Indexing for Life

Deliverable D4.2: Production versions of the CoL Piping and Cross-mapping Service & Tools

Workpackage 4

Wouter Addink, Ayco Holleman, Peter Schalk

24 June 2013

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

Introduction

The objective for this deliverable was to create robust, sustainable and documented versions from the Catalogue of Life (CoL) Cross-mapping and Piping services and tools that have been created in WP11 and WP12. The production versions are available for testing at:

- http://dev.4d4life.eu:8085/CrossMappingPortlet/ (Cross-mapping GUI)
- http://dev.4d4life.eu:8085/CrossMapping/XMapService?wsdl (Cross-mapping Service)
- http://dev.4d4life.eu/piping-tool/php/ (Piping tool GUI)
- http://dev.4d4life.eu/piping-tool/webservice/ (Piping tool webservice)

(login with user admin/ password admin)

The versions currently in use can be found at http://www.i4life.eu/i4lifewebsite/run-services/

Code and code documentation are stored in the CoL Subversion software repository at svn://dev.4d4life.eu. Specification documents can be found at the i4Life website at http://www.i4life.eu/i4lifewebsite/projectdocuments/.

Short description of the Piping services production version

The Piping services consist of several components or 'pipelines', these are taxon and name placement services. The services also include tools for processing and monitoring. The components developed are:

i) Taxon & Name Supply-side Pipeline (developed in task 12.1)
A protocol and procedure for pulling a data set containing a particular batch of supplied name-records and taxon-records thought to be absent from the CoL, with their associated higher taxa tags, into the 'reception buffer', ready for piping the data to the Global Species Database (GSD) providers in the CoL network. The reception buffer is a database warehouse containing the datasets and annotations that are piped between the Global partners, GSD providers and CoL assembly.

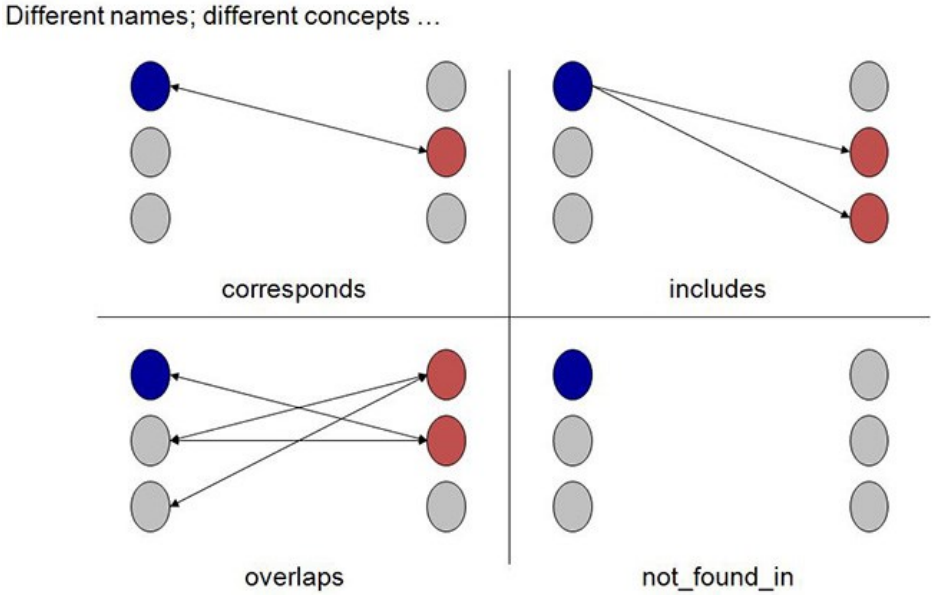
ii) Taxon & Name Distribution-side Pipeline (developed in task 12.2)
A tool set with services for serving supplied Names & Taxa from the CoL reception buffer to the appropriate GSD system, and for managing this process.

iii) Processing Monitor (developed in task 12.3)
A protocol and service for a) partitioning and distributing incoming data elements between the GSD-sectors maintained taxonomically by different GSD organisations within the CoL GSD network, and b) making the records of successful and rejected name and taxon positioning tasks available for collection by the instigator.

Workflows, use cases and implementation are described in detail in the Piping Tools Overview document available at http://www.i4life.eu/i4lifewebsite/projectdocuments/.

Short description of the Cross-mapping services production version

The cross-mapping services allow to compare two checklists and calculate the difference between them: taxa present in one checklist, but not in the other, or a partial overlap (see figure).



