
# Milestone M4.33

# XML Transformations CDM export

**Leading partner**: BGBM

**Compiled by**: Lorna Morris and Sybille Bürs

**Date**: May, 31th 2013

## Introduction

The aim of this milestone is to transform an XML representation of part of a CDM database (e.g. taxon family or genus) into MediaWiki wikitext format [1]. The function of MediaWiki is to provide a publication tool that can, for example, provide long-term stable versions of the CDM content. Here we have focused on designing the structure for the taxonomic content of the wiki pages and developing an XSLT transformation to produce the wikitext sources for the MediaWiki pages. In milestone 4.37 we will expand on the work from this milestone to automate the pipeline for the CDM to MediaWiki and test the pipeline with larger datasets.

MediaWiki is free and licensed under the GNU General Public License (GPL) and as a tool that is used for creating the popular web-site Wikipedia it is likely that it will be maintained in the foreseeable future. This is an important consideration to help future-proof our data against changes in technology. Exporting data from the CDM platform to MediaWiki may serve as a perfect way to fulfil the requirement of providing stable and accessible versions within a constantly changing data environment [2].

## CDM export to XML

The generation of the XML output from CDM, can either be triggered from the user interface of the Taxonomic Editor or by directly calling a Java test class within the CDM library.

The latest version of the Taxonomic Editor is available from:

http://dev.e-taxonomy.eu/download/taxeditor/

To export XML from the Taxonomic Editor the following steps are carried out:

1. Connect to a CDM data source. For testing our pipeline we used the Flora of Central Africa.
2. Select a taxon from the Taxon navigator panel on the left. We selected Ericaceae Durande for testing purposes.
3. From the ‘General’ menu select the ‘Generate PDF’ wizard. Follow the steps in the wizard, clicking ‘Finish’ to trigger the XML export process.

The source code for the XML export and harvesting and testing is available from EDIT's subversion repository:

http://dev.e-taxonomy.eu/trac/browser/trunk/cdmlib/cdmlib-print

XML harvesting makes use of CDM library REST web services to collect the taxon and associated data (e.g. synonymy, descriptive data, polytomous keys, references) and assemble the data into an XML file.

The same XML harvesting process is used for generating a print publication and for MediaWiki publication. In each case a different stylesheet is applied to the XML output to produce the required format. The different stylesheets are also available from EDIT's subversion repository:

http://dev.e-taxonomy.eu/trac/browser/trunk/cdmlib/cdmlib-print\src\main\resources\stylesheets\mediawiki

## Transformation policies

The purpose of the export is to provide MediaWiki as a versioned publication tool for long-term availability of online published content from CDM databases. Exports from the CDM and updates to the MediaWiki pages should take place at regular intervals, for example yearly, depending on the requirements of the individual data provider.

It is not expected that data providers will edit the data directly in MediaWiki, as the focus is on publication and archiving. However, looking further into the future the MediaWiki content could be serviced as a stand-alone system therefore we ensured that appropriate formatting make the generated wiki code human readable.

For these reasons we used MediaWiki templates rather specific wikitext markup within the XSLT where possible. These templates act as semantic markup and provide their own corresponding layout that is performed in the MediaWiki when the page is shown. We created an entire set of new MediaWiki Templates all prefixed with "EDIT" that provide all the formatting/layout and semantic information for the taxonomic elements. To avoid using MediaWiki standard templates in a taxon page we created a wrapper template around the standard templates where necessary.

The advantage is that you can easily change the layout of the MediaWiki pages by changing these “EDIT” templates and do not have to change any standard templates of the wiki or redo the complete transformation workflow.

