Indexing for Life

D2.7 Taxonomic Integration: Integration of combined species catalogues across the Global Organisations

Work package 2

Alastair Culham

With contributions from Craig Hilton-Taylor and Andrew Jones

31 October 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:	1 November 2010
End Date:	31 October 2013

D2.7 Taxonomic Integration: Integration of combined species catalogues across the Global Organisations.

Introduction

The primary aim of the i4Life project under WP2 was to establish and integrate a Virtual Research Community that would enhance and unify taxonomic enumeration and recording of the global biota among the major global biodiversity programmes, building on the Catalogue of Life (CoL) as a common platform. This taxonomic integration has been achieved through the development of access to regularly updated versions of the CoL that can be integrated into the search and data delivery of our global partners and through the cross mapping and piping of names between the CoL and those partners and further through upload of the complete CoL in Darwin Core Archive format. It was recognised from the beginning of the project that this taxonomic integration was to be through the provision of the CoL taxonomy alongside the existing systems used by our partners such that it would enhance their capabilities and not generate competition with any pre-existing classification systems they used. This made the inclusion of the CoL taxonomy politically acceptable to organisations and is also central to the approach of the CoL, that the system presented is a reference system that does not claim to be the only correct taxonomic system for any set of species.

The initial phase of this deliverable was achieved through scoping discussions among i4Life partners on the needs of the Global Biodiversity Programmes (GBPs) and of the CoL. Together the CoL and GBPs are the Global Organisations. Much of the work of integrating the CoL taxonomy into those other data portals is already reported in their separate deliverables. What was evident was that each GBP had a slightly different bespoke need for the use of the CoL ranging from a way of reducing null search results (IUCN Red List) to providing a backbone taxonomy for the data provided (GBIF). Presented here is an overview of the workflow that allows that taxonomic integration, with examples to illustrate how this integrated system is delivered.

The i4Life workflow now has the GBPs integrated into its data flow as part of the iterative process of review used to build the CoL (Figure 1). This iterative process of data flow from the GBPs through the Cross-mapper to the piping tool, or direct to the piping tool, and then via the Global Species Databases (GSDs) to the CoL and again back to the GBPs gives all partners a mutual interest in sharing data that allow taxonomic choice to be made in the development of species lists. It also allows annotation of records that are rejected from CoL and incorporation of new records that are to be integrated into the taxonomic system of the CoL. This process is the core function of the virtual research community.

The initial functioning of the Cross-mapping tool, Piping tool and Download tool are described in their appropriate deliverables (D4.1 Download tool, D11.3 Cross-mapping tool, D12.1 Piping tool [Placement service]) available on the i4Life web site. The production versions of the Cross-mapping and Piping tools are described in D4.2 and so will not be discussed in more detail here. These tools and their role has been publicised via the CoL blog (http://blog.catalogueoflife.org/).

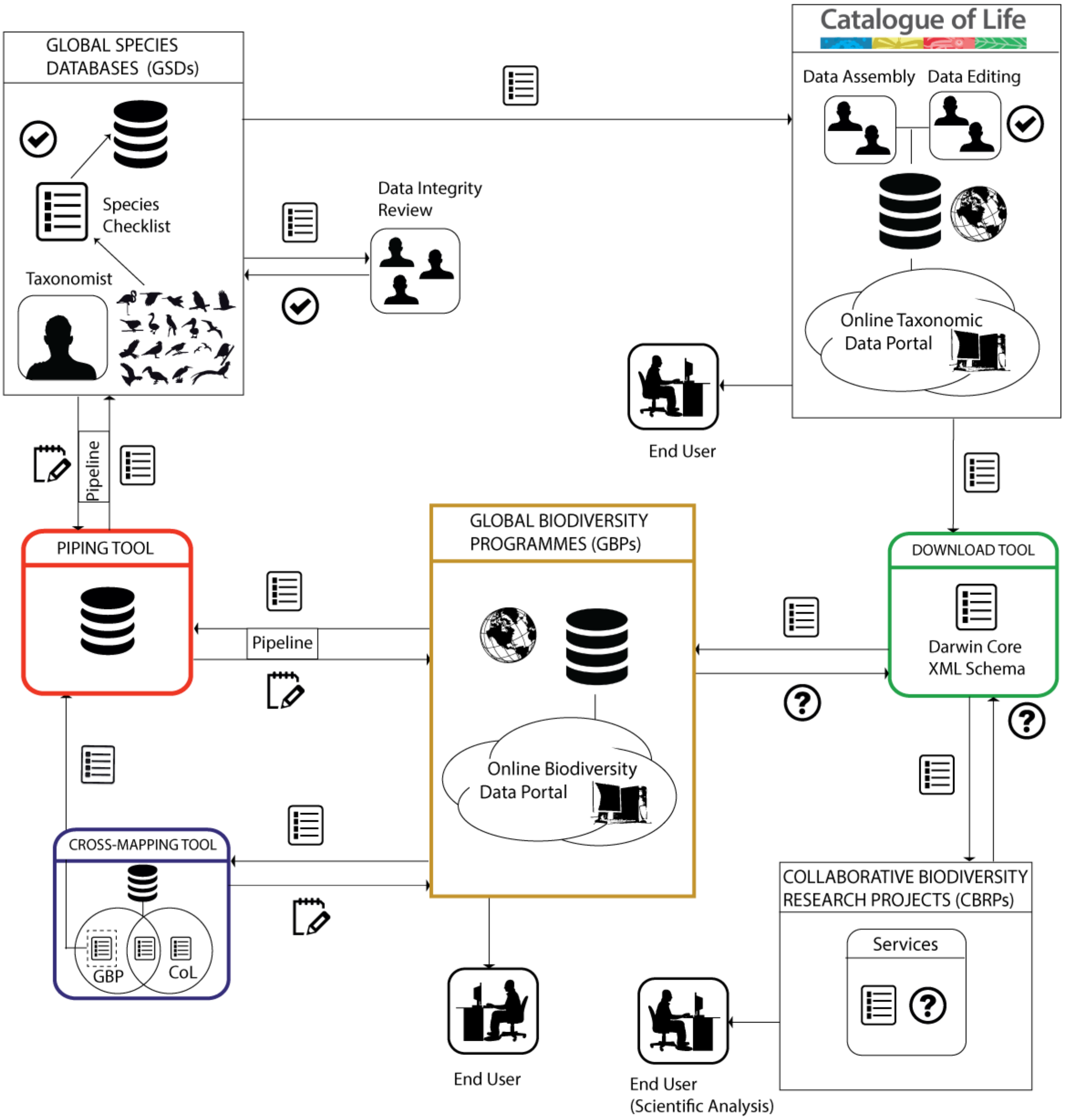
GLOBAL SPECIES DATABASES (GSDs)
Species Checklist
Taxonomist
Data Integrity Review
Catalogue of Life
Data Assembly
Data Editing
Online Taxonomic Data Portal
End User
Pipeline
PIPING TOOL
GLOBAL BIODIVERSITY PROGRAMMES (GBPs)
Pipeline
Online Biodiversity Data Portal
DOWNLOAD TOOL
Darwin Core XML Schema
CROSS-MAPPING TOOL
GBP
CoL
End User
End User (Scientific Analysis)
COLLABORATIVE BIODIVERSITY RESEARCH PROJECTS (CBRPs)
Services

Figure 1 – Data flow in the Catalogue of Life/i4Life system

Development of name placement processes

The CoL is an edited metadatabase now comