GBIF Vocabularies

GBIF publishes a series of vocabularies at <http://rs.gbif.org/vocabulary>. These were created to provide controlled values for properties in the core and extension definitions of Darwin Core archives at <http://rs.gbif.org/core/> and <http://rs.gbif.org/extension/> respectively. This document is intended to foster discussion on how GBIF can best manage such vocabularies and their use in Darwin Core archives and future exchange formats, e.g., the outcome of the W3C CSV on the web working group .

# How are vocabularies used?

Vocabularies are required for encoding data in Darwin Core archives. The following table is an extract from a simple DwC archive consisting of just a single (core) table based on the Occurrence core. Each row in the table is a record of the occurrence of a taxon. More formally, we can state that each row is an instance of class Occurrence. A class is a recognisable entity whose members share certain properties (attributes). The DwC archive core definition for class Occurrence[[1]](#footnote-1) contains many properties, three of which are shown in the table: occurrenceID, lifeStage and sex.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | occurrenceID | lifeStage | sex | … |
| 1 | bh\_002 | adult | female | … |
| 2 | bf\_023 | juvenile | male | … |

Vocabularies are used for the following:

1. To define the classes of entities (e.g., Occurrence, Taxon, Location);
2. To define the properties associated with a class, e.g., instances of the Occurrence class can have, among others, an occurrenceID, lifeStage and sex; instances of the Location class can have a decimalLatitude;
3. To define the permissible values for a particular property, e.g., instead of just allowing free text, the values for the sex property could be restricted to a controlled list (vocabulary) consisting of “male”, “female”, “hermaphrodite”.

# Darwin Core archives

A Darwin Core Archive (DwC-A) consists of a core CSV file, zero or more extension CSV files, a meta.xml file and an EML.xml file (Figure 1). The core and extension files contain the actual data and are related in a one-to-many manner. The meta.xml file describes the structure of the CSV files including any mappings from column headers to vocabulary terms. The EML.xml provides data set level metadata according to the Ecological Metadata Language (EML) specification.



Figure 1. The components of a Darwin Core archive.

# Vocabularies in Darwin Core archives

Each core and extension is defined in a simple XML format devised by GBIF with both human readable HTML and also XML versions available, e.g., <http://rs.gbif.org/core/dwc_taxon.xml> . A definition consists of a list of permissible vocabulary terms for the core or extension. For example, in the case of the Taxon core, these are largely drawn from the DwC Taxon class (Figure 2).

Cores and extensions can be viewed as class definitions with associated **properties**, e.g., class Taxon has as one of its properties dwc:scientificName. A property in a class definition may be restricted to a greater or lesser extent in the values it can have (i.e., its **range**). The main options, with increasing restriction are:

1. Free text
2. Free text following a pre-defined format, e.g., a date format like DD-MM-YYYY
3. Selection of a value from a controlled list

The third option, a controlled list of terms, is commonly called a **controlled vocabulary**. At its simplest, a controlled vocabulary consists of a flat list of concepts (terms) but more complex hierarchical structures as exemplified by **taxonomies** and **thesauri** are also an option. Concepts in vocabularies are typically assigned identifiers in the form of URIs. An URI unambiguously identifies a concept thus allowing it to be used widely across the web. Resolving a concept URI returns its definition so a prospective user can be guided on use.

As for vocabularies like Darwin Core, controlled vocabularies can be “published” on the web in a well known format such as RDF, SKOS or OWL. Such formats are useful not just for ontology based applications, e.g., by using these formats, vocabularies are more easily discovered, consumed and used in other applications including sites that aggregate terminology to promote re-use, e.g., Linked Open Vocabularies (<http://lov.okfn.org/dataset/lov/>).

# taxon-core.png

Figure 2. Vocabulary terms in the Taxon core.

# DwC archive cores and extensions in use

GBIF currently uses just two cores – Occurrence and Taxon (<http://rs.gbif.org/core/>).The properties for these two cores (classes) are drawn from the Darwin Core and Dublin Core vocabularies (Table 1). In contrast to the two cores, many more extensions have been deployed (currently 16 - see <http://rs.gbif.org/extension/>) (Table 2) and while Darwin Core and Dublin Core are the main vocabularies used for properties, four additional vocabularies are also used in these extensions: IUCN terms, WGS84 Geo Positioning, GBIF terms, and the Germplasm Term Vocabulary.

Table 1. Darwin Core Archive cores currently in use. The (property) vocabularies used by each extension are listed.

|  |  |  |
| --- | --- | --- |
| **Group** | **Core** | **Vocabularies used** |
| dwc | dwc\_occurrence.xml | <http://rs.tdwg.org/dwc/terms/><http://purl.org/dc/terms/>  |
| dwc\_taxon.xml | <http://rs.tdwg.org/dwc/terms/><http://purl.org/dc/terms/> |

Table 2. Darwin Core Archive extensions currently in use. The (property) vocabularies used by each extension are listed.

|  |  |  |
| --- | --- | --- |
| **Group** | **Extension** | **Vocabularies used** |
| dwc | identification.xml | <http://rs.tdwg.org/dwc/terms/>  |
| measurement\_or\_facts.xml | <http://rs.tdwg.org/dwc/terms/>  |
| resource\_relation.xml | <http://rs.tdwg.org/dwc/terms/>  |
| gbif/1.0 | description.xml | <http://purl.org/dc/terms/>  |
| distribution.xml | <http://rs.tdwg.org/dwc/terms/><http://iucn.org/terms/><http://rs.gbif.org/terms/1.0/>  |
| identifier.xml | <http://purl.org/dc/terms/>  |
| images.xml | <http://purl.org/dc/terms/><http://www.w3.org/2003/01/geo/wgs84_pos>  |
| multimedia.xml | <http://purl.org/dc/terms/>  |
| references.xml | <http://purl.org/dc/terms/>  |
| speciesprofile.xml | <http://rs.gbif.org/terms/1.0/><http://rs.tdwg.org/dwc/terms/>  |
| typesandspecimen.xml | <http://rs.tdwg.org/dwc/terms/><http://rs.gbif.org/terms/1.0/><http://purl.org/dc/terms/>  |
| vernacularname.xml | <http://rs.tdwg.org/dwc/terms/><http://purl.org/dc/terms/> <http://rs.gbif.org/terms/1.0/>  |
| germplasm | GermplasmAccession.xml | <http://purl.org/germplasm/germplasmTerm> <http://www.w3.org/2003/01/geo/wgs84_pos> <http://rs.tdwg.org/dwc/terms/>  |
| MeasurementScore.xml | <http://purl.org/germplasm/germplasmTerm> <http://rs.tdwg.org/dwc/terms/>  |
| MeasurementTrait.xml | <http://purl.org/germplasm/germplasmTerm> <http://rs.tdwg.org/dwc/terms/>  |
| MeasurementTrial.xml | <http://purl.org/germplasm/germplasmTerm> <http://rs.tdwg.org/dwc/terms/><http://www.w3.org/2003/01/geo/wgs84_pos>  |
| Eol(http://eol.org/schema/) | media\_extension.xml | <http://purl.org/dc/terms/><http://rs.tdwg.org/dwc/terms/><http://rs.tdwg.org/audubon_core/><http://iptc.org/std/Iptc4xmpExt/1.0/xmlns/><http://rs.tdwg.org/ac/terms/> <http://ns.adobe.com/xap/1.0/><http://ns.adobe.com/xap/1.0/rights/><http://eol.org/schema/agent/><http://www.w3.org/2003/01/geo/wgs84_pos><http://eol.org/schema/reference/>  |
| reference\_extension.xml | <http://purl.org/dc/terms/><http://eol.org/schema/reference/><http://purl.org/ontology/bibo/><http://schemas.talis.com/2005/address/schema>  |
| sandbox | audubon.xml | <http://purl.org/dc/elements/1.1><http://rs.tdwg.org/ac/terms><http://ns.adobe.com/xap/1.0/rights><http://ns.adobe.com/photoshop/1.0><http://purl.org/dc/terms><http://iptc.org/std/Iptc4xmpExt/2008-02-29><http://ns.adobe.com/xap/1.0><http://ns.adobe.com/exif/1.0><http://rs.tdwg.org/dwc/terms> |

As described earlier, a core or extension may restrict the range of values of a particular property by use of a controlled vocabulary. GBIF has created several such controlled vocabularies. These are organised at <http://rs.gbif.org/vocabulary/> under various groupings and also listed in Table 3. Each vocabulary consists of several terms which may be drawn from other vocabularies or newly minted by GBIF. The number of terms minted by GBIF per vocabulary is listed together with the cores and extensions that use the vocabulary. Table 4 provides an alternate view listing each of the cores and extensions with the controlled vocabularies they use.

