ViBRANT status report

Deliverable D4.2, Ontology Tools

Milestone M4.12, Liaison and networking with ontology experts and existing ontology providers

Milestone M4.14, Define further milestones in the light of usage and feedback

Leading partners: Global Biodiversity Information Facility (GBIF)

Compiled by: Dag Endresen, Éamonn O Tuama and David Remsen

November 2011
Summary

This document provides a summary of the contributions from the Global Biodiversity Information Facility (GBIF) to the Virtual Biodiversity Research and Access Network for Taxonomy (ViBRANT). ViBRANT is a European Union FP7 funded project running December 2010 to 2013 that will support the development of virtual research communities involved in biodiversity science.

The development of new milestones for 2012 is provided with this document.

Introduction

This part describes the efforts made by GBIF to consult with ontology experts and existing ontology providers during the first period of the ViBRANT project. Experts from the biodiversity information community provided advise on the best practice for supporting the use of vocabularies and ontologies in the community.

Related to milestone M4.12 Liaison and networking with ontology experts and existing ontology providers (Deadline July 2011).

The GBIF task group report on metadata implementation recommended the use of controlled vocabularies where the terms would be identified by persistent identifiers (Jones et al., 2010, page 22-23, recommendation R53-R57). “The identifier of the vocabulary should use existing identifiers from other registries where possible. If one does not exist, then GBIF should construct and publish the identifier.” (…) “At present there is no well-defined and consistent means of referencing an identifier of a vocabulary or a vocabulary term. The proposed GBIF registry should provide an unambiguous citation method for each vocabulary and the terms they contain.” (…) “Some vocabularies will be global in use but some will be domain specific. To ensure compatibility across all metadata records, it is important that users use the appropriate and community agreed vocabularies” (Jones et al., 2010, page 23).

The GBIF task group report on persistent identifiers provided a recommendation for GBIF to “take a leadership role in encouraging the use of metadata vocabularies for information in the GBIF data portal and extending the role of the data portal by hosting resources related to the use of identifiers, such as the TDWG vocabularies” (Cryer et al., 2010, page 14, recommendation 12).

The GBIF beginner’s guide to persistent identifiers stated the importance “to reuse, where appropriate, the vocabularies and schemas that other communities have developed, to aid interoperability and save reinventing the wheel.” On the same page this guide also commented: “Because biodiversity informatics is a fairly specialized area of expertise, it is likely that a large proportion of the vocabularies and ontologies required for this domain will need to be developed within this community” (Richards et al., 2011, page 20).
The GBIF task group for Knowledge Organisation Systems (KOS) produced a white paper with a focus on recommendations on the use of vocabularies and ontologies for biodiversity informatics (Catapano et al., 2011).

DCMI 2011 pre-conference on vocabulary management
In September 2011 GBIF represented the biodiversity informatics community at the Dublin Core annual conference for metadata and vocabulary management in Hague, Netherlands. The pre-conference that we attended, addressed the best practices for maintaining a federated knowledge organization system (KOS) with a common vocabulary of terms used by a decentralized network. The most important lesson learnt was that communities of domains other than ours were also just starting to work with very many of the same issues we face. The best practices for maintaining a federated KOS are under development in the Dublin Core (DCMI) community and the early experiences from the biodiversity informatics community will provide important input to this process.

TDWG 2011 KOS Symposium
GBIF organized a special symposium at the BIS TDWG conference in New Orleans, 17-21 Oct 2011 (see abstract in Annex A). A series of presentations introduced recent activities related to KOS in the GBIF work programme, the current status and history of the TDWG vocabularies, and the management of the Darwin Core set of terms. The presentations left little time for discussion around the core issue of identifying key actors and activities that would enable TDWG, GBIF and related communities of practice to develop a road map on how to engage all players in creating a global infrastructure for the development, maintenance and governance of such vocabularies. However some of the issues were considered in the subsequent TDWG Technical Architecture Group (TAG) meeting and, as an outcome, GBIF proposes to create a task group within the
TAG to take this work forward. This will entail submitting a charter outlining the purpose of the task group with expected deliverables and timelines.

**Action:** GBIF, in consultation with community to develop charter for a “terms and vocabularies” management interest group; initial draft charter to be circulated by Dec 31, 2011.
**Background, GBIF KOS systems**

This section provides a background on the KOS systems established to support the GBIF infrastructure including an overview to base further work and proposed milestones on.

