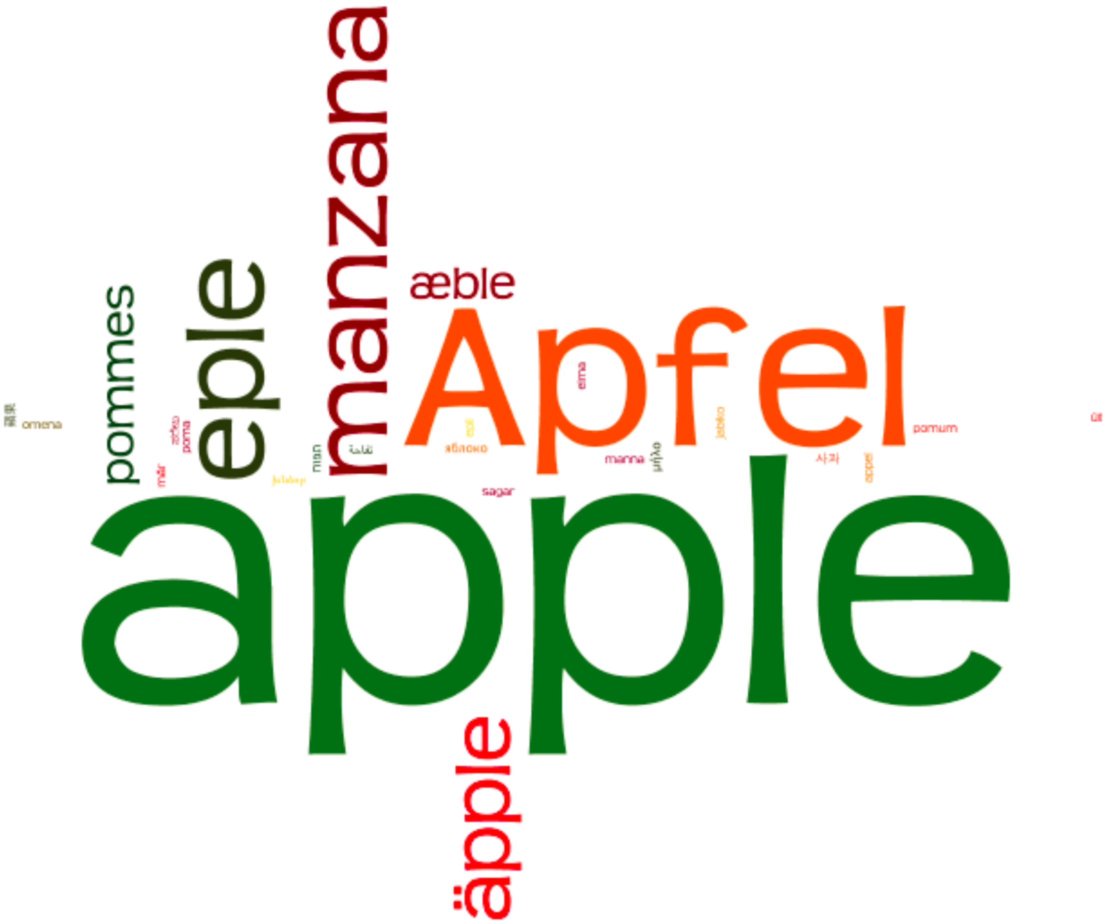
GBIF
www.gbif.org

Biodiversity Knowledge Organization System: Proposed Architecture

Version 0.3

Dag Terje Filip Endresen[1], Éamonn Ó Tuama[2], David Remsen[3]

March 2012 (*draft discussion document*)



[1] Knowledge Systems Engineer, Global Biodiversity Information Facility (GBIF)
[2] Senior Programme Officer for Inventory, Discovery, Access (DADI), GBIF
[3] Senior Programme Officer for the Electronic Catalogue of Names of Known Organisms (ECAT), GBIF

Suggested citation:
GBIF (2012). Biodiversity Knowledge Organization Systems: Proposed Architecture. Version 0.3. Authors: Dag Terje Filip Endresen, Éamonn Ó Tuama, David Remsen. Copenhagen: Global Biodiversity Information Facility (GBIF).