The requirement for readability forced us to include newlines within the wikitext output from the transformation. However we had to avoid any indentations or additional newlines that would be interpreted as MediaWiki layout marks by the MediaWiki parser. Figure 1 shows an example of a small MediaWiki page source.

```
{{EDIT_TOC}}{{EDIT_Taxotree| parentTaxon=Internal:Ericaceae}}

{{EDIT_Section|title=Synonomy}}

{{EDIT_Taxon|Agarista }}{{EDIT_Taxon_Author|G.Don}}
{{EDIT_Reference|name=ab2689a1-53d0-4c7c-bf33-6bed8a27e2fb|content={{aut|}} A general system of gardening and botany. 3: London,
Rivington. }}{{EDIT_Reference|name=ff11b5bd-32bb-4556-9025-8680c3ea6116|content={{aut|Judd}} A taxonomic revision of the American
species of Agarista (Ericaceae). 65: 255–342. }}{{EDIT Heterotypic Synonym|1|Agauria (A.P.DC.) Hook.f. }}

{{EDIT_Highlevel_Feature|Description}}
{{EDIT_Nested_Feature|name=Lifeform|description=Lifeform|elements=
{{EDIT_Nested_Feature_Element|Arbres, arbustes ou arbrisseaux.}}}}
{{EDIT_Nested_Feature|name=Leaves|description=Leaves|elements=
{{EDIT_Nested_Feature_Element|Feuilles alternes ou subopposées.}}}}
{{EDIT_Nested_Feature|name=Inflorescences|description=Inflorescences|elements=
{{EDIT_Nested_Feature_Element|Inflorescences axillaires ou terminales, en racèmes ou en panicules, avec bractées caduques;
pédicelles généralement à 2 bractéoles.}}}}
{{EDIT_Nested_Feature|name=Flower|description=Flower|elements=
{{EDIT_Nested_Feature_Element|Fleurs 5-mères; calice articulé sur le pédicelle, persistant, à lobes imbriqués; corolle
cylindrique à urcéolée; étamines 10, en 2 verticilles, insérées à la base de la corole, à filet aplati, à anthère aves des pores
terminaux et elliptiques; ovaire supère, 5-loculaire, à disque nectarifère; stigmate tronqué ou capité. }}}}
{{EDIT_Nested_Feature|name=Fruits|description=Fruits|elements=
{{EDIT_Nested_Feature_Element|Capsules à déhiscence irrégulière.}}}}
{{EDIT_Nested_Feature|name=Seeds|description=Seeds|elements=
{{EDIT_Nested_Feature_Element| Graines minuscules.}}}}

{{EDIT_Feature|name=Distribution|elements=
{{EDIT_Feature_Text|Genre essentiellement d'Amérique du Sud, comprenant 31 espèces dont une seule en Afrique, présente dans la
Flore.}}}}

{{EDIT_Reference_Section}}

[[Category:Agarista]]
```

**Figure 1: Screenshot of MediaWiki source code generated by XSLT transformation of CDM XML content.**

## Page Contents

The taxon pages are named with the taxon name and we also provide a parameter in the XSLT that can be modified to create a MediaWiki prefix, e.g. in our testwiki we use the parameter “Internal” to prefix the pages to keep them private.

A taxon page has the following sections:

- Table of contents (typical MediaWiki style)
- A “Taxonomy” box that shows the higher taxon and the tree of all lower taxa as links to navigate the taxonomy. (Figure 2)
- Synonymy (Figure 3)

- Keys, if the taxon relates to any (Figure 4)
- Descriptive features* including descriptive images (Figure 5)
- References (Figure 6)

* There are several feature types for descriptive data in the CDM e.g. Habitat, Vernacular name and general description (which can be divided into further feature types, e.g. Lifeform, Flowers). Each features is associated with a free text description.


