Indexing for Life

Deliverable D4.1: Enhanced CoL Download Service

Workpackage 4

Ruud Altenburg, Wouter Addink, Peter Schalk

31 May 2011

Capacities Programme of Framework 7: EC e-Infrastructure Programme – Virtual Research Communities - INFRA-2010-2

Grant Agreement No:	261555
Project Co-ordinator:	Dr Alastair Culham
Project Homepage:	http://www.i4Life.eu
Duration of Project:	36 months
Start Date:	November 2010
End Date:	November 2013

Objective

The objective for this deliverable was to create an enhanced Catalogue of Life Download service to serve the global programme partners in the project with Catalogue of Life data in a standardized, structured and easy to use format; optimized for their needs. The service has been implemented according to the specification made in Work Package 2, see Deliverable D2.1 Download and Piping Tools Specifications. The service is available at: http://dev.4d4life.eu/dca_export/.

Implementation route followed

Ideas for implementations were presented and discussed at the first i4Life project meeting in Reading. This resulted in decisions made on the service protocol, the data format (DarwinCore Archive), on full and incremental downloads, and on credit requirements.

Following the meeting, a draft specification was produced under Work Package 2 in collaboration with the University of Reading. This was discussed at the next i4Life meeting at EBI, Cambridge in March. The download format requires harmonisation of the DwC-A structure with the global partner's organisation data needs. To facilitate this, a workgroup was established for this purpose. The workgroup is led by Work Package 4 and will be active in May and June 2011. A selection of partners in the project participates in the workgroup:

- W Addink, ETI
- S Riviere, EBI
- D Remsen and M Doering, GBIF
- V Robert, CBS
- J Ragle, IUCN
- N Hoffman, BGBM
- V Didziulis, UR
- A Jones, CU
- M Sitko, UR

A prototype of the download service was created in collaboration with D Remsen, GBIF. The prototype was demonstrated at the University of Reading in Work Package 2. The discussions resulted in a more complete specification for the download service which is described in deliverable D2.1. Based on this specification the first version of the download service was created, and made operational on May 30, 2011. It is expected that the download format established by the workgroup may require future changes once the tool is rolled out across the partner network. Hence, updates and enhancements of the download services are planned later in year 1. During the implementation, the specification in D2.1 was updated (updated documents are made available through the project's website).

For the implementation and documentation of the download service, the development infrastructure created in the 4D4Life project was used. Code is stored in the central CoL software repository (location: svn://dev.4d4life.eu/DCA_Export/branches/0.3). For easy communication between members of the download format workgroup a mailing list has been established and is in use (i4lifewp2_dwc@eti.uva.nl).

Short description of the service

The service can be used to download data from the Catalogue of Life in DarwinCore Archive (DwC-A) format in UTF8 (see http://rs.tdwg.org/dwc/terms/guides/text/index.htm). It is an online service with a web-interface. Users can either download the whole CoL dataset, or download a selection. The whole dataset (first option) is already created in advance since this takes several hours to create. The service contains an option to create this file by the Secretariat. Selections of data (second option) are created dynamically ('on the fly'). Users can monitor progress when the download is being created. For a large selection this may take several minutes, for a small selection this can be a few seconds. The service uses a webservice in the background for compatibility with the services-based E-2 infrastructure.

A selection can be made for 1.) a 'horizontal' slice from the dataset (select all data from all source databases for a taxon group, like all data for taxa in family Aves). The selection can also be 2.) 'vertical': download only taxon data 'block' and distribution data 'block' but no references data. A description of all supported blocks can be found in the specification in D2.1. Combination of 'horizontal' and 'vertical' selection is also possible.

This advanced selection mechanism allows partners to select exactly the part of CoL data they need, and in their preferred format. The data in the generated download files are compliant with the 'three levels of credit' used in the Catalogue of Life: information is included about the full product name, name of the contributing database and, when available, the latest taxonomic scrutiny.