF-13: Demonstration of publishing of sample-based data through the GBIF network

Due: 31 Dec 2014 (some activities proposed for 2015 implementation)

Lead: Éamonn Ó Tuama

Team: Markus Döring, Kyle Braak, Tim Robertson, Olaf Bánki

Planning 2015 campaign: Olaf Bánki, Mélianie Raymond, Alberto González Talaván

# Achievements

The following items relating to JIRA item GBIF-13 (sample data) have been completed as of October 2014:

1. Capturing requirements: extensive consultation with community on how to capture essential information on sample data in Darwin Core, including GBIF organised workshop (May 2013) the TDWG and EU BON mailing lists, the training event at the GEO BON AGM (Crete, April 2014) and the EU BON – GEO BON workshop on Essential Biodiversity Variables (Leipzig, Oct 2014.
2. Identification and definition of five new vocabulary terms required for encoding sample data: sampleSize, sampleSizeUnit, quantity, quantityType, parentEventID.
3. Submission of the five new terms for consideration for ratification in the Darwin Core; this involved creating five separate issues in the GitHub tracker and formally announcing the request for change on the TDWG Content and TDWG TAG mailing lists.
   1. <https://github.com/tdwg/dwc/issues/9>
   2. <https://github.com/tdwg/dwc/issues/10>
   3. <https://github.com/tdwg/dwc/issues/11>
   4. <https://github.com/tdwg/dwc/issues/12>
   5. <https://github.com/tdwg/dwc/issues/13>
4. Definition of a controlled vocabulary (unit\_of\_measurement.xml) for the permitted values of sampleSizeUnit[[1]](#footnote-1).
5. Encoding sample data in a Darwin Core Archive by creation of a new Event core[[2]](#footnote-2) and an Occurrence extension[[3]](#footnote-3)
6. Development of an IPT instance[[4]](#footnote-4) for testing publication of sample data.
7. Six proof-of-concept sample data sets published using the IPT together with metadata (in collaboration with EU BON partners):
   1. Rhine Main aquatic invertebrates
   2. Rhine Main macrophytes biodiversity
   3. Reef-Life survey of reef fishes
   4. Gialova lagoon brackish water survey
   5. Lepidoptera samples from Hannu Saarenmaa
   6. Butterflies monitoring scheme in Israel
8. Organisation of a symposium[[5]](#footnote-5) on sample data standards at the TDWG 2014 conference.
9. Sample based data plans were promoted at each of the GBIF regional nodes meetings thus ensuring that the GBIF community is aware of these activities.

# Sample data in GBIF portal

The main outstanding item relating to Jira GBIF-13 is to demonstrate indexing, discovery and access to sample data sets via the GBIF portal. Once a number of dependencies are satisfied, it is expected that this will proceed in two phases (I and II). These are outlined below.

## Dependencies

Demonstrating indexing, discovery and access is dependent on the following conditions:

1. The definitions for Event core and Occurrence extension are stable.
2. Any new vocabulary terms used in Event or Occurrence tables are “published” with stable identifiers (URIs).
3. A stable version of the unit\_of\_measurement vocabulary is available.
4. The IPT is capable of handling the Event core.
5. Bone fide (not proof-of-concept) sample based datasets are available from GBIF nodes and published using the extended Darwin Core standard.

## Adapting the GBIF portal for sample data

Enabling the indexing, discovery and access to sample-based datasets will be achieved through two phases of implementation (I and II). A third phase of deployment will further the work of sample data mobilization with full involvement of the Participant Nodes and with the involvement of the wider community possible.

2015

2016

2015

2014

**GB22**

Training for nodes?

**Phase I**

* Finish planning
* Identify documentation requirements

**GBIF.org**allows discovery and download

**Phase II**

* Update/produce documentation
* Create interest group(s)
* Promote approach
* Test mobilization
* Refine standard

**EU BON**Training in Finland

3 months

**Deployment**

* Promote beyond already convinced groups
* Include in Capacity Enhancement support programme

**Phase I**: Support basic registration of datasets (expected by early 2015)

1. Registry supports “sample” type;
2. Dataset search supports filtered search to “sample” based data sets only;
3. Searching of metadata of sample data sets is supported;
4. Extension records depicting occurrences are indexed and available for search through GBIF.org and the API as per existing occurrence indexing processes.

**Phase II**: enhanced indexing of sample records, and discovery of datasets (2nd half 2015)

1. The GBIF portal is adapted to provide filtering of data sets based on criteria related to sample data (e.g. sampling protocol, event ID, quantity type). Example use cases include:
   1. Filter search to “sample” data only.
   2. Find data sets that use the same sampling methodology (protocol) - which sampling events are comparable across time and space?
   3. Check if individual record is part of a sample data set - recognize occurrence records from a single sampling event.
   4. Find data sets that are members of a series (e.g. a long term monitoring series).
   5. Filter data on the quantity type (e.g., individuals, biomass, biovolume, or a scale such as Braun Blanquet, Domin, etc).
   6. Support arbitrary mapping of sample data.
   7. In response to a download request, the GBIF portal provides access to the full original records in the sample data sets.
2. Where feasible, the data set metadata will be annotated with keywords, to improve discovery of data sets. For example, data sets with distinct eventID elements on the records warrant the eventID being present as a keyword on the metadata to aid data set discovery.