Table 3. GBIF vocabularies organised by grouping. For each, the number of terms minted by GBIF and any DwC-A core or extension using the vocabulary are listed.

|  |  |  |  |
| --- | --- | --- | --- |
| **Group** | **Vocabulary**  | **GBIF Terms** | **Used by Core/Extension** |
| ac | rating.xml | 7 |  |
| basic | boolean.xml  | 2 | vernacularname.xml |
| dcterms | type.xml |  | dwc-occurrence.xml;multimedia.xml |
| dwc | basis\_of\_record.xml |  | dwc-occurrence.xml |
| eml | update\_frequency.xml | 11 |  |
| eol | datatype.xmllicense.xmlmimetype.xmlsubject.xml |  |  |
| gbif | agent\_role.xml | 15 |  |
| dataset\_subtype.xml | 8 |  |
| dataset\_type.xml | 4 |  |
| description\_type.xml | 38 |  |
| establishment\_means.xml | 6 | distribution.xml |
| life\_form.xml | 6 |  |
| life\_stage.xml | 10 | distribution.xml;vernacularname.xml |
| nomenclatural\_code.xml | 6 | dwc\_taxon.xml |
| nomenclatural\_status.xml | 29 | dwc\_taxon.xml |
| occurrence\_status.xml | 7 | distribution.xml |
| preservation\_method.xml | 14 |  |
| rank.xml | 38 | dwc\_taxon.xml |
| reference\_type.xml | 10 | references.xml |
| resource\_type.xml | 8 |  |
| sex.xml | 4 | vernacularname.xml |
| taxonomic\_status.xml | 6 | dwc\_taxon.xml |
| type\_designation\_type.xml | 9 | typesandspecimen.xml |
| type\_status.xml | 41 | typesandspecimen.xml |
|  |  |  |
| germplasm | AcquisitionSourceType.xmlBiologicalStatusType.xmlStorageCondityionType.xml |  |  |
| ggbn | permit\_status.xmlpreservation\_temperature.xmlquality.xml |  |  |
| gisin/v2.0 | abundance.xmldistribution.xmlharmful.xmlorigin.xmlpersistence.xmlpresence.xmlregulatory.xmlspreadrate.xmltrend.xml |  |  |
| iso | 639-1.xml |  | vernacularname.xml |
| 639-2.xml |  |  |
| 3166-1\_alpha2.xml |  |  |
| iucn | habitat.xml | 127 | speciesprofile.xml |
| threat\_status.xml | 9 | distribution.xml |
| mixs | env\_package.xmlinvestigation\_type.xml |  |  |
| sn2000 | habitat.xml | 23 |  |
| un | cites\_appendix.xml | 3 | distribution.xml |

Table 4. IPT cores and extensions and the vocabulary schemes they use.