*Related to deliverable D4.2 Ontology Tools (November 2012).*

**GBIF Vocabularies Service**

The GBIF Vocabulary Service¹ (Figure 1) was developed by GBIF in collaboration with the Natural History Museum of London and based on the Scratchpads² platform of the European Distributed Institute of Taxonomy (EDIT)³. The Scratchpads are developed on the Drupal platform. The first prototype for the GBIF Vocabularies was ready in 2009. The GBIF Vocabularies Service provides a collaborative environment for the development of domain specific extensions (attribute or element vocabularies) to the standard Darwin Core vocabulary (TDWG, 2009) and the development of controlled value vocabularies for these terms. There is also an identified (long-term) need to provide a similar service to define vocabularies for terms and controlled values for data types other than the Darwin Core Occurrences and Taxon that are relevant to biodiversity informatics (GBIF, 2006, 2011). “In anticipation of the integration and serving of future data types, GBIF will work closely with partners to enable data integration and interoperability across phenotypic, genomic, taxonomic, geospatial and ecosystem domains” (GBIF, 2011, page 7).

The GBIF Vocabularies Service provides a web service interface (API) to access the vocabulary terms and definitions in XML or in tab delimited text format.

---

¹ GBIF Vocabularies, [http://vocabularies.gbif.org](http://vocabularies.gbif.org)
² Scratchpads, [http://scratchpads.eu/](http://scratchpads.eu/)
³ EDIT, [http://e-taxonomy.eu/](http://e-taxonomy.eu/)
**GBIF Resources Registry**

Agreed upon vocabularies need to be registered in the GBIF Resources Registry\(^4\) in order to be available to the GBIF infrastructure including the tools and toolkits such as the GBIF Integrated Publishing Toolkit (IPT)\(^5\) and the GBIF harvesting and indexing toolkit (HIT)\(^6\). The Darwin Core standard defines the basic core terminology for the two core types\(^7\) Occurrence and Taxon. The Darwin Core also defines some other relevant classes or data units/entities/types that are registered in the GBIF Resources Registry as “Extensions” to the core types\(^8\). In addition some other vocabularies for basic data types have been defined by GBIF\(^9\). And even more vocabularies for new data types such as multimedia\(^10\), EOL

---

\(^4\) GBIF Resources Registry, [http://rs.gbif.org](http://rs.gbif.org)


\(^7\) Darwin Core, core types, [http://rs.gbif.org/core/](http://rs.gbif.org/core/)

\(^8\) Data types defined by DwC, [http://rs.gbif.org/extension/dwc/](http://rs.gbif.org/extension/dwc/)

\(^9\) Data types defined by GBIF, [http://rs.gbif.org/extension/gbif/1.0/](http://rs.gbif.org/extension/gbif/1.0/)
data object\textsuperscript{11}, invasive species status\textsuperscript{12} and plant genetic resources\textsuperscript{13} have been developed and included in the GBIF Resources Registry.

Finalized approved controlled value vocabularies\textsuperscript{14} for many of the Darwin Core terms (including terms in extensions) are also registered in the GBIF Resources Registry. External controlled value vocabularies from other domains such as the ISO 3166-1 country codes\textsuperscript{15} are also registered here for accessibility and ease of use in the GBIF infrastructure.

It should be noted that the GBIF Secretariat has to actively maintain the GBIF Resources Registry. There are (at the present) no automatic procedures to publish new extensions and vocabularies developed using the GBIF Vocabularies Service to the GBIF Resources Registry. GBIF participants and partners are required (for the moment) to make contact with the GBIF Secretariat for new vocabularies to be added to the GBIF Resources Registry.

The GBIF Resources Registry provides a web service interface (API) to access the vocabulary terms and definitions as XML, and a human user interface using style sheets (xsl and css) to provide the XML definitions in a more accessible presentation.

