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

# 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, eventSeriesID.
3. Encoding sample data in a Darwin Core Archive by creation of a new Event core[[1]](#footnote-1) and an Occurrence extension[[2]](#footnote-2)
4. Development of an IPT instance[[3]](#footnote-3) for testing publication of sample data.
5. 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
6. Organisation of a symposium[[4]](#footnote-4) on sample data standards at the TDWG 2014 conference.

# 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. A process for achieving this is outlined in the remainder of this document.

## Dependencies

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

1. Bone fide (not proof-of-concept) sample based datasets are available from GBIF nodes for publishing.
2. The definitions for Event core and Occurrence extension are stable.
3. Any new vocabulary terms used in Event or Occurrence tables are “published” with stable identifiers (URIs).

## 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.

**Phase I**: support basic registration of datasets (expected by end 2014, provided datasets available)

1. The GBIF registry supports registration of a new dataset type, e.g., “sample-based”.
2. Extension records depicting occurrences are indexed and available for search through GBIF.org and the API as per existing occurrence indexing processes.
3. Searching metadata of sample data sets is supported.
4. Dataset search (only) supports filtered search to “sample” based data sets only, along with the full text of any provided metadata (already operational)

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

1. The GBIF indexing process is adapted to recognise sample records (e.g. record based indexing) including certain key fields.
2. 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. Sample based occurrence records are identifiable on maps using a distinct filter.
	7. In response to a download request, the GBIF portal provides access to the full original records in the sample data sets.
3. 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.

## Uptake and testing

Once a stable technical infrastructure is in place (by Dec 2014), 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)

# Next steps

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **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>) | 31 Dec 2014 | EOT; MD, TR |
| 2 | Review documentation needs for sample data  | 31 Dec 2014 | EOT; AGT; KB |
| 3 | Devise plan for running campaign with GBIF nodes | 31 Dec 2014 | EOT; OB |
| 4 | Prepare work plan for portal enhancements for sample data | 31 Dec 2014 | EOT; TR |

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