**Phase III: Deployment.** Adoption by the community. Nodes mobilize sample data as one of the data types enabled by GBIF.

## Uptake and testing

Once a stable technical infrastructure is in place (by early 2015), the wider GBIF community can be approached to inaugurate the use of the system for publishing of bone fide sample data sets. This process can include:

1. Running a campaign with GBIF nodes to publicise and encourage uptake.
2. Aligning with requirements of the SEP2D project relating to sample data.
3. Aligning with requirements of EU BON relating to sample data, in particular Work Package 5 (testing and validation of concepts, tools and services at EU BON test sites)
4. Collaborating with the GEO BON office on mobilising time-series (sample) data.

# Milestones to end 2014

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Item** | **Deadline** | **Agents** |
| 1 | Provide stable identifiers (URIs) for the new vocabulary terms either as additions to Darwin Core or using the GBIF namespace (e.g., <http://rs.gbif.org/terms/gbif/sampleSize>). This is dependent on the TDWG ratification procedures. | Dependent on DwC ratification procedure – earliest is ca end Feb 2015 | EOT; MD, TR |
| 2 | Define a controlled vocabulary (unit\_of\_measurement.xml) for permitted values of sampleSizeUnit | 31 Dec 2014 | EOT; MD |
| 3 | Review documentation needs for sample data | 31 Dec 2014 | EOT; AGT; KB |
| 4 | Devise plan for running campaign with GBIF community | 31 Dec 2014 | EOT; OB; AGT, MR |
| 5 | Prepare work plan for portal enhancements for sample data | 31 Dec 2014 | EOT; TR |

# Stable infrastructure for sample data

In order to roll out wider testing of the IPT for publishing sample data sets, the following need to be in place:

1. The enhancements to the IPT for sample data as developed in the IPT prototype at <http://eubon-ipt.gbif.org> must be implemented in a new release of the standard IPT by Q1 2015 (due in release version 2.3).
2. The definitions for the new Event core[[6]](#footnote-6) and Occurrence[[7]](#footnote-7) extension must be in a stable state and moved from the sandbox to production.
3. The core and extension definitions, in turn, depend on the five newly proposed terms (sampleSizeValue, sampleSizeUnit, organismQuantity, organismQuantityType, and parentEventID) being stable - meaning each term has a stable, resolvable identifier in the form of a URI.

## Stable URIs for terms

Ideally, each term should first be provided with a stable identifier in the form of a resolvable URI before being widely promoted and taken up by data publishers. In this case, the options are to use the Darwin Core namespace under the assumption that the terms once submitted for inclusion in the Darwin Core vocabulary will be accepted, or, as an alternative, use the GBIF namespace. For example,

|  |  |
| --- | --- |
| **Darwin Core** | **GBIF Terms** |
| http://rs.tdwg.org/dwc/terms/sampleSizeValue | http://rs.gbif.org/terms/gbif/sampleSizeValue |
| http://rs.tdwg.org/dwc/terms/sampleSizeUnit | http://rs.gbif.org/terms/gbif/sampleSizeUnit |
| http://rs.tdwg.org/dwc/terms/organismQuantity | http://rs.gbif.org/terms/gbif/organismQuantity |
| http://rs.tdwg.org/dwc/terms/organismQuantityType | http://rs.gbif.org/terms/gbif/organismQuantityType |
| http://rs.tdwg.org/dwc/terms/parentEventID | http://rs.gbif.org/terms/gbif/parentEventID |

As a matter of priority, and following the feedback and support received at the TDWG conference, the process to seek ratification of the five new terms was initiated on 7th November. This process could be completed by late January 2015, assuming the terms do not prove controversial. In the event of the latter happening, the terms can be registered, instead, under the GBIF namespace. By introducing a versioning system for DwC-A cores and extensions, we allow the possibility for them to evolve. Thus, if the terms are later ratified as part of DwC, a new and revised core and extension can use them.

Update (23/01/2015): the terms went through some revision based on discussion on the TDWG Content list. These changes were consolidated and re-posted to the list on Dec 23rd and, barring further changes after a public review period of 30 days, will be submitted to the TDWG Executive for their decision after Jan 23rd.

