Global Biodiversity Information Facility

www.gbif.org

PROJECT SCOPING FOR DWC-ARCHIVE METAFILE CREATION WEB APPLICATION

Project Scoping

A. General Information

Project Title:	MetaMaker web application for creation of a XML Metafile to support text-based publication of DarwinCore Archive files.	***Project Working Title:***	Metamaker
Secretariat	GBIFS	***Proponent Group:***	ECAT
Version	Description	***Change By***	Date
1.0	Initial scoping draft	David Remsen	1 June 2010

GBIF MetaMaker web application initial scoping document

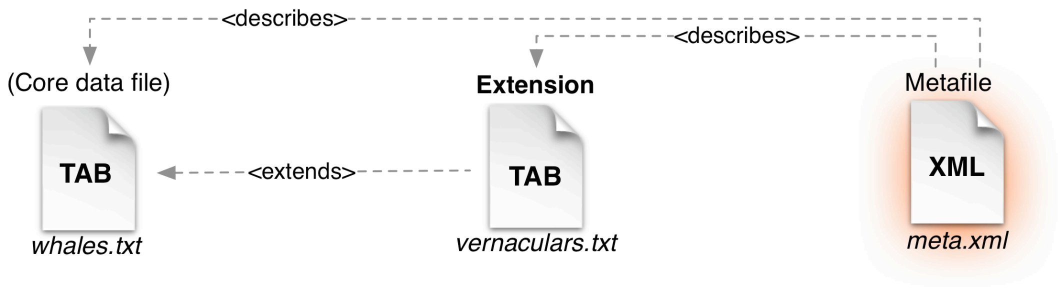
Introduction

The Darwin Core is body of standards. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. The Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, and samples, and related information.[1]

The Global Biodiversity Information Facility (GBIF) has developed a data exchange format for publishing annotated checklist data that is based on the ratified Darwin Core terms[2] and the Darwin Core text guidelines[3]. This approach provides a solution that is simple and extensible. The format uses plain text files, in a tabular format that is familiar to anyone accustomed to working with electronic spreadsheets. This simplified format, however, is linked to processes that support consistency, stability, and interoperability.

A GNA Darwin Core Archive is a set of one or more files that collectively, provide the means to publish relatively rich information commonly found in biodiversity databases that relate to occurrences of species or annotated species checklists. Data can be published according to a standard set of terms using a data format methodology described in the Darwin Core text guidelines. A key component of this methodology is the creation and use of a special data description file referred to as a *metafile*.

A ***metafile*** that serves as a directory for the data files in the archive. This file provides a standard way of listing the fields in each file and the relationships between them. A metafile is ***required*** when an archive includes any extension files or if the core data file lacks a header row containing defined field elements.



Use Case

The web application serves a need for data managers seeking simplified mechanisms for publishing biodiversity data. The current suite of data biodiversity data publishing solutions impose a number of limitations or requirements on users that affect uptake or specific application. Publication of taxonomic data to GBIF particularly, is effectively limited to the Darwin Archive format solution.

The DarwinCore Archive format presents a simple text-based format that conforms to basic data export methods familiar to many data managers. Both core data and extension data can be fairly

[1] What is Darwin Core? http://rs.tdwg.org/dwc/
[2] Darwin Core Terms: A Quick Reference Guide - http://rs.tdwg.org/dwc/terms/index.htm
[3] Darwin Core Text Guidelines - http://rs.tdwg.org/dwc/terms/guides/text/index.htm

easily configured as repeatable and fast data export operations. The most complex component of the archive format is the metafile, an XML format that is not as amenable to manual creation.

The metafile, however, is an XML document that is relatively complex to manually compose for many data managers. The objective of this project is to provide a simpler and mediated mechanism to create and validate a metafile and thereby facilitate a simplified data publishing solution for the GBIF network.

