ViBRANT status report (M4.20)

Deliverable D4.2, Ontology Tools

M4.20 - Ontology Tools: new prototype (GBIF) Glossary of Terms registry

Develop a new prototype (GBIF) Glossary of Terms registry for terms used by the ViBRANT and GBIF infrastructure (deadline: 31 April 2012).

Leading partner: Global Biodiversity Information Facility (GBIF)

Compiled by: Dag Endresen, Éamonn Ó Tuama, Burke Chih-Jen Ko, and Gregor Hagedorn (JKI)

April 2012

Summary
This report provides the progress on developing a new information system for exploring terms and concepts declared by the relevant KOS (knowledge organization system) information resources such as the RDF Vocabularies, Darwin Core extensions and controlled value vocabularies in relation to the ViBRANT milestone M4.20. The Virtual Biodiversity Research and Access Network for Taxonomy (ViBRANT) is a European Union FP7 funded project running from December 2010 to 2013 that will support the development of virtual research communities involved in biodiversity science.

Keywords: Controlled value vocabulary; Darwin Core; Darwin Core archive; Darwin Core extension; Darwin Core vocabulary; Knowledge Organization System (KOS); Resource Description Framework (RDF); RDF Vocabulary; Term.

Introduction
The GBIF (Global Biodiversity Information Facility) data-sharing infrastructure is largely based on sharing datasets described by the Darwin Core[1] data standard (TDWG 2009a; Wieczorek et al. 2012). This means that biodiversity datasets are mapped to terms included in Darwin Core, but it does not mean that concepts in the source dataset without a match to any of the terms in the Darwin Core standard cannot be shared. A mechanism for Darwin Core extensions was established to describe such concepts in the source datasets that did not match the Darwin Core terms.

For the purpose of achieving efficient sharing of biodiversity datasets the *Darwin Core archive*[2] (Döring et al. 2011) data exchange format was developed based on the *Darwin Core text guidelines*[3] (TDWG 2009b). The Darwin Core archive is designed as a star schema with extensions hanging off a core entity. Two core types have been developed, one core type for sharing information on taxa (such as species checklists) and another core type for sharing information on occurrences (such as museum specimens). Each of the core types has properties declared that include the relevant subset of terms from the Darwin Core standard. In context of the Darwin Core archives, the *Darwin Core extensions* were designed to provide information for other data entities than taxa and occurrences. The entities of Darwin Core extensions are linked in a one-to-one or one-to-many relation to the core entities. The rationale for Darwin Core extensions were partly as a mechanism to include (in Darwin Core archives) data properties mapped to terms from other data standards than the Darwin Core

[1] http://rs.tdwg.org/dwc/
[2] http://en.wikipedia.org/wiki/Darwin_Core_Archive
[3] http://rs.tdwg.org/dwc/terms/guides/text/index.htm

standard. For the data properties listed by Darwin Core extensions, the recommendation is to reuse the URIs for terms declared as published and machine-readable (RDF-based) data standards (RDF Vocabularies).

However, many of the relevant terms and concepts are not yet defined by established and published data standards, and most of the existing data standards in the biodiversity community are not (yet) based on the RDF technology for appropriate machine-readable implementations. Therefore, within the scope of the ViBRANT project, efforts been initiated to promote the expression of existing terms and concepts in use by the biodiversity community as RDF vocabulary resources (Figure 1). We suggest that all relevant and finalized RDF vocabulary resources, and in particular those used when building new Darwin Core extensions and controlled value vocabularies, will be registered and that a copy is deposited at the GBIF Resources Repository. The GBIF Resources Repository would therefore include terms that are ready to use and recommended for reuse. Developers of new Darwin Core extensions and controlled value vocabularies are advised to be careful when including terms not (yet) available from the GBIF Resources Repository (Figure 2).