**Figure 2: The Taxonomy box from the Ericaceae family Mediawiki page.**

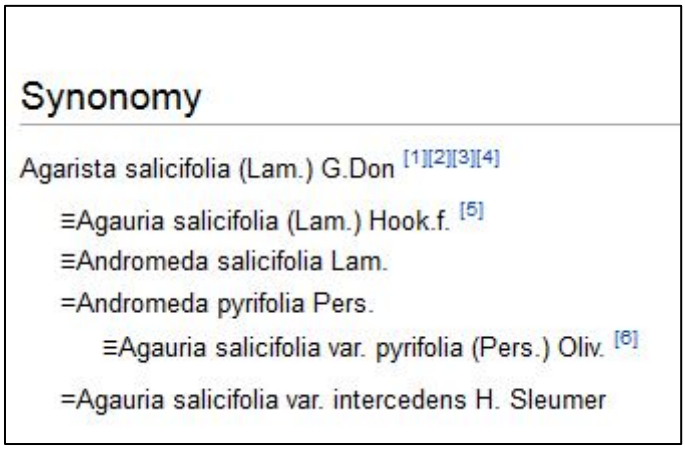



**Figure 3: The Synonymy section of the MedaWiki page.**

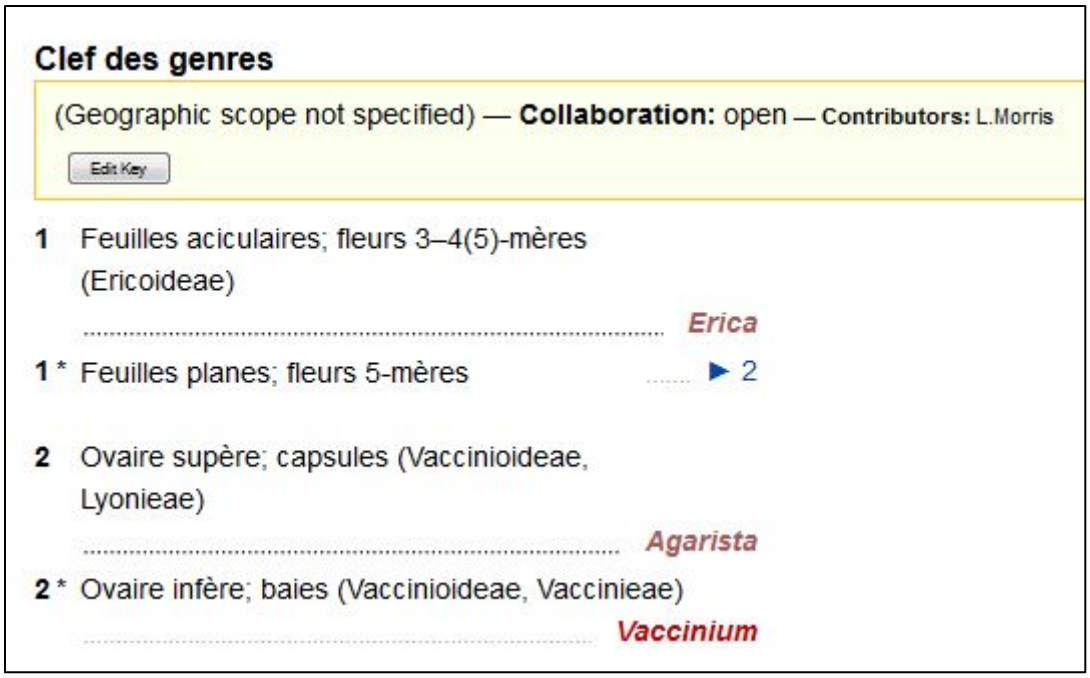



**Figure 4: The polytomous key section from the Ericaceae family MediaWiki page.**

## Description

***Lifeform:*** Arbre ou arbuste sempervirent, atteignant jusqu'à 20 m de haut; tronc pouvant atteindre un diamètre de 25 cm, à écorce grise ou brune, subéreuse, fortement crevassée et rougeâtre avec l'âge; aubier blanc; rameaux cassants, pubescents à cils simples parfois entremêlés de cils glanduleux.