A “cross-map” will enable the relationships between lists of species and other taxa in one species information system to be related to those in another species information system. This is a fundamental element of making it possible to perform data aggregation and complex analyses which require the use of data from multiple, diverse species information systems.

The primary relationships to be maintained in a cross-map are relationships between the taxa, and not between the individual names. This means that, given a scientific name, the taxon that it is associated with can be determined in one taxonomy, and the corresponding taxon/taxa in another can be provided by the cross-map. The cross-mapping services are described in detail in Deliverable 2.2 (Enhanced Cross-Map Service Specification).

Created cross-maps can be viewed online or exported as a zip archive with CSV files. The zip archive can be processed further by the Piping services.

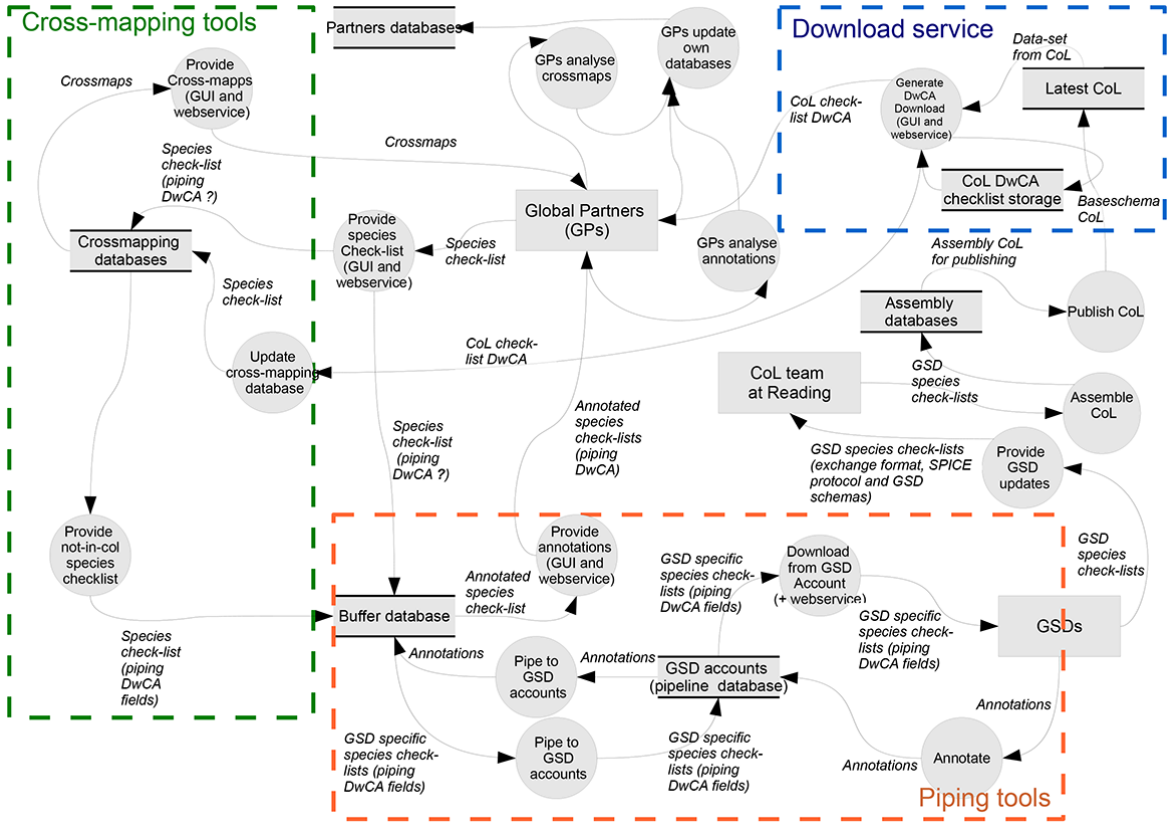
The Cross-mapping services consist of two components:

i) A backend responsible for storing the checklists, calculating a cross-map using Taxonmatcher algorithms and generating a report of the cross-map operation. The backend is exposed through a SOAP web service.
ii) A GUI implemented as a web application based on the Google Web Toolkit.

The cross-mapping services follow the Service Oriented Architecture (SOA) paradigm, with the web application never accessing the backend directly – only through the SOAP service. Any other application can call the service just as well to carry out cross-map operations for its own purposes.