|  |  |
| --- | --- |
| **IPT** | **Controlled Vocabulary** |
| **Cores** |  |
| dwc\_occurrence.xml | [type.xml](http://rs.gbif.org/vocabulary/dcterms/type.xml)[basis\_of\_record.xml](http://rs.gbif.org/vocabulary/dwc/basis_of_record.xml) |
| dwc\_taxon.xml | [rank.xml](http://rs.gbif.org/vocabulary/gbif/rank.xml)[nomenclatural\_code.xml](http://rs.gbif.org/vocabulary/gbif/nomenclatural_code.xml)[taxonomic\_status.xml](http://rs.gbif.org/vocabulary/gbif/taxonomic_status.xml) [nomenclatural\_status.xml](http://rs.gbif.org/vocabulary/gbif/nomenclatural_status.xml) |
| **Extensions** |  |
| description.xml | none |
| distribution.xml | [life\_stage.xml](http://rs.gbif.org/vocabulary/gbif/life_stage.xml) [occurrence\_status.xml](http://rs.gbif.org/vocabulary/gbif/occurrence_status.xml) [threat\_status.xml](http://rs.gbif.org/vocabulary/iucn/threat_status.xml) [establishment\_means.xml](http://rs.gbif.org/vocabulary/gbif/establishment_means.xml) [cites\_appendix.xml](http://rs.gbif.org/vocabulary/un/cites_appendix.xml) |
| identifiers.xml | none |
| images.xml | none |
| multimedia.xml | [type.xml](http://rs.gbif.org/vocabulary/dcterms/type.xml)  |
| references.xml | [reference\_type.xml](http://rs.gbif.org/vocabulary/gbif/reference_type.xml)  |
| speciesprofile.xml | [habitat.xml](http://rs.gbif.org/vocabulary/iucn/habitat.xml)  |
| typesandspecimen.xml | [type\_status.xml](http://rs.gbif.org/vocabulary/gbif/type_status.xml) [type\_designation\_type.xml](http://rs.gbif.org/vocabulary/gbif/type_designation_type.xml) |
| vernacularname.xml | [639-1.xml](http://rs.gbif.org/vocabulary/iso/639-1.xml) [sex.xml](http://rs.gbif.org/vocabulary/gbif/sex.xml) [life\_stage.xml](http://rs.gbif.org/vocabulary/gbif/life_stage.xml) [boolean.xml](http://rs.gbif.org/vocabulary/basic/boolean.xml)  |

# Management of cores, extensions and vocabularies

Cores and extensions are best managed separately from the vocabularies they use. The former can be considered as application schemas supporting the Darwin Core Archive format whereas vocabularies themselves will normally have much wider applicability beyond DwC archives.

## Cores

1. Continue to use <http://rs.gbif.org/core/> for publishing core definitions.
2. A core should be versioned as the definition may change over time through deprecation or addition of properties. The versioning can be implemented via the URI for the definition, e.g., http://rs.gbif.org/core/1.0/dwc\_taxon.xml. Current cores are not versioned:
	1. <http://rs.gbif.org/core/dwc_taxon.xml>
	2. <http://rs.gbif.org/core/dwc_occurrence.xml>
3. The current plain XML format for cores is adequate. Consider whether there are any advantages in using RDF/OWL as alternatives.
4. Consider whether with the decommissioning of vocabulaies.gbif.org a dedicated application for creating a core is needed.

## Extensions

1. Continue to use <http://rs.gbif.org/extension/> for publishing extension definitions.
2. An extension should be versioned as the definition may change over time through deprecation or addition of properties. Current extensions are already versioned via their URI:
	1. <http://rs.gbif.org/extension/gbif/1.0/distribution.xml>
	2. <http://rs.gbif.org/extension/gbif/1.0/vernacularname.xml>
	3. <http://rs.gbif.org/extension/gbif/1.0/typesandspecimen.xml>
3. The current plain XML format for extensions is adequate. Consider whether there are any advantages in using RDF/OWL as alternatives.
4. Consider whether with the decommissioning of vocabulaies.gbif.org a dedicated application for creating an extension is needed.

## Vocabularies

Vocabularies as discussed in this document are groupings of concepts where, in some cases, an rdf:type of property or class can be applied to an individual concept. In this sense, the management of a vocabulary like Darwin Core (mostly a collection of properties of a small set of classes) is no different to a “controlled vocabulary” like Life Stage (a simple collection of concepts, each of which can be treatable as skos:concept).