\textsuperscript{10} Audubon (multimedia), \url{http://rs.gbif.org/sandbox/extension/audubon.xml}
\textsuperscript{11} EOL data object, \url{http://rs.gbif.org/sandbox/extension/eol_dataobject.xml}
\textsuperscript{12} GISIN Invasive species status, \url{http://rs.gbif.org/sandbox/extension/gisin_2.0_species_status.xml}
\textsuperscript{13} Data types for plant genetic resources, \url{http://rs.gbif.org/extension/nordgen/0.1/germplasm.xml}
\textsuperscript{14} Controlled value vocabularies, \url{http://rs.gbif.org/vocabulary/}
\textsuperscript{15} ISO 3166-1, country codes, \url{http://rs.gbif.org/vocabulary/iso/3166-1_alpha2.xml}
Planned further developments at GBIF

This section provides further developments planned for the GBIF KOS work program and provides the background and explanation for the new milestones.

Related to milestone M4.14 Define further milestones in the light of usage and feedback (Deadline November 2011).

GBIF Glossary of Terms used in Biodiversity Informatics

GBIF has published a glossary of terms used in bionomenclature (Hawksworth, 2010) and has, in collaboration with SilverBiology\(^\text{16}\), developed a web application for improved access to this glossary\(^\text{17}\). The source code for this software application is available from a Google Code repository\(^\text{18}\). A new prototype interface to the GBIF Resources Registry was developed based on the same software as developed for the Bionomenclature browser (Figure 2). This new interface\(^\text{19}\) presents the schemas defined and made available to the GBIF infrastructure.

---

\(^{16}\) SilverBiology LLC, [http://www.silverbiology.com/](http://www.silverbiology.com/)

\(^{17}\) Terms used in bionomenclature, [http://bionomenclature-glossary.gbif.org/](http://bionomenclature-glossary.gbif.org/)


\(^{19}\) GBIF Resources Browser, [http://tools.gbif.org/resource-browser/](http://tools.gbif.org/resource-browser/)
GBIF plans to provide a similar service for the presentation of the basic biodiversity terms used by the GBIF infrastructure. This service would provide a flat glossary of terms including all terms from controlled vocabularies. The glossary of terms would include the definitions and persistent identifiers (PIDs) for all data categories such as classes/groups, attribute/element terms and controlled values or code lists in use by various parts of the GBIF infrastructure. Each term would be included once and identified by an existing persistent identifier (PID) such as, for example, an URI, LSID or DOI. If no such existing persistent identifier is available for a term, one will be created for the term in question. If the existing persistent identifier is resolvable then the definition and labels for different languages can be automatically retrieved and cached by the (new) GBIF Glossary of Terms.

This new GBIF Glossary of Terms will provide a backbone for the vocabularies and eventual domain ontologies developed within the GBIF network. GBIF partners will be urged to provide the definitions for all new terms to this GBIF Glossary. The GBIF Secretariat will also maintain the glossary to include missing terms when such terms are identified. The GBIF Glossary of Terms will be provided with an API to allow programmatic access to the definition and labels in use for a term.

The GBIF Glossary of Terms will not provide or define the context of any term such as the class a term could belong to or any one specific vocabulary that the term might belong to. Also the GBIF Glossary of Terms will not make explicit if an individual term is a class definition, a controlled element/attribute term or a controlled value term. The same term from the GBIF Glossary of Terms could thus be used as a class, a controlled element/attribute term, or as a controlled value by different vocabularies and ontologies. A feature of the system to tag the terms should, however, be included. Tagging a term should not be seen as normative or limiting for the context that the term could be used in but rather aid in discovery of the term.

The initial scoping document for this Glossary service will be completed by the end of December and made available to the GBIF community for a first round of public review.

**GBIF Vocabularies Service**

The GBIF Vocabularies Service\(^\text{20}\) is at the present running from the Scratchpads servers hosted by the Natural History Museum in London. The Vocabularies Service has recently been successfully replicated to development servers hosted at the GBIF Secretariat in Copenhagen with the assistance from the Scratchpads team. The objective is, during November, to finalize the migration of the GBIF Vocabularies Service from the servers in London to new servers in Copenhagen. This migration to GBIF servers is motivated by two main reasons. The response time and the performance of the GBIF Vocabulary Service is sub-optimal when

\(^20\) GBIF Vocabularies, [http://vocabularies.gbif.org/](http://vocabularies.gbif.org/)
hosted from the same platform as the other Scratchpads. This is because of the sharing of server resources such as system memory and processing capacity between the Scratchpads. The second reason is the need for some (minor) further developments of features for the GBIF Vocabularies Service.