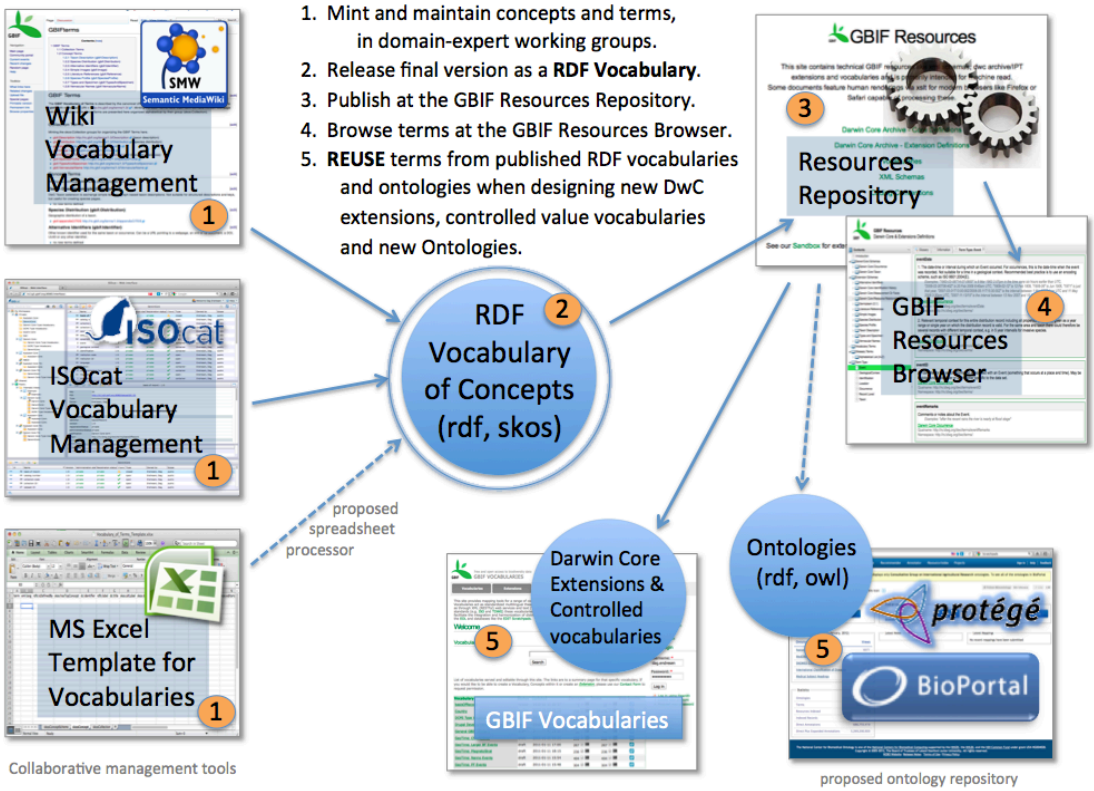

Figure 1: Overview of the workflow for management and uptake of RDF vocabularies of basic terms and concepts.