1. The foremost requirement for a vocabulary term is that it should have an identifier in the form of a normative URI.
2. A vocabulary term URI should be resolvable to an information page that provides a definition of the term.
3. Related terms are typically grouped together and published as a named vocabulary in a normative document.
4. The normative document can take a number of formats. For example,
	1. Darwin Core is defined in RDF at <http://rs.tdwg.org/dwc/rdf/dwctermshistory.rdf>
	2. Audubon Core is defined in a (HTML) web page at: <http://terms.tdwg.org/wiki/Audubon_Core_Term_List>
	3. The FOAF specification is in a (HTML) web page at: <http://xmlns.com/foaf/spec/> .
5. The normative document can serve as the resolution point for its vocabulary terms, e.g., the Audubon Core term accessURI identifier (<http://rs.tdwg.org/ac/terms/accessURI>) is resolved to <http://terms.tdwg.org/wiki/Audubon_Core_Term_List#ac:accessURI>. FOAF terms (e.g., <http://xmlns.com/foaf/0.1/Person>) resolve to the spec page <http://xmlns.com/foaf/spec/>. Alternatively, a separate document can be set up for resolving terms, e.g., terms in Darwin Core (e.g. <http://rs.tdwg.org/dwc/terms/lifeStage>) resolve to a special document <http://rs.tdwg.org/dwc/terms/index.htm#lifeStage>. The use of fragment identifiers in resolution is optional.
6. Providing a machine readable version of a vocabulary in RDF/SKOS/OWL (or possibly JSON-LD) helps to ensure that the terms are more easily discovered, consumed and used in other applications including sites that aggregate terminology. It would enable future versions of DwC archives to be more self contained by importing full information on a term (e.g., labels, definitions and examples in multiple languages) instead of just using holding the URI (in the meta.xml file) and relying on on-the-fly resolution at processing time should more information on a term be required and with no guarantee of being able to access it in a standard format.

## Managing vocabularies in the GBIF namespace

There are issues with term resolution for several of the vocabularies at <http://rs.gbif.org/vocabulary/gbif/>. These are listed in Appendix 1. In addressing these issues, we should consider the following options.

1. Continue to use <http://rs.gbif.org/vocabulary/gbif/> for publishing the normative document of a vocabulary in the form of a web page. Use the normative document as the resolution page for vocabulary terms. Consider implementing fragment identifiers which are currently not used, e.g., <http://rs.gbif.org/vocabulary/gbif/life_stage/juvenile> resolves to <http://rs.gbif.org/vocabulary/gbif/life_stage.xml>.
2. Create a GitHub page for each vocabulary to capture comments and suggestions for ongoing development of the vocabulary.
3. Create a machine readable version of each vocabulary in RDF/SKOS/OWL and make available in GitHub.
4. A more radical alternative is to decommission <http://rs.gbif.org/vocabulary> and create a GitHub page as the normative document for each vocabulary.
5. Assess the status of <http://rs.gbif.org/terms/1.0/>. The vernacularname.xml extension uses three properties from this (isPlural, isPreferredName, organismPart). At least, ensure all terms have definitions. Should they resolve to a more human friendly web page (as for controlled vocabularies)?

# Appendix 1

Table . Term resolution issues for vocabularies.