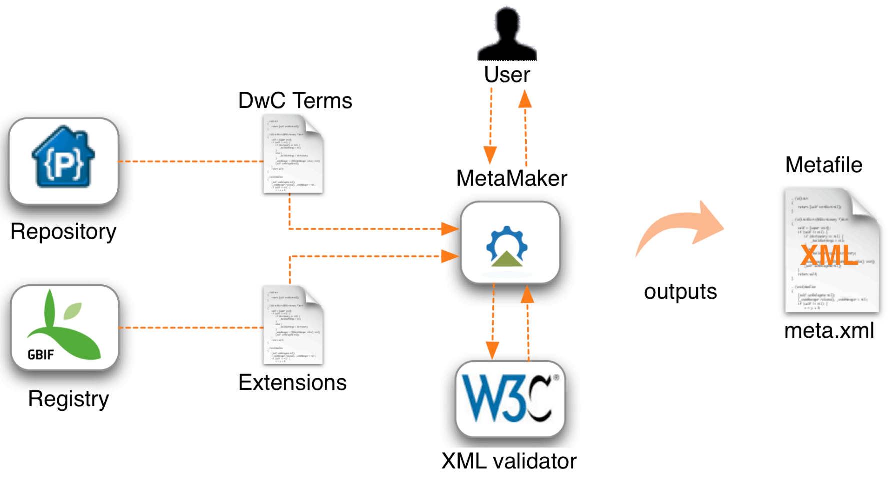
General Functionality

A general description of the functionality of the web application is as follows:

Configuration Requirements: PHP/Javascript (ExtJS)

General Functional Description

1. The web application asynchronously (relative to the following user interaction) loads data from the GBIF network.
    a. Sets of Darwin Core terms
    b. Definitions of DarwinCore Extensions.
2. The user accesses the web application
3. The user selects and configures the core data file via the web interface.
4. The user selects and configures zero or more extensions according to their requirements. During the selection and configuration process, an XML document is being composed and can be viewed at any time by selecting a tab.
5. The user can validate the resultant XML file to confirm it is properly configured.
6. The user saves a copy of the output metafile via a simple copy and paste operation or through a “Save file” option.



More specific functional requirements are listed below and organized by the major metafile components. Interpretation of these requirements is supported by some screen design figures included in this document. These figures include labels that are inserted in the descriptions below to indicate their placement in the draft design. They are referenced by their Figure number followed by an integer. Example “Figure 1.2” refers to Figure 1 with and a label “2”.

The MetaMaker Web Application

The objective of this project is to create a web application that enables a user to generate a validated metafile. The scope of this metafile is defined by the XML schema located at http://darwincore.googlecode.com/svn/trunk/text/tdwg_dwc_text.xsd and provided as an Annex to this document. In general, *this schema is used to describe the set of files* that comprise a Darwin Core Archive.

The files that are described in the metafile include:

- A *core data file* that indicates the main class of data being served in the archive, it’s physical properties, and the specific fields of data that it will contain.

- Zero or more *extension files* that extend the core, their physical properties, and the specific fields they will contain.

- An optional *resource metadata document*, referenced by a URI.

Core Data File Requirements

The *core data file* indicates the main class of data being served in the archive, it’s physical properties, and the specific fields of data that it will contain. There are two options regarding available classes of data

1. Occurrence (Collections, Species Observations)
2. Taxon (Species Checklist, Nomenclatural list, etc).

The core data file is a regularly delimited text file composed of rows and columns representing fielded text. Each class of data is linked to a set of terms that would be presented to a user who would then select the terms that represent the data fields to be exported from a source database to a text file.

The web application interface would provide the means to:

A. Select the core class of data being served (Occurrence or Taxon). A panel in the application would present the two options , perhaps using radio buttons to ensure an exclusive selection. **See Figure 1.1**

B. Once selected, present the user with the list of terms tied to the class and a mechanism to

1. Select/de-select a term indicating it will be represented in the data file. **See Figure 1.2**

2. Order/re-order the term among a list of terms to indicate its sequence from 1 to N where (N+1) is the total number of columns being published in the core data file. Position 0 (the first position) is reserved for the default term that serves as the primary key for the data file.

    a. For Taxon Class the first element is the taxonID

    b. For Occurrence Class the first element is occurrenceID

    The user should be able to move the term element manually (drag and drop) and the numeric index indicator will be re-ordered in response. **See Figure 1.2**

3. Enter static (global) values for a given term via a data entry field associated with the term.

C. Input physical attributes of the core data file that is being described. Parameters that can be set by the user include. **See Figure 1.4**

- File name/location. There are two options for a user.
  - Simple filename when the file will be included in the same location (filesystem directory) as the metafile. (Example: "whales.txt")
  - URI when the file is located on a remote web server. (Example: http://www.gbif.org/test/data/whales.txt)
- File-encoding options. An enumerated list will provide options for users.
- Field delimiter. An enumerated list will provide options for users.
- Fields enclosed by. An enumerated list will provide options for users.
- Line ending character(s). An enumerated list will provide options for users.
- A Boolean to indicate if the first row of the file contains data.

D. Add new blank rows to the list of terms that will be included in the core data file. The tabular format of the core or extension data files provided by a data publisher may legitimately contain data elements that are not included in the set of standard terms. These non-standard fields may occur as any column in the published data. This option allows a user to insert a blank row and position it at the relevant column location to account for these non-mapped columns. They will be indicated as "Unmapped data." **See Figure 1.5**

E. View additional metadata pertaining to the term. The list (table) of standard terms will be displayed with additional, read-only, metadata contained in the source file. This includes a "Data Type" attribute and a "Required" Boolean. **See Figure 1.6**

- For Required terms, the term should be automatically selected and the user not able to de-select it.

Source and format of the core data class elements

The web application requires access to a list of Darwin Core terms and their properties for each of the content classes supported in the core data file (Occurrence and Taxon data). These data will be accessible as text files in a common format, where each file represents the scope of the particular class. These files will be maintained on the project source repository.

Extension file requirements

The web application will enable users to select relevant extensions from a list of extensions that are acquired from the GBIF network[4]. The extensions will be organised and presented as resources grouped into folders that are organised by namespace. The requirements for Extensions are nearly exactly the same as those for the core data files described above with a few exceptions.

A. Multiple extensions can be selected so a checkbox selection mechanism should be implemented. Any extension that is selected will have a corresponding <extensions> element in the output metafile.

[4] Registry access to Extension list - http://gbrds.gbif.org/registry/ipt/extensions.json

B. A checked extension can be selected for configuration. This means that a selected extension can be the target of a related tabbed display that provides configuration details as are outlined in Sections B-E in the core data requirements above. It would be useful for there to be some sort of visual indicator or flag, set by the user, to indicate that an extension has been configured to help facilitate the process. This indicator would enable the user to identify selected extensions that have not gone through a configuration process.

Metadata requirements

The web application will allow a user to identify a remote or local metadata document by filename or URI. This value will be input as a parameter of the <archive> element in the output metafile. A metadata document is not required.

Output

The sole output of the MetaMaker web application will be a XML file that validates against the previously identified metafile schema. Output options include the means to select, copy and paste the contents of the XML display component of the web application as well as a standard "Save as" option that stores a copy of the document on a local filesystem.

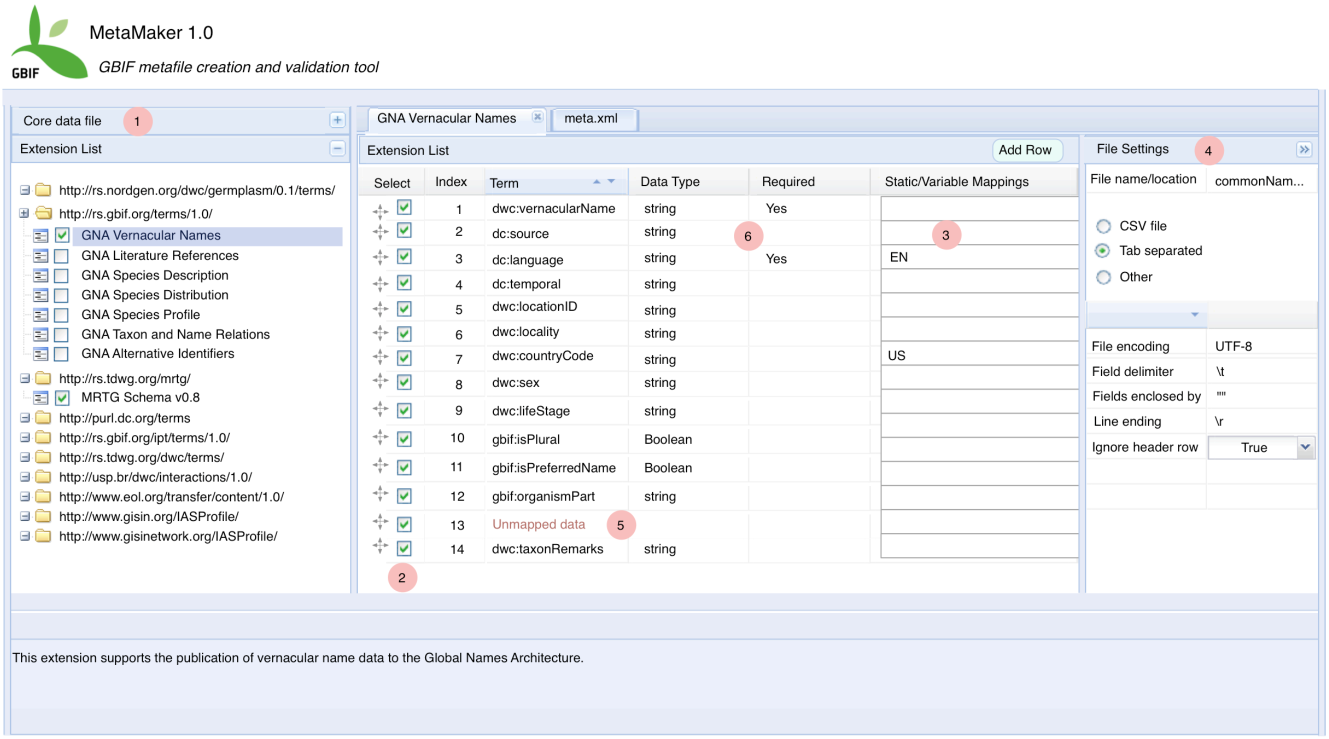
A number of public XML validation services are available:

- http://www.stg.brown.edu/service/xmlvalid/
- http://www.xmlvalidation.com/
- http://www.w3schools.com/Dom/dom_validate.asp
- http://www.validome.org/xml/

Development requirements

The development of the web application will have the following Phases

- Initial consultation to identify all required components.
    a. Identify and test input data sources
    b. Select and test a XML validation service
    c. Review and lock in wireframe/interface design and end-user interaction
- Primary Development
- Release of Alpha Candidate
    a. The purpose of this release is to assess the initial requirements and interface ideas, practice the intended workflow, etc.
    b. Verifies output in testing real uses cases among a select group of reviewers.
- Alpha evaluation and feedback
    a. This release is subject to re-evaluation of user-interfaces based on initial testing by reviewer.
- Post-Alpha Development based on feedback
- Beta Release
    a. The beta release evaluates the implementation of alpha recommendations
    b. Verifies output in testing real uses cases among participants.
- Pre-release evaluation and Testing
- Post-release installation



MetaMaker 1.0

GBIF metafile creation and validation tool

Core data file

Extension List

- http://rs.nordgen.org/dwc/germplasm/0.1/terms/
- http://rs.gbif.org/terms/1.0/
  - ☑ GNA Vernacular Names
  - □ GNA Literature References
  - □ GNA Species Description
  - □ GNA Species Distribution
  - □ GNA Species Profile
  - □ GNA Taxon and Name Relations
  - □ GNA Alternative Identifiers
- http://rs.tdwg.org/mrtg/
  - ☑ MRTG Schema v0.8
- http://purl.dc.org/terms
- http://rs.gbif.org/ipt/terms/1.0/
- http://rs.tdwg.org/dwc/terms/
- http://usp.br/dwc/interactions/1.0/
- http://www.eol.org/transfer/content/1.0/
- http://www.gisin.org/IASProfile/
- http://www.gisinetwork.org/IASProfile/

GNA Vernacular Names | meta.xml

Extension List — Add Row

Select	Index	Term	Data Type	Required	Static/Variable Mappings
☑	1	dwc:vernacularName	string	Yes	
☑	2	dc:source	string		
☑	3	dc:language	string	Yes	EN
☑	4	dc:temporal	string		
☑	5	dwc:locationID	string		
☑	6	dwc:locality	string		
☑	7	dwc:countryCode	string		US
☑	8	dwc:sex	string		
☑	9	dwc:lifeStage	string		
☑	10	gbif:isPlural	Boolean		
☑	11	gbif:isPreferredName	Boolean		
☑	12	gbif:organismPart	string		
☑	13	Unmapped data			
☑	14	dwc:taxonRemarks	string		

File Settings

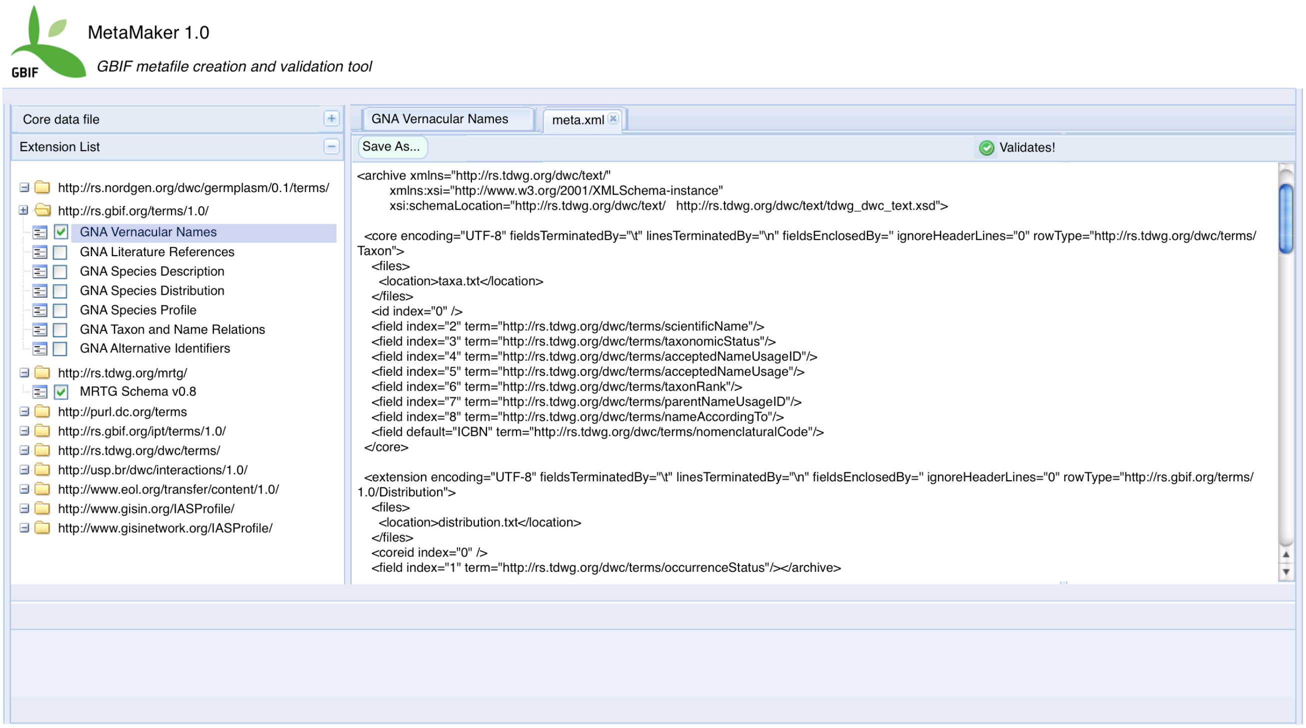
File name/location: commonNam...

○ CSV file
● Tab separated
○ Other

File encoding	UTF-8
Field delimiter	\t
Fields enclosed by	""
Line ending	\r
Ignore header row	True

This extension supports the publication of vernacular name data to the Global Names Architecture.

Figure 1 – Selection and Configuration of files



MetaMaker 1.0

GBIF metafile creation and validation tool

Core data file

Extension List

- http://rs.nordgen.org/dwc/germplasm/0.1/terms/
- http://rs.gbif.org/terms/1.0/
  - ☑ GNA Vernacular Names
  - □ GNA Literature References
  - □ GNA Species Description
  - □ GNA Species Distribution
  - □ GNA Species Profile
  - □ GNA Taxon and Name Relations
  - □ GNA Alternative Identifiers
- http://rs.tdwg.org/mrtg/
  - ☑ MRTG Schema v0.8
- http://purl.dc.org/terms
- http://rs.gbif.org/ipt/terms/1.0/
- http://rs.tdwg.org/dwc/terms/
- http://usp.br/dwc/interactions/1.0/
- http://www.eol.org/transfer/content/1.0/
- http://www.gisin.org/IASProfile/
- http://www.gisinetwork.org/IASProfile/

GNA Vernacular Names | meta.xml

Save As... — ✓ Validates!

```
<archive xmlns="http://rs.tdwg.org/dwc/text/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://rs.tdwg.org/dwc/text/   http://rs.tdwg.org/dwc/text/tdwg_dwc_text.xsd">

  <core encoding="UTF-8" fieldsTerminatedBy="\t" linesTerminatedBy="\n" fieldsEnclosedBy='' ignoreHeaderLines="0" rowType="http://rs.tdwg.org/dwc/terms/
Taxon">
    <files>
      <location>taxa.txt</location>
    </files>
    <id index="0" />
    <field index="2" term="http://rs.tdwg.org/dwc/terms/scientificName"/>
    <field index="3" term="http://rs.tdwg.org/dwc/terms/taxonomicStatus"/>
    <field index="4" term="http://rs.tdwg.org/dwc/terms/acceptedNameUsageID"/>
    <field index="5" term="http://rs.tdwg.org/dwc/terms/acceptedNameUsage"/>
    <field index="6" term="http://rs.tdwg.org/dwc/terms/taxonRank"/>
    <field index="7" term="http://rs.tdwg.org/dwc/terms/parentNameUsageID"/>
    <field index="8" term="http://rs.tdwg.org/dwc/terms/nameAccordingTo"/>
    <field default="ICBN" term="http://rs.tdwg.org/dwc/terms/nomenclaturalCode"/>
  </core>

  <extension encoding="UTF-8" fieldsTerminatedBy="\t" linesTerminatedBy="\n" fieldsEnclosedBy='' ignoreHeaderLines="0" rowType="http://rs.gbif.org/terms/
1.0/Distribution">
    <files>
      <location>distribution.txt</location>
    </files>
    <coreid index="0" />
    <field index="1" term="http://rs.tdwg.org/dwc/terms/occurrenceStatus"/></archive>
```

Figure 2 – Metafile display and validator