**NCBO BioPortal platform**

The GBIF KOS task group recommended “**GBIF should deploy an instance of the BioPortal platform for biodiversity ontologies as a complement to the GBIF Vocabularies Server**” (Catapano et al., 2011, page 32, recommendation 4). The BioPortal provides a platform to publish and share community developed vocabularies and ontologies. The BioPortal does not provide a platform to author or modify such vocabularies and ontologies. Using the BioPortal platform users can provide and discuss the mapping and relationships between terms used in different vocabularies and ontologies. The present GBIF infrastructure deals with DwC Occurrence data and with DwC Taxon data. The BioPortal will provide a platform to expand to **new data types**. This expansion would be community driven and would not guarantee the uptake of all new data types by other parts of the GBIF infrastructure. The implementation of a BioPortal instance for biodiversity informatics will also support the integration of terminology from other communities, in particular the Genomics Standards Consortium (GSC).

---

New Milestones for 2012

*Milestone M4.14 Define further milestones in the light of usage and feedback (Deadline November 2011).*

M4.2.a Ontology Tools
Migrate the GBIF Vocabularies Service from the Scratchpads server hosted by the Natural History Museum in London to new servers hosted by the GBIF Secretariat in Copenhagen for improved performance.
**Deadline:** 31 November 2011.

M4.2.b Ontology Tools
Prepare scoping document on a KOS architecture for the GBIF network including the requirements for a new ontology tool to provide a new GBIF Glossary of Terms used in Biodiversity Informatics.
**Deadline:** First draft scoping document finalized by 31 December 2011.

M4.2.c Ontology Tools
Develop a new prototype (GBIF) Glossary of Terms registry for terms used by the ViBRANT and GBIF infrastructure.
**Deadline:** April 2012

M4.2.d Ontology Tools
Present a new prototype Biodiversity BioPortal based on the NCBO BioPortal. This platform will be used for hosting vocabularies and ontologies for new data types (in addition to the DwC Occurrence and DwC Taxon).
**Deadline:** June 2012
References


Annex A TDWG Conference 2011 KOS Symposium Abstract

Convenors: Éamonn Ó Tuama, Dag Terje Filip Endresen, David Remsen

Title: Establishing a support infrastructure for Knowledge Organisation Systems (KOS) in biodiversity informatics

Presentations:

1. Introduction: the GBIF context (Éamonn Ó Tuama)
2. GBIF KOS task group report (Bob Morris)
3. The GBIF KOS work programme: prioritised requirements and proposed solutions (Dag Endresen)
4. BioPortal as an ontology management system (Hilmar Lapp)
5. The TDWG Vocabularies (Greg Whitbread)
6. Managing the Darwin Core set of terms (John Wieczorek)

Abstract Body:
Vocabularies and ontologies are types of Knowledge Organisation System (KOS) that range from simple glossaries and dictionaries through classification schemes, taxonomies, thesauri and ontologies [1]. As such, they are an essential foundation for data interoperability within the biodiversity informatics domain. Recent TDWG conferences have acknowledged the need to adopt principles and best practices for developing ontologies and to draw, wherever possible, on available community resources [2]. GBIF also recognises this need by including KOS related activities in the GBIF Strategic Plan 2012-2016 [3] and in its ongoing work programmes. Key activities to date have been the establishment of the prototype GBIF Vocabularies Service [4] and the publication of a white paper Recommendations for the use of Knowledge Organisation Systems by GBIF [5].

The focus of this symposium is on identifying key actors and activities that will enable TDWG, GBIF and related communities of practice to develop a road map on how to engage all players in creating a global infrastructure for the development, maintenance and governance of such vocabularies (the term “vocabularies” is used here in a generic sense to represent the various types of KOS). Several brief presentations will set the background for our discussions. After introducing the broader GBIF context, the recommendations of the GBIF KOS task group and their uptake in the GBIF work programme will be outlined. Two systems for managing vocabularies and ontologies will then be introduced – the GBIF Vocabularies Service [4] and the BioPortal [6]. This will be followed by a summary of the history, management and plans for the TDWG vocabularies which form the foundation for KOS in biodiversity informatics, and conclude with an account of experiences in managing the Darwin Core as an example of a widely used set of terms under active discussion in the community.