***Leaves:*** Feuilles à pétiole pubescent à glabrescent, de 0,5–1 cm de long: limbe généralement étroitement elliptique à largement ovale, assez coriace, de 2–10 cm de long et 0,5–4 cm de large, à base cunéée à subcordée, à bords entiers, à sommet arrondi, mucroné ou acuminé, à face supérieure brillante et glabre, à face inférieure glauque et glabre sauf la nervure principale.

***Inflorescences:*** Inflorescences en racèmes 15–35-flores, de 5–15 cm de long; pédicelles de 2–6 cm de long, généralement pubescent; bractéoles de 1–1,5 mm de long, pubescentes.

***Flowers:*** Fleurs: calice vert suffusé de rouge, à lobes triangulaires, d'environ 2,5 mm de long, à face externe éparsément pubescente et à bords ciliés; corolle cupuliforme à urcéolée, verdâtre à jaunâtre, parfois teintée de rouge à la base, de 7–10 mm de long et 4–5 mm de diamètre, glabre, à lobes triangulaires de moins de 1 mm de long; étamines orangées, à filet pubescent; ovaire globuleux, d'environ 2 mm de diamètre, éparsément pubescent; style de 6–8 mm de long, persistant, glabre; stigmate capité.

***Fruits:*** Capsules vert foncé à rougeâtres, de 4–7 mm de diamètre.

***Seeds:*** Graines jaunes.

***Figures:***

Figure 1 — *Agarista salicifolia*: A, rameau; B, fleur; C, coupe longitudinale de la fleur; D, face interne de [...]

## Distribution

R.D.Congo: IX: Bequaert 6147; Chapin 457; Christiaensen 2435; Deru 468; Devred 3762; Donis 3922; Ern 75; Germain 3277; Ghesquière 5069; Hendrickx 3222; Humbert 7619; Keremera 6; Lebrun 4710, 5413, 9420; A. Léonard 350, 3574; Michel 5589; Michelson 983; Pierlot 428, 538, 2051, 2560, 2834; Rossignol 186; Stauffer 349. XI: Desenfans 4097; de Witte 2550, 3869; Lisowski, Malaisse & Symoens 3585, 5613, 12710; Malaisse 4191; Quarré 5548; Schmitz 6547, 7779. Rwanda: IX: Auquier 2636; Bouxin 208, 509, 742; Bridson 344; Mildbraed 1024; Renier 235; Reynders 128; Troupin 11513, 13238. Burundi: IX: Christiaensen 2393; Claessens s.n.; Lewalle 1616; Reekmans 4437, 8547; Robyns 2307. X: Declerck 83, Lewalle 597; Reekmans 9364. Cameroun, Bioko, Uganda, Kenya, Tanzanie, Angola, Zambie, Malawi, Mozambique, Madagascar, Réunion, Maurice.

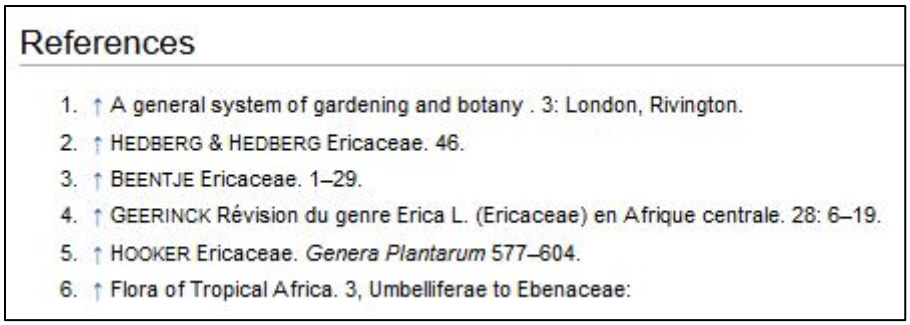
## Habitat

Forêts-galeries, forêts de montagne; lisières forestières; marais broussailleux, savanes arbustives; formation de bruyères, bambousaies; cratères et plaines de lave; entre 1400 et 3000 m d'altitude.

## Vernacular Name

Tshihondo (Mashi); Mukarakara (Kinande); Kijojo (Kifulero); Nyabafumbwe (Mashi); Kidjodjo (Kinyindu); Umukarakara (Kinyarwanda); Kishushuti (Kinyarwanda); Kishasha (Kitembo); Umutandura (Kinyarwanda); Gishusha (Kihunde); Kihomba (Mashi); Mushushuti (Kinande); Igishushuti (Kirundi); Igiwundwa (Kirundi); Kitanduli (Kinyarwanda); Kashasha (Mashi); Tshiniabawubwe (Kinyabongo)

**Figure 5: An example of the descriptive section from the Agarista salicifolia MediaWiki page, showing a thumbnail link to an image file.**



References

1. ↑ A general system of gardening and botany . 3: London, Rivington.
2. ↑ HEDBERG & HEDBERG Ericaceae. 46.
3. ↑ BEENTJE Ericaceae. 1–29.
4. ↑ GEERINCK Révision du genre Erica L. (Ericaceae) en Afrique centrale. 28: 6–19.
5. ↑ HOOKER Ericaceae. *Genera Plantarum* 577–604.
6. ↑ Flora of Tropical Africa. 3, Umbelliferae to Ebenaceae:

**Figure 6: The reference section from the Agarista salicifolia MediaWiki page.**

## Navigation between Taxa in the Taxon tree

MediaWiki pages can be placed in categories to enable related pages to be grouped [5]. Categories may belong to other categories in a hierarchy, thus making use of categories provided us with a useful way of navigating between higher and lower taxa in the taxonomic hierarchy. An advantage of adopting this practice is that the relations between categories in MediaWiki are present automatically. We used the MediaWiki parser function, to provide the tree of lower taxa in the taxonomy box.

Each wiki page can be assigned to one or more categories. However categories are currently only used to provide the navigation functionality in the taxonomic hierarchy, although in future implementations non-taxonomic categories to classify taxa (e.g. “Redlist”) could be added.

### Implementaion

We created a MediaWiki category for each taxon. These categories build up the taxonomic relations (a lower taxon category is assigned to its higher taxon category). Each taxon page is attached to its corresponding category. The categories have no content except for a redirect link to the corresponding taxon page. On the taxon pages we present a “Taxonomy” box. It has a link to the higher taxon as well as a tree presenting the complete lower taxonomy. The tree is created using the parser function *categorytree*.

## Workflow

The following steps are performed in succession:

1. Using the CDM export services we create an XML file that contains all data.
2. The XSLT transformation is applied to the XML file using the Saxon XSLT engine.
3. The XSLT transformation generates a single output file containing all the MediaWiki pages and categories. These are embedded within an XML wrapper which contains tags that provide metadata for a wiki import. The specific MediaWiki templates used by the pages are kept in a different file.
4. Import of the XML output file and the templates file into a MediaWiki web-site completes the process [3].

## Further work

For the next milestone (M4.37) we plan to optimise the export of data from the CDM and automate the entire workflow. For example a CDM service that triggers CDM export from a selected CDM data source, transforms it to wikitext format and imports the file into a MediaWiki. We will also run the workflow on a larger dataset.

Although we do not show the full taxonomic tree in the Taxonomy box, the information is present in the system and navigation up and down the tree is possible manually. Depending on user feedback we could modify the implementation to show the entire tree if required.

We will also extend the content in the MediaWiki pages, for example use the REST map web services to generate maps showing the geographical location of specimens [6].

## References


1. http://www.mediawiki.org/wiki/Manual:What_is_MediaWiki%3F
2. http://www.pensoft.net/journals/zookeys/article/2166/
3. http://www.mediawiki.org/wiki/Manual:Importing_XML_dumps
4. http://www.mediawiki.org/wiki/Help:Templates
5. http://www.mediawiki.org/wiki/Help:Categories
6. http://dev.e-taxonomy.eu/trac/wiki/MapRestServiceApi