Language: English

Global Biodiversity Information Facility (GBIF)
Free and open access to biodiversity data
GBIF Secretariat, Universitetsparken 15, DK-2100 Copenhagen, Denmark
Home page: http://www.gbif.org

Virtual Biodiversity Research and Access Network for Taxonomy (ViBRANT)
Development of a virtual research community for managing biodiversity data on the web
Home page: http://vbrant.eu
Funded by the European Union (EU) 7th framework programme (FP7)

Biodiversity Knowledge Organization System, Architecture

Scoping document prepared by Dag Endresen, David Remsen, and Éamonn Ó Tuama, Global Biodiversity Information Facility (GBIF).

Summary

This document provides a proposed architecture for the new Knowledge Organization System (KOS) for biodiversity information resources to be hosted by GBIF. The proposed KOS architecture includes the following key components: “Vocabulary of Terms” (1), “Term Browser” (2), “Extensions and Code lists for the Darwin Core Archives” (3), domain Ontologies (4), and a “Resources Repository” (5). This document presents the overall architecture and how these conceptual building blocks are linked together. Some of the possible software tools that can be used to implement each of these components are also mentioned.

This document contributes to the development of a new Knowledge Organization System (KOS) for management of terms and concepts used for the description of biodiversity information resources.

Table of Contents

Abbreviations: knowledge organization system (KOS); persistent identifier (PID); resource description framework (RDF); RDF schema (RDFS); simple knowledge organization system (SKOS); uniform resource identifier (URI)

Introduction

Vocabularies and ontologies are types of Knowledge Organisation Systems (KOS). An introduction, including the different types of KOS, was presented by the US Council on Library and Information Resources; *"The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge management"* (Hodge, 2000). Controlled vocabularies provides a list of preselected terms associated with a description of its meaning (Harping, 2010). With this document *vocabularies* and *vocabularies of terms* are meant to refer to controlled vocabularies of authorized terms persistently identified by HTTP URIs that resolve to a machine readable description expressed using the resource description framework (RDF).

Background

The Biodiversity Information Standards (TDWG) organization provides a number of standards for the description of biodiversity information resources. Many of these standards are presented as an XML schema or in other XML formats and include the definition of terms and concepts designed for description of biodiversity information resources. Examples of TDWG standards presenting such a list of terms include the Darwin Core[4] (Wieczorek et al., 2012), Access to Biological Collection Data (ABCD)[5], Structured Descriptive Data (SDD)[6], Taxonomic Concept Transfer Schema (TCS)[7], Natural Collections Descriptions (NCD)[8], and Audubon Core[9]. These standard lists of terms sometimes include their own version of terms for the same or very similar concepts. This is of course because these standards were not designed to re-use existing terms. Among the TDWG standards only Darwin Core are expressed using the Resource Description Framework[10] (RDF) notation; and only this standard together with the Audubon Core identifies each individual term by persistent HTTP URIs.

Agreed standards with well-defined syntax and semantics[11] have made it possible to integrate distributed datasets maintained by data publishers from all over the world. Agreed standards for the data exchange protocol and user-friendly data publishing toolkits such as DiGIR[12], BioCASE[13], TapirLink[14], and the GBIF Integrated Data Publishing Toolkit (IPT)[15] enabled a dynamic information system where the federated network of biodiversity datasets are indexed on a regular basis by the GBIF data portal[16]. The Nodes Portal Toolkit (NPT)[17], currently under development as a community driven project, will assist the GBIF participants with building their own data portals to easily integrate datasets and other biodiversity information resources that are shared with the GBIF infrastructure.