## Tasks

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Item** | **Status** | **Due by** |
| 1 | Define Event core and Occurrence extension for DwC-A | completed | Dec 31 2014 |
| 2 | Define unit\_of\_measurement controlled vocabulary | completed | Dec 31 2014 |
| 3 | New release of IPT handles sample data |  | Dec 01 2014 |
| 4 | Submit new terms for ratification in Darwin Core | completed | Dec 01 2014 |
| 4 | Devise versioning system for DwC-A cores and extensions |  | Q3 2015 |
| 5 | Devise versioning system for controlled vocabularies |  | Q3 2015 |
| 6 | Devise framework for publishing GBIF vocabularies |  | Q3 2015 |

# Documentation needs for sample data

In order to promote and support the publication of sample-based data sets, all documentation relating to the IPT and the GBIF EML metadata profile will require updating. There may also be needs for specific training materials on how to use the IPT for publishing sample data.

## Documents

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Document** | **Task** | **Status** |
| 1 | GBIF IPT sample data primer | Describes the new vocabulary terms required for Event core and Occurrence extension and structure of the Darwin Core archive with several examples included | Draft |
| 2 | Use cases | Extend primer with typical use case for sample based data |  |
| 3 | GBIF Metadata Profile, How-to guide | Update with ref to projectID attribute |  |
| 4 | GBIF Metadata Profile, Reference guide | Update with ref to projectID attribute |  |
| 5 | GBIF IPT v. 2 User Manual | Update with ref to the new flexibility around cores and extensions |  |
| 6 | Basic data publishing guides | Update anything that refers to the types of data that can be published through the network and how to do it. |  |
| 7 | “How to publish” schema about sample data | Create resource, similar to the existing ones about metadata, occurrence and checklists. |  |
| 8 | Basic How-to guide for publishing sample based data | Technical description on how to publish this kind of data |  |
| 9 | Documentation for Nodes about common challenges when promoting the mobilization of sample based data | Concentrate in the social and cultural challenges. |  |
| 10 | Basic documentation about how to use sample-based data obtained through GBIF | New data type, new uses. |  |
| 11 | Update GBIF.org | Update of all sections of GBIF.org referring to the types of data that can be published |  |

# A campaign to promote publishing of sample data

Once a stable informatics infrastructure is in place by Dec 2014, GBIF can start the mobilization of datasets by interested parties. 2015 will be used to refine the process, develop strategies to promote the digitization, publication and use of sample based data, and train those in charge of promoting the publication of such data,

After this pilot phase, the deployment phase can start to widely publicise the new developments and further encourage uptake for mobilisation and publication of sample data.

## Working with the GBIF Nodes

The nodes will be involved in each phase as follows:

**Phase I**:

1. Involve highly engaged nodes that want to push this work forward to actively participate in the conceptualization phase by providing input and feedback, example datasets, etc.
2. Discuss the plans with the NSG (end of 2014)
3. Include the topic as one of the recommended options for the GB22 Nodes training.

**Phase II**:

1. Prepare announcement for circulation to Nodes list (Jan 2015) alerting them to the plans
2. Prepare news item (Jan 2015)
3. Create interest group within the nodes community to help in the mobilization of sample based datasets.
4. Promote regional projects, or projects Nodes have an interest in as further pilots for publishing sample-based data and mapping this data to the standard.
5. Start the involvement of international organizations and societies working with sample-based data.
6. Organize workshop at GB22, if approved by the Nodes Committee AND if the modifications to GBIF.org are ready well in advance (min. 3 months).
   1. Invite several Node Managers to participate as trainers
   2. Preparatory meeting to train the trainers
   3. Development of specific training materials in several languages.
7. Document any barriers to publishing sample-based data and capture requirements for use and extraction of data out of the portal. This could actually lead to a more formal mobilization strategy around specific types of sample-based data.

**Phase III**: Deployment

1. Invite all Nodes to support the publishing of sample-based data as one of the official data types supported by GBIF.
2. Include the organization of capacity enhancement support actions around the mobilization of sample-based data.

## Working with the SEP2D project

Promote the use of the IPT within the SEP2D project for publishing sample based data, in particular, forest plots, to the GBIF network.

## Working with EU BON

Work Package 5 takes the concepts, tools and services developed in WP 1 and 2 and tests and validates them using designated EU BON test sites (e.g., the Rhine Main Observatory, an ILTER site). The IPT adapted for publishing sample data is one such tool and this should offer an opportunity to demonstrate its usefulness, allowing GBIF to provide views into, at least, a subsection of the kinds of data managed at ILTER sites.

1. <http://rs.gbif.org/sandbox/vocabulary/gbif/unit_of_measurement.xml> [↑](#footnote-ref-1)
2. <http://rs.gbif.org/sandbox/core/dwc_event.xml> [↑](#footnote-ref-2)
3. <http://rs.gbif.org/sandbox/extension/event_occurrence.xml> [↑](#footnote-ref-3)
4. <http://eubon-ipt.gbif.org> [↑](#footnote-ref-4)
5. <https://mbgserv18.mobot.org/ocs/index.php/tdwg/2014/schedConf/trackPolicies#track6> [↑](#footnote-ref-5)
6. <http://rs.gbif.org/sandbox/core/dwc_event.xml> [↑](#footnote-ref-6)
7. <http://rs.gbif.org/sandbox/extension/event_occurrence.xml> [↑](#footnote-ref-7)