Role of the Piping and Cross-mapping services

One of the major aims of the i4Life project is to establish a set of workflows for data sharing and synergistic cooperation among major biodiversity programmes around the world. The piping and cross-mapping services, together with the download service, form the information system that supports the workflows for this virtual research community.



The cross-mapping services allow global partners to compare their name-list against the CoL or checklists from other partners. The piping services pipe the cross-mapping results to the Global Species Databases (GSDs) in the CoL network for evaluation (annotated data can be piped back to the global partners) and adding missing taxa to the GSDs. This way, taxonomic elements in the global partners' checklist not presently in the CoL will be added and placed within the CoL. This makes the CoL a more useful resource for the global partner. Global partners use the CoL as a 'taxonomic backbone' for their services. Taxonomic elements in global partners' checklists entering the CoL through these workflows will also become available to the other partners using the CoL – thus raising the level of information available to all.

Work done to make the services ready for production

All code created for the services was collected and put under version control in the CoL SVN software repository located at ETI. Steps for installation, software and hardware requirements and dependencies were identified in collaboration with the developers and documented in installation documents. 'Fresh' installations of crossmapper and piping tools were performed to test the installation procedure and installation documentation was fine-tuned. Code was reviewed and code documentation was improved and completed.

Functionality of the piping services was tested and revealed some issues that are currently being fixed by the group at Reading University; there was no missing functionality. The updated version

will be used for production. Identified bugs, together with required code improvements, were discussed with the developers. It was noted that the programming method (Procedural PHP) chosen for the code base of the piping tools is not suitable for automated testing or generation of code documentation other than the provided inline code documentation. Since it is unlikely that the Piping service code will become open source or will get major future code additions and since the number of lines of code is fairly small, this is not seen as a major issue that would require rewriting of the code. The inline code documentation is sufficient for future maintenance of the tool.

Functionality of the cross-mapping services was tested (V5 of the PHP Cross-mapper with Perl Taxon matcher), which is available for test at http://dev.4d4life.eu/crossmapping-services/ (login admin/admin), user guide at http://i4life.cs.cf.ac.uk/xmappingDocuments/UserGuide.pdf) and it was concluded that major changes and enhancements were needed for production and future maintenance, also taking into account the foreseen future use of the cross-mapping services outside the i4Life project. A report with findings was created and discussed with Cardiff University (responsible for the Cross-mapper implementation) and the project leader (University of Reading). It was decided that for the production version of the Cross-mapper a new code base was going to be used (implemented in Java and already under development based on experiences with the PHP cross-mapper service), developed as Service Oriented Architecture (SOA) with a Soap service as back-end and improved user interface. This version will also be used in the OpenBio project, where it will be embedded in a GCube cloud environment, so CoL will only have to maintain one code base for the cross-mapper service.

The cross-maps created by the Java production version were compared with the cross-maps created by the earlier PHP version (taxon matcher algorithms implemented were the same, but ported from Perl to Java), discussed with developers at Cardiff University and fixed to make them produce comparable results (except that the production version creates bi-directional cross-maps, not only a crossmap from one checklist to the other). Installation steps, software and hardware requirements and dependencies were identified and documented. Maven builds for deployment into production and for future development and maintenance in Eclipse IDE were ironed out. Code documentation (javadoc) was generated and published (http://dev.4d4life.eu/crossmapper_java/javadoc/).

Both Piping services and Cross-mapping services were deployed at a development server at Naturalis for maintenance and future development and testing. Naturalis will be the new host for the ICT infrastructure for CoL production in the near future (end of 2013).

Plans for future enhancements

Two possible enhancements were identified:

- Workflow enhancement: allow DwCA provider services from global partners (when existing) to be registered in the crossmapper, so that manual upload of checklists for comparison is no longer necessary.
- Workflow enhancement: allow cross-map consumer services to be registered in the crossmapper, so that manual downloads of crossmaps are no longer necessary for those consumers. This could especially be used to integrate the cross-mapping services with the piping tool services without tying them to each other. The piping tool services currently are the prime consumer of crossmap results.

In the OpenBio project, the crossmapping service and piping service will be tested for other use cases then in i4Life, where cross-maps are created by other users than the global partners. For instance, it might be useful for a GSD provider to cross-map their GSD checklist to another checklist (not necessarily the CoL) and have the results sent to that checklist. Results of these tests and making the tools available for new users may lead to new enhancements in the future.