[4] Darwin Core, http://rs.tdwg.org/dwc/

[5] Access to Biological Collection Data (ABCD), http://www.tdwg.org/standards/115/

[6] Structured Descriptive Data (SDD), http://www.tdwg.org/standards/116/

[7] Taxonomic Concept Transfer Schema, http://www.tdwg.org/standards/117/

[8] Natural Collections Descriptions (NCD), http://www.tdwg.org/standards/312/

[9] Audubon Core, http://species-id.net/wiki/Audubon_Core

[10] Resource Description Framework (RDF), http://www.w3.org/RDF/

[11] By syntax we mean the encoding, e.g., XML, etc., usually determined by a schema; semantics, on the other hand, refers to the meaning attached to properties (i.e., fields, attributes) that are expressed in the encoding.

[12] Distributed Generic Information Retrieval (DiGIR), http://digir.net/

[13] Biological Collection Access Service for Europe (BioCASE), http://biocase.org

[14] TapirLink, http://sourceforge.net/projects/digir/files/TapirLink/

[15] GBIF Integrated Data Publishing Toolkit, http://code.google.com/p/gbif-providertoolkit/

[16] GBIF Data Portal, http://data.gbif.org

[17] GBIF Nodes Portal Toolkit (NPT), http://code.google.com/p/gbif-npt/

Well-defined standards for sharing information resources within the biodiversity information community may, however, create a barrier for integration of information resources from other scientific communities. Information resources described using standards from other communities will generally require transformation of the data format as well as cross-mapping of the concepts/terms being used. When transforming data from one data standard to another, some information will often be lost. Loss of information is also typical for the transformation of the local data format (e.g. the database schema used by an individual Natural History collection) to a standard format (such as the Darwin Core). This loss of information is not only related to lack of terms in the standard to match all of the data properties maintained by the local system, but also related to differences between the semantic meaning of standards terms and the local data properties.

The Semantic Web with its Linked Data principles (Berners-Lee, 2006; Bizer et al., 2009) promotes the use and re-use of common concepts and vocabularies of terms each identified by globally unique and persistent HTTP URIs. Universal identification using persistent identifiers (PIDs) for machine-readable concepts supports re-use of terms and concepts across scientific communities when sharing datasets. Such best practices are the first step to improve cross-domain data interoperability. The next step in this process towards *"global interoperability of datasets"* is the application of universal Persistent Identifiers (PIDs) to the entities that constitute biodiversity resources. For example, when individual entities (such as the observation or occurrence of a species) are identified by PIDs, then federated data sources can re-use these PIDs to make explicit that they share information on the same entities. When data categories identified by PIDs are used for annotation of data properties, then public ontologies can be developed to describe the semantic relationship between such data categories to allow for semantic integration of datasets.

Provision of an aggregated, flat list of the terms used in biodiversity informatics including their definition and associated persistent identifier (PID) was proposed by the GBIF KOS[18] Task Group to support the re-use of terms for common concepts (Catapano, 2011). This list would underpin an application known as the proposed Term Browser that facilitates easy viewing of all terms and their definition and should constitute a foundational component in the GBIF KOS Architecture.

GBIF KOS Architecture

The components of the proposed GBIF Knowledge Organization System (KOS) Architecture are: Vocabulary of Terms (1), Resources Repository (2), Term Browser (3), Resources Browser (4), and Ontology Repository (5) (see Figure 1). These components are described in further details below. The fundamental building blocks of the GBIF KOS Architecture are Terms and Concepts. *Concepts* refer to an abstract idea of an entity or property while *terms* are the names or labels associated with these concepts. Different communities can use a suite of different software tools to re-use or mint new terms and organize them in a "Vocabulary of Terms". By *Vocabulary of Terms* we mean a dictionary or list of basic terms including the definition, but no ontological relationships. Each tool should allow expression of the vocabulary in RDF using the RDF Schema (RDFS) and Simple Knowledge Organization System (SKOS) vocabulary. A Vocabulary processor tool can be developed to harmonize and transform vocabularies expressed in other formats. These vocabularies can be registered (and a copy of the vocabulary deposited)

[18] Knowledge Organization System (KOS)

at the GBIF Resources Repository where Darwin Core Archive extensions and code lists are also stored. The Term Browser provides a human accessible portal interface for exploring terms from all registered vocabularies. The Term Browser can be designed in a similar manner to the existing GBIF Resources Browser that enables exploration of the terms and concepts included in the Darwin Core Archive extensions and code lists. New Biodiversity vocabularies (of terms) are recommended to re-use, wherever possible, the terms included in a ratified and published Vocabulary of Terms.

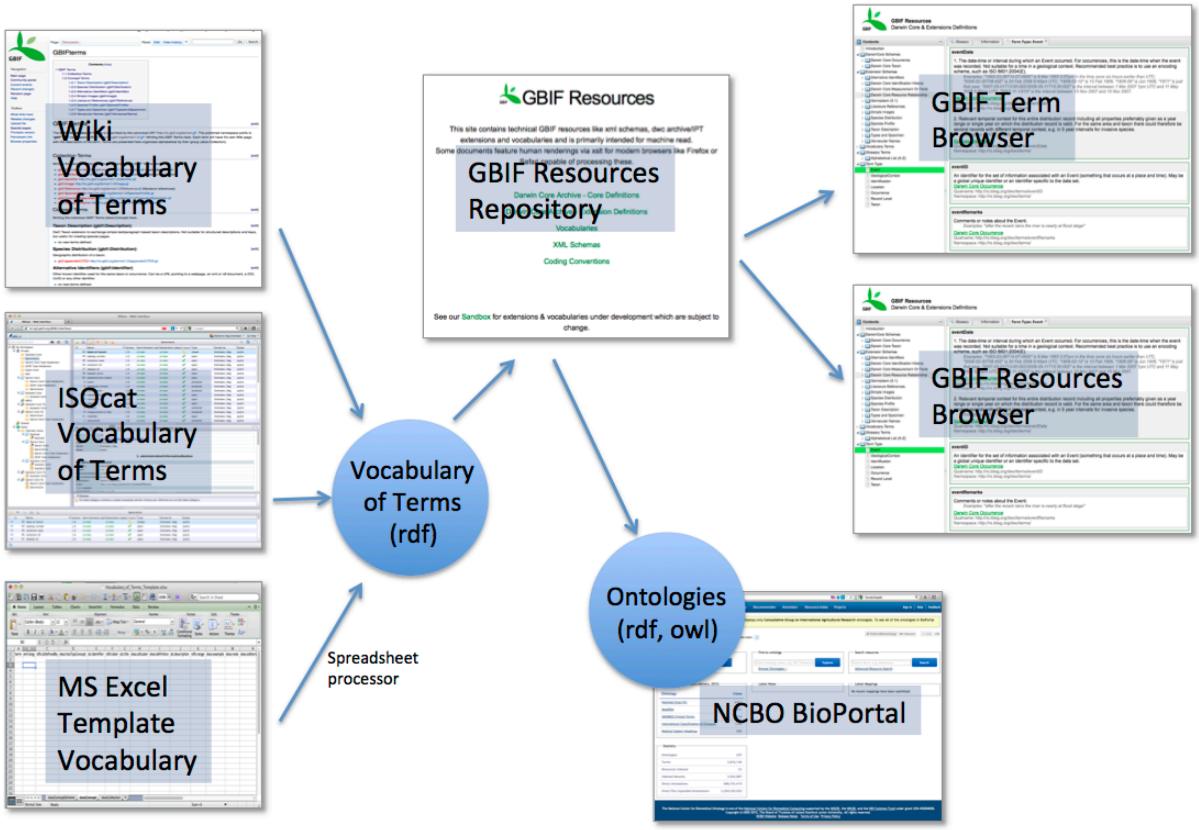


Figure 1: Overview of the proposed architecture for the resources repository and the term vocabularies.

Vocabulary of Terms

The management of terms for biodiversity information resources is often made in a collaborative manner by a group of experts. Each expert group publishes the terms and concepts managed under their control as a vocabulary (i.e. a collection or list of terms). This practice can be made a general best practice principle, including the recommendation to create new terms as part of an existing Vocabulary of Terms that is maintained and managed by a dedicated task group. The ratification of new terms could then be organized as part of the process to ratify a new version of the entire Vocabulary of Terms.

A uniform resource identifier (URI) should be assigned to identify each of the individual terms. If no HTTP resolvable URI is available for an individual term, then a new URI will be created for this term. If other types of persistent identifiers (PIDs) are available for a term, then this new tool should support the recording of this PID when adding new terms to the Vocabulary of Terms.

Proposal (1): Develop guidelines and best practices for the management of a Vocabulary of Terms by a formalized task group.

Proposal (2): When importing existing terms from a previous collection of terms into a Vocabulary of Terms, a new URI identifier will be created if one is not already provided.

Proposal (3): Each Vocabulary of Terms should be described by a separate document or resource using the RDF notation.

Proposal (4): Develop one system for presentation and discussion of new (and existing) terms using the Semantic Media Wiki platform (Krötzsch et al., 2007), and another system based on the GBIF Resources Repository (see later) for publishing the final and ratified versions of terms. (Terms to be published as members of a named and managed Vocabulary of Terms).

GBIF Vocabulary Server

The GBIF Vocabulary Server[19] (Harman et al., 2009) was designed to create extensions and code lists to the core data types supported by the Darwin Core Archive format (Döring et al., 2011). The GBIF Vocabulary Server has basic features that can be used to create new terms. However, these terms are created as part of an "extension" (following the Simple Darwin Core text guidelines[20]) or as part of a "vocabulary" (code list or controlled value-vocabulary). These "extensions" and "vocabularies" are designed for the GBIF Integrated Data Publishing Toolkit (IPT) and the Darwin Core Archive format[21]. The GBIF Vocabulary Server is well suited for building "extensions" and "vocabularies" for the GBIF IPT and the Darwin Core Archives, but should be limited in scope to only re-use terms defined using other tools. The interface should be updated to make this limitation explicit.

Question (1): Are the features and capabilities of the GBIF Vocabulary Server to define new basic terms and concepts useful and appropriate to keep - or should the GBIF Vocabulary Server be limited to allow only re-use of terms from a ratified and published Vocabulary of Terms?

Proposal (5): Limit the scope of the GBIF Vocabulary Server to only re-use terms and concepts defined by a Vocabulary of Terms and published at the GBIF Resources Repository (or by the [new] GBIF Term Browser).

Proposal (6): Initiate further developments of the GBIF Vocabulary Server to draw terms and concept from the ratified and published term vocabularies. This would allow ONLY the re-use of terms and concepts and cancel the requirement to provide terms and their definitions directly in the GBIF Vocabulary Server interface.

GBIF Resources Repository

The present GBIF infrastructure uses the GBIF Resources Repository[22] as the normative source for the official version of the Darwin Core Archive "extensions" and "vocabularies" that are created by tools such as the GBIF Vocabulary Server. Deployment versions of the "extensions" and "vocabularies" are uploaded to the GBIF Resources Repository. This repository can be extended in functionality as a repository for the official and ratified version of a Vocabulary of Terms. The goal of this new repository

[19] GBIF Vocabulary Server, http://vocabularies.gbif.org
[20] Simple Darwin Core Format, http://rs.tdwg.org/dwc/terms/simple/index.htm
[21] Darwin Core Text Guide, http://rs.tdwg.org/dwc/terms/guides/text/index.htm
[22] GBIF Resources Repository, http://rs.gbif.org

service is to provide improved access to promote the re-use of these terms when building new “extensions” and “vocabularies” (code lists) for the Darwin Core Archive format used by the GBIF IPT - as well as when building new ontologies for biodiversity information resources.

Proposal (7): Add a new section to the GBIF Resources Repository for publishing a Vocabulary of Terms (collection of basic terms). The URL to the terms section could be http://rs.gbif.org/terms/ and each vocabulary could be added as a new directory folder following the format http://rs.gbif.org/terms/[VOCABULARY]/.

Proposal (8): Include the final version of a Vocabulary of Terms to the GBIF Resources Repository only after ratification by a community such as the Biodiversity Information Standards (TDWG).