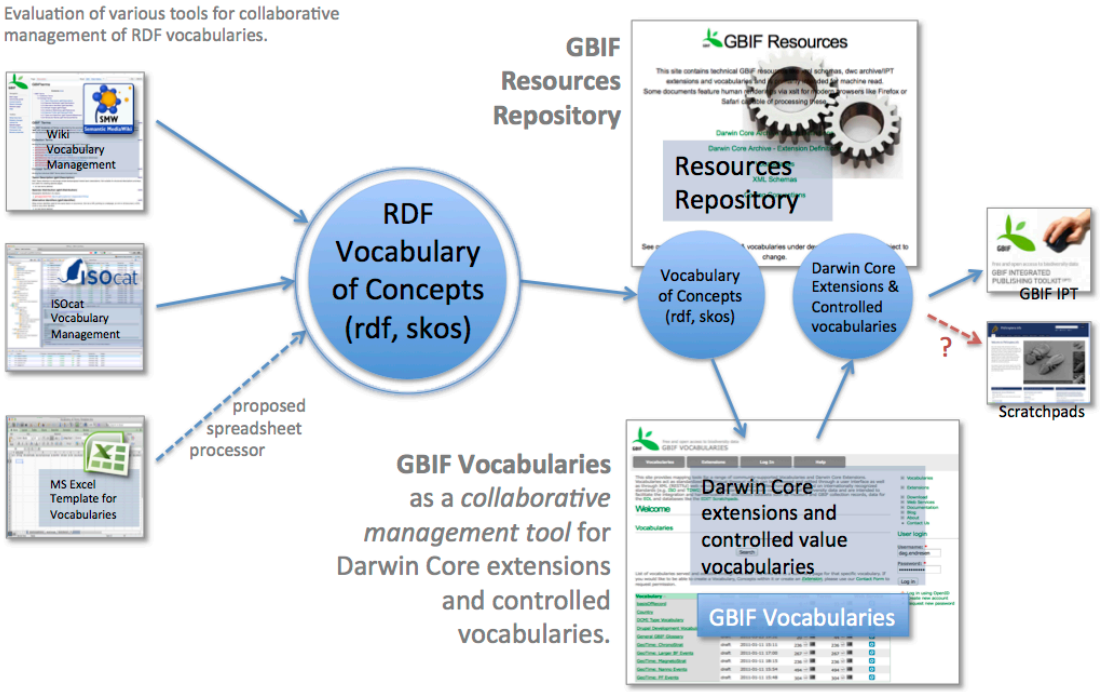

Figure 2: GBIF Vocabularies Server as a tool for the collaborative management of Darwin Core extensions and controlled value vocabularies.

As these RDF vocabulary resources are developed, the users and data publishers will need a user-friendly service to discover recommended terms to avoid duplicated definitions for the same concept and to maximize the reuse of the existing RDF vocabulary terms when new Darwin Core extensions and controlled value vocabularies are developed. For this purpose we describe here two different approaches for providing user-friendly discovery and access to information about the vocabulary terms and concepts. The first proposed service (1) is a community-driven description of terms using a Semantic Wiki. The second service (2) provides a human interface to the terms published at the GBIF Resources Repository[4].

Community-driven semantic wiki (BioVocabulary)

The semantic media wiki platform provides a user-friendly tool for collaborative development and presentation of KOS resources. The semantic wiki interface is easy to use and is therefore likely to lower the barrier for domain experts without experience with KOS management software tools to contribute. The Species-ID[5] platform hosted by the biowikifarm[6] was established as a semantic wiki in the context of the ViBRANT project. A prototype demo instance of the

[4] http://rs.gbif.org/terms/
[5] http://species-id.net/wiki/
[6] http://biowikifarm.net/meta/

semantic wiki was also installed at http://kos.gbif.org/wiki/ for testing purposes. A permanent community-driven "Terminology Wiki" based on the semantic wiki platform could be developed based either on the GBIF demo site or as a new semantic wiki on the biowikifarm platform. The biowikifarm has a long-term strategy with a technical service supported by the SNSB (Munich, Germany) and the BGBM (Berlin, Germany), two German museum IT-centers. The Species-ID and the prototype demo hosted by GBIF provide the first prototype examples of how a community-driven wiki forum for biodiversity terms could be implemented (Figure 3). The semantic wiki software is easy to install and configure. The challenge is to find a solution for long-term persistence of the web site and to attract the user community to use, update and maintain the content of the resources. A wiki forum for terms is well suited as a platform for the community to discuss the meaning and use of existing terms as well as the need for new (or adapted) terms. New terms can be proposed for discussion as easily as creating a new wiki page, and experienced users might join in to suggest that any proposed new terms are overlapping with an existing term.

GBIF Resources Browser

The resources published at the GBIF Resources Repository are primarily made available as an API for machines. However, all resources are styled using cascading style sheet (CSS) and XML transformations (XSLT) to provide a human readable version. Machines and humans can therefore explore the terms recommended by the GBIF Resources Repository directly from the repository site. To improve on the human access interface a searchable Resources Browser[7] service was established in 2010 in collaboration with SilverBiology[8] (Figure 4). The source code for this software application is available from a Google Code repository[9]. When the first RDF vocabularies for biodiversity information resources are developed and deposited to the GBIF Resources Repository, the Resources Browser should list these resources in the search and browse interface for this service (Figure 3). The services provided from the GBIF Resources Repository and the GBIF Resources Browser will be limited to present information about the terms included in RDF vocabularies, Darwin Core extensions and controlled value vocabularies that are published to the GBIF Resources Repository. The information provided from these services is for technology developers (and as machine readable services) only, while the information presented at the wiki forum will be more aimed at other users.