|  |  |
| --- | --- |
| **Vocabulary** | **Issue** |
| * <http://rs.gbif.org/vocabulary/gbif/reference_type.xml>
* <http://rs.gbif.org/vocabulary/gbif/taxonomic_status.xml>
* <http://rs.gbif.org/vocabulary/gbif/preservation_method.xml>
* <http://rs.gbif.org/vocabulary/gbif/agent_role.xml>
* <http://rs.gbif.org/vocabulary/gbif/dataset_subtype.xml>
* <http://rs.gbif.org/vocabulary/gbif/dataset_type.xml>
* <http://rs.gbif.org/vocabulary/gbif/description_type.xml>
* <http://rs.gbif.org/vocabulary/eml/update_frequency.xml>
 | Instead of the underscore for the vocabulary name, e.g., agent\_role, the term URIs use camel case, e.g., [http://rs.gbif.org/vocabulary/gbif/](http://rs.gbif.org/vocabulary/gbif/agentRole/editor)**[agentRole](http://rs.gbif.org/vocabulary/gbif/agentRole/editor)**[/editor](http://rs.gbif.org/vocabulary/gbif/agentRole/editor). Changing the URI to [http://rs.gbif.org/vocabulary/gbif/**agent\_role**/editor](http://rs.gbif.org/vocabulary/gbif/agent_role/editor) enables resolution. |
| * <http://rs.gbif.org/vocabulary/un/cites_appendix.xml>
 | The URI is of the form <http://rs.gbif.org/vocab/cites/appendixI> and needs to be <http://rs.gbif.org/vocabulary/un/cites_appendix/appendixI>. |
| * <http://rs.gbif.org/vocabulary/basic/boolean.xml>
 | The URI is of form <http://rs.gbif.org/vocab/boolean.xml> and needs to be <http://rs.gbif.org/vocabulary/basic/boolean.xml>  |
| * <http://rs.gbif.org/vocabulary/iso/639-1.xml>
 | The term URIs, e.g., <http://iso.org/639-1#aa> produce Not Found error |
| * <http://rs.gbif.org/vocabulary/iso/639-2.xml>
 | The term URIs, e.g., <http://iso.org/iso639-2/alpha3#aar> produce Not Found error |
| * <http://rs.gbif.org/vocabulary/iso/3166-1_alpha2.xml>
 | The term URIs, e.g., <http://iso.org/iso3166-1/alpha2#AD> produce Not Found error. Update to 3166, e.g., <https://www.iso.org/obp/#iso:code:3166:AD>  |
| * <http://rs.gbif.org/vocabulary/sn2000/habitat.xml>
 | URIs of form [http://rs.gbif.org/vocabulary/sn2000/habitat/aq:br](http://rs.gbif.org/vocabulary/sn2000/habitat/aq%3Abr) not found |
| * <http://rs.gbif.org/vocabulary/iucn/habitat.xml>
 | URIs of form <http://rs.gbif.org/vocabulary/iucn/habitat/1.1> not found |
| * <http://rs.gbif.org/vocabulary/ggbn/permit_status.xml>
 | URIs, e.g., <http://data.ggbn.org/schemas/ggbn/vocabulary/permit_status/available_upon_request> not found |
| * <http://rs.gbif.org/vocabulary/ggbn/preservation_temperature.xml>
 | URIS, e.g., <http://data.ggbn.org/schemas/ggbn/vocabulary/preservation_temperature/Celsius> not found |
| * <http://rs.gbif.org/vocabulary/ggbn/quality.xml>
 | URIs, e.g., <http://data.ggbn.org/schemas/ggbn/vocabulary/quality/low> not found |
| * <http://rs.gbif.org/vocabulary/eol/mimetype.xml>
 | URIs, e.g., <http://www.eol.org/IPT/vocabulary/mimetype/image/jpeg> not found |
| * <http://rs.gbif.org/vocabulary/dwc/basis_of_record.xml>
 | URIs, e.g., <http://rs.tdwg.org/dwc/dwctype/PreservedSpecimen> not found |

# Appendix 2. Term vs Concept

Vocabularies (including glossaries, dictionaries, taxonomies, thesauri, ontologies, etc) are collections of concepts, also loosely known as terms. However, strictly speaking, concepts differ from terms (Figure 3).

“thesauri show the relationships between **concepts** – units of thought – and distinguishes these from the **terms** that are used to label these concepts. These terms may be in one or more languages, and one term per language is chosen as a preferred term for each concept. One or more additional terms for the same concept may be recorded in the thesaurus as non-preferred terms.” [1]



Figure . Distinguishing term from concept. Source [2].

# Concept types

RDF, SKOS and OWL encode concepts in a machine readable format. While SKOS recognises just concepts and collections of concepts [3]. For example,

“cowMilk” is of type concept with a preferred term (label) in English of “cow milk”:

ex:cowMilk rdf:type skos:Concept;

 skos:prefLabel "cow milk"@en;

 skos:broader ex:milk.

The concept “cowMilk” is a member of the concept collection “\_:b0” with the preferred label in English of “milk by source animal”.

\_:b0 rdf:type skos:Collection;

 skos:prefLabel "milk by source animal"@en;

 skos:member ex:cowMilk;

RDF distinguishes other types, in particular property and class.