Term Browser

A new glossary of basic terms will provide an overview to support the identification of terms that can be re-used. The Dublin Core Metadata Initiative (DCMI) and the TDWG Darwin Core (DwC) could provide a general model and act as a guideline for how to maintain and publish the description of terms and concepts.

Proposal (9): Create a new portal interface providing an overview of basic terms and concepts (Glossary of Terms). The target software tool should maintain a flat list (or lists) of individual terms. Terms and their definitions should be retrieved from the (proposed) new section of the GBIF Resources Repository for Terms (Vocabulary of Terms).

Question (2): Should the same software tool (portal interface to terms and concepts) also support the minting of new terms including the description (definition) for these new terms?

GBIF Resources Browser

The online user interface to the Terms Used in Bionomenclature[23] (Hawksworth, 2010) was used as a model to build a portal (GBIF Resources Browser) to the terms and concepts included to the “extensions” and “vocabularies” (code lists)[24]. GBIF and SilverBiology developed this tool using the Ext JS JavaScript framework[25] and the source code was made available at Google Code[26]. This tool can be updated to provide a new and more general presentation of “Terms Used in Biodiversity” based on the vocabularies registered in the GBIF Resources Repository.

Discussion forum

The Biodiversity KOS architecture should include a threaded discussion forum. Each discussion post or thread could be annotated using the URI (or PID) for the relevant term(s). The discussions could sometimes be relevant to multiple terms. Relevant email discussions from the TDWG mailing lists could perhaps be imported and annotated with the relevant term(s) to be included in the threaded archive of discussions available for each term.

[23] Terms Used in Bionomenclature, http://bionomenclature-glossary.gbif.org/

[24] GBIF Resources browser, http://tools.gbif.org/resource-browser/

[25] Ext JS, http://www.sencha.com/products/extjs/

[26] Source Code, http://code.google.com/p/terms-of-bionomenclature/

Proposal (10): Use the GBIF Community Site as the discussion forum.

Question (3): Will the GBIF Community Site support annotation using the PID/URI for the relevant terms?

Proposal (11): As a supplement to the GBIF Community Site, use the Semantic MediaWiki platform as a discussion platform, and as an archive for the most important discussions leading to the description of a new term.

TDWG Ontologies

The description of semantic relationships between the terms and concepts used to describe resources is outside of the scope for the "Vocabulary of Terms". Appropriate description of semantic relationships can be addressed effectively by the development of domain ontologies. The TDWG Ontology[27] was initiated by the first meeting of the TDWG technical architecture group (TAG) in April 2006 (TDWG TAG 2006a). The TDWG Ontology was designed to follow a three-layered approach. The first layer was called "**base ontology**" and proposed to describe fundamental abstract concepts from which all classes and properties of the TDWG ontologies would be based. The second layer was the "**core ontology**" to describe classes and properties for the most common and general concepts within the TDWG community. The third layer was the "**domain ontologies**" with concrete classes and properties for use by resources such as the application schemes (TDWG TAG, 2006b). See also the description of the TDWG Ontology by Roger Hyam (2009).

Proposal (12): Provide guidelines and recommendations to base the domain ontologies on the re-use of terms defined by ratified and published Vocabularies of Terms. The re-use of terms from other communities would also be a best practice recommendation.

Question (4): Does the "base ontology" and "core ontology" still fulfill the envisioned critical function as the base for the "domain ontologies", or could a best practice recommendation include the re-use of common and general concepts defined outside of the TDWG community.

Workflow

1. New terms and concepts are collaboratively developed using various software tools such as, for example, the Semantic MediaWiki or Drupal-based tools.
2. Terms and concepts for description of biodiversity information resources are managed as part of a Vocabulary of Terms controlled by an expert group. [Dependencies: Guidelines for the development of Vocabulary of Terms].
3. A community such as the Biodiversity Information Standards (TDWG) will ratify the final version of a Vocabulary of Terms including its publication using a persistent namespace.
4. The ratified version for each new Vocabulary of Terms is registered and deposited at the GBIF Resources Repository.
5. A new Term Browser (as proposed by this document) will read the terms from the GBIF Resources Repository and provide a portal interface to discover the terms and identify terms for re-use when creating new resources such as an "extension" or a "vocabulary" for the Darwin Core Archives or when designing a new ontology for biodiversity information resources.

[27] TDWG Ontology, http://rs.tdwg.org/ontology/

6. Darwin Core Archive extensions and code lists will be designed to re-use terms from one of the ratified and published Vocabulary of Terms.
7. Best practices for the development of Biodiversity Ontologies could recommend to re-use terms from a ratified and published Vocabulary of Terms.

References

Basca C, Corlosquet S, Cyganiak R, Fernández S, and Schandl T (2008). Neologism: Easy Vocabulary Publishing. *In*: Proceedings of the 4th Workshop on Scripting for the Semantic Web, Tenerife, Spain, June 02, 2008, CEUR Workshop Proceedings, ISSN 1613-0073. Available online at http://CEUR-WS.org/Vol-368/paper10.pdf, verified 3 Feb 2012.

Berners-Lee T (2006). Design issues, Linked data [online]. World Wide Web Consortium (W3C), Massachusetts Institute of Technology, Cambridge, MA, USA. Available at http://www.w3.org/DesignIssues/LinkedData.html, verified 3 Feb 2012. Last updated 18 June 2009.

Bizer, C, Heath T, and Berners-Lee T (2009). Linked data, the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3): 1–22. DOI: 10.4018/jswis.2009081901

Catapano T, Hobern D, Lapp H, Morris RA, Morrison N, Noy N, Schildhauer M, and Thau D (2011). Recommendations for the use of knowledge organization systems by GBIF. Released on 4 February 2011. Global Biodiversity Information Facility (GBIF), Copenhagen. Available at http://www.gbif.org/orc/?doc_id=2942&l=en, verified 10 Feb 2012.

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

Harman KT, Hyam R, Remsen DP (2009). Vocabularies – managing them. p. 10-11 In: Weitzman, A.L., Proceedings of TDWG (2009), Montpellier, France. Biodiversity Information Standards (TDWG). Available at http://www.tdwg.org/proceedings/article/view/605, verified 1 Feb 2012.

Harping P (2010). Introduction to controlled vocabularies: Terminology for art, architecture, and other cultural works. Online edition. Getty Research Institute, Los Angeles, CA. ISBN: 978-1-60606-026-1. Available at: http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/index.html, verified 10 Feb 2012.

Hawksworth DL (2010). Terms used in bionomenclature: The naming of organisms (and plant communities). Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-09-7. Available at http://www.gbif.org/orc/?doc_id=2430&l=en, verified 8 Feb 2012.

Hodge G (2000). Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files. Council on Library and Information Resources, Washington DC. Available at: http://www.clir.org/pubs/reports/pub91/contents.html, verified 10 Feb 2012.

Hyam R (2009). Managing the managing of the TDWG ontology [online blog]. Available at http://www.hyam.net/blog/archives/643, verified 8 Feb 2012.

Krötzsch M, Vrandecic D, Völkel M, Haller H, Studer R (2007). Semantic Wikipedia. Journal of Web Semantics 5: 251–261. Doi: 10.1016/j.websem.2007.09.001

TDWG TAG (2006a). TDWG Technical Architecture Group 11th to 13th April 2006 eSI Edinburgh. Technical Architecture Group (TAG), Biodiversity Information Standards/Taxonomic Databases Working Group (TDWG). Available at http://www.tdwg.org/uploads/media/TAG-1_Report_02.pdf, verified 8 Feb 2012.

TDWG TAG (2006b). TDWG core ontology meeting 16th to 18th May 2006 eScience Institute, Edinburgh, UK. Technical Architecture Group (TAG), Biodiversity Information Standards/Taxonomic Databases Working Group (TDWG). Available at http://www.tdwg.org/uploads/media/TDWG_TAG_Ontology_Report_02.doc, verified 8 Feb 2012.

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