[7] GBIF Resources Browser, http://tools.gbif.org/resource-browser/
[8] SilverBiology LLC, http://www.silverbiology.com/
[9] Google Code, http://code.google.com/p/terms-of-bionomenclature/

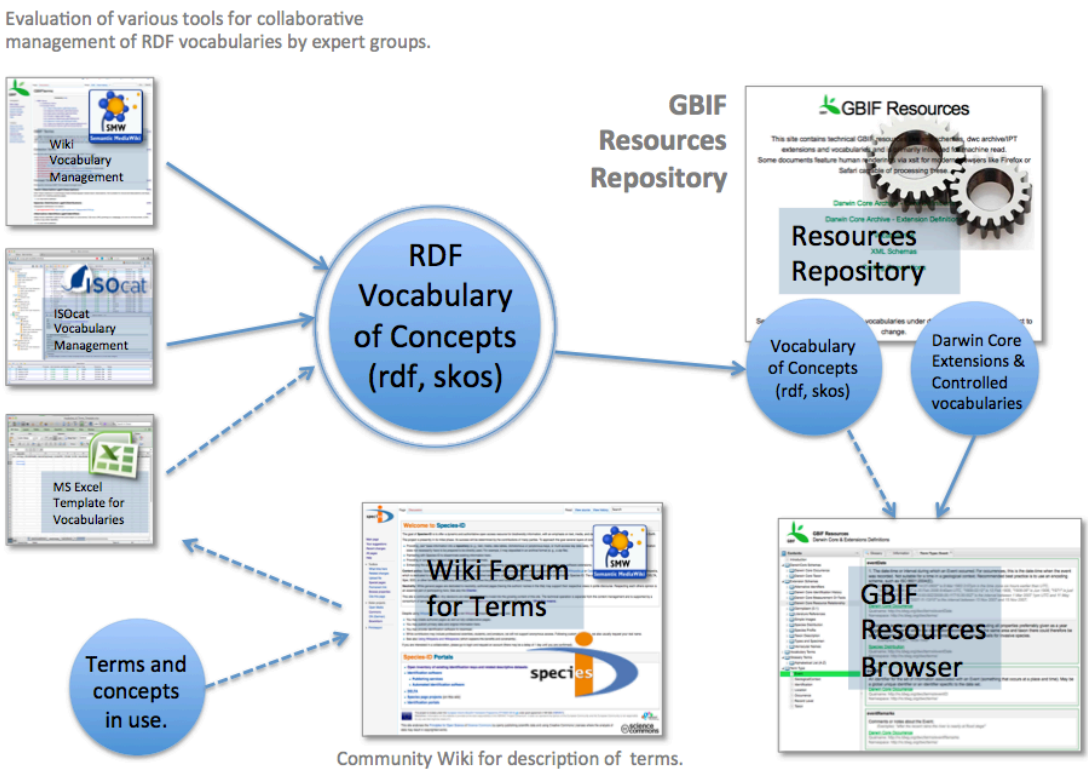

Figure 3: A new community-driven Wiki Forum for Terms implemented as a semantic wiki. The Species-ID wiki (http://species-id.net/) provides a prototype demo example, but the final place may be an independent wiki instance, either on the biowikifarm or at GBIF. The GBIF Resources Browser (http://tools.gbif.org/resource-browser/) provides more technical information and machine-readable services for terms and concepts from the RDF vocabularies, Darwin Core extensions and controlled value vocabularies that are published at the GBIF Resources Repository (http://rs.gbif.org/).