<rdf:type rdf:resource="<http://www.w3.org/1999/02/22-rdf-syntax-ns#Property>"/>
vs
<rdf:type rdf:resource="<http://www.w3.org/2000/01/rdf-schema#Class>"/>

The concepts within the Darwin Core vocabulary are largely of type property, grouped under a small set of organising classes. So, e.g., dwc:Taxon is a class and one of the properties (attributes) that members of that class can have is dwc:taxonID. The RDF document for Darwin Core defines these concepts as follows:

<rdf:Description rdf:about="http://rs.tdwg.org/dwc/terms/**Taxon**">

<rdfs:label xml:lang="en">Taxon</rdfs:label>

…

<rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#**Class**"/>

</rdf:Description>

<rdf:Description rdf:about="http://rs.tdwg.org/dwc/terms/**taxonID**">

<rdfs:label xml:lang="en">Taxon ID</rdfs:label>

…

<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#**Property**"/>

</rdf:Description>

# GBIF vocabularies

Specification of the GBIF vocabularies thus includes:

1. Identification of the most common and useful classes of entities in the biodiversity domain, e.g., Taxon, Occurrence, Event, Location, Collection, Collector, etc.
2. Identification of the most common and useful properties (i.e., attributes) of those classes, e.g., instances of the Taxon class can have taxonID, vernacularName, nomenclaturalCode, taxonomicStatus, etc.
3. Identification of the permissible values for the various properties; the range of permissible values is typically encoded in a controlled vocabulary where each item (concept) is expressed as a skos:Concept.

This is the RDF data model [4] consisting of triple statements of the structure:

<subject> <predicate> <object> (i.e. subject – property – value)

International Resource Identifiers (IRI) must be used to identify the subject and predicate in a triple statement. The object may optionally use an IRI or may be just a literal. The subject will typically be an instance of some class (as may the object) – hence the need to identify the main classes in a domain like biodiversity and use persistent unique identifiers to their instances. Predicates should be drawn from published vocabularies where the concepts (in this case, properties) have IRI identifiers, e.g., the DwC concept of taxonID has an IRI:

<http://rs.tdwg.org/dwc/terms/taxonID>

In the context of Darwin Core archives, the values for the object are likely to be literals. That can be free text or taken from a controlled vocabulary where the label for the concept in English is used. However, the IRI for the concept could be used instead and parsed to obtain the label in a more appropriate language.

[1] Will, L. (2012). The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus. Bulletin of the American Society for Information Science and Technology 38(4): 48-51; <http://onlinelibrary.wiley.com/doi/10.1002/bult.2012.1720380413/abstract>

[2] Dextre Clarke, S.G. and L. Zeng (2012). From ISO 2788 to ISO 25964: the evolution of thesaurus Standards towards Interoperability and data modeling. ISQ Information Standards Quarterly 24(1): 20-26; <http://www.niso.org/publications/isq/2012/v24no1/clarke/>

[3] <http://www.w3.org/TR/skos-primer/>

[4] <http://www.w3.org/TR/rdf-primer/>

# Appendix 3

These are examples of how others have expressed controlled vocabularies in RDF/SKOS/OWL.

[EOT] See, for example, the EU announcement [1]. In this example, the controlled values for asset descriptions are enumerated under the property [2]. This seems O.K. for a small number of values but may become a bit unwieldy if you needed to include all the world’s languages.

For another, more comprehensive variation, see the controlled vocabularies for EU languages [3] countries [4]. This appears to include a lot of versioning information.

Another example from US Library of Congress; For ISO languages and countries we could just use something like [5]/[6] and [7]/[8] or create our own version. What's interesting here is that they don't point out to an ISO namespace (like we did).

[1] <http://id.loc.gov/vocabulary/iso639-1.skos.rdf>

[2] <http://id.loc.gov/vocabulary/iso639-1/da.skos.rdf>

[3] <http://id.loc.gov/vocabulary/countries.skos.rdf>

[4] <http://id.loc.gov/vocabulary/countries/dk.skos.rdf>

[5] <https://joinup.ec.europa.eu/asset/adms/document/skos-used-publish-controlled-vocabularies-defined-adms-web>

[6] <https://joinup.ec.europa.eu/svn/adms/ADMS_v1.00/ADMS_SKOS_v1.00.rdf>

[7] <https://open-data.europa.eu/en/data/dataset/language>

[8] <https://open-data.europa.eu/en/data/dataset/country>

[TR] some further examples of controlled vocabulary expression:

1. using a custom schema and owl <http://downlode.org/Code/RDF/ISO-3166/>
2. using [genomes.org](http://genomes.org) for the country class <http://datahub.io/dataset/iso-3166-2-data>
1. <http://rs.gbif.org/core/dwc_occurrence.xml> [↑](#footnote-ref-1)