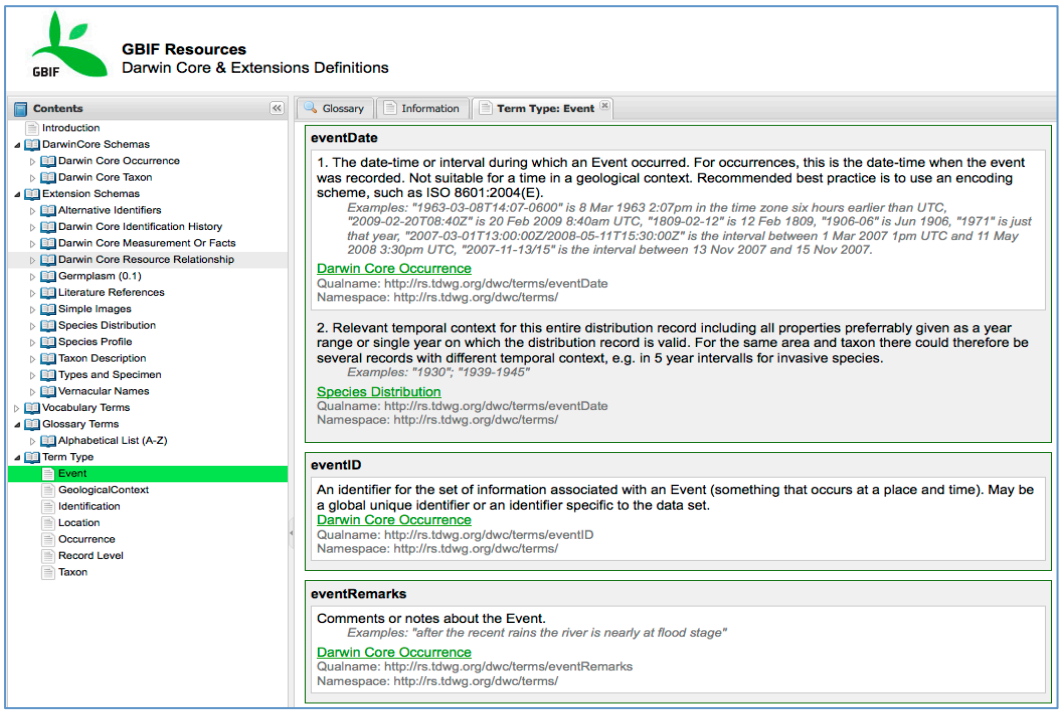

Figure 4: GBIF Resources Browser (http://tools.gbif.org/resource-browser/), for Darwin Core extensions and controlled value vocabularies registered at the GBIF Resources Registry (http://rs.gbif.org/).

Conclusions

The first prototype solutions for providing human-centric information and discussion forum for biodiversity vocabulary terms are presented at http://species-id.net and at http://kos.gbif.org/wiki/. Further work will focus on identifying the most appropriate web hosting implementation to ensure the long-term persistence of such a service. The biowikifarm web environment at http://biowikifarm.net seems to be the most appropriate alternative. We propose to install the technical backbone implementation for the recommended and published machine-readable vocabulary resources (RDF vocabularies) on the existing GBIF Resources Repository as http://rs.gbif.org/terms/. The first prototype implementation for RDF vocabularies of terms has been demonstrated using the sandbox at the GBIF Resources Repository at http://rs.gbif.org/sandbox/terms/. The GBIF Resources Browser (http://tools.gbif.org/resource-browser/) includes a glossary of terms for the resources included at the GBIF Repository of terms and machine readable services more aimed at technology developers than the more human-centric wiki forum.

See also the minutes from the ViBRANT WP4 planning meeting at GBIF on 3 April 2012 available at "http://vbrant.eu/content/vibrant-wp4-project-coordination-meeting".

References:

Döring M, Robertson T, and Remsen D (2011). Darwin Core Archive Format, Reference Guide to the XML Descriptor File. Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark. 16 pp.

TDWG (2009a). Darwin Core [Online resource]. Contributors: Wieczorek J, Döring M, Giovanni RD, Robertson T, and Vieglais D. TDWG Darwin Core Task Group, Biodiversity Information Standards (TDWG). Available at http://www.tdwg.org/standards/450/ and at http://rs.tdwg.org/dwc/, verified 26 April 2012.

TDWG (2009b). Darwin Core text guidelines [Online resource]. Contributors: Robertson T, Döring M, Wieczorek J, Giovanni RD, and Vieglais D. TDWG Darwin Core Task Group, Biodiversity Information Standards (TDWG). Available at http://rs.tdwg.org/dwc/terms/guides/text/index.htm, verified 26 April 2012.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni RD, Robertson T, and Vieglais D (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. doi: 10.1371/journal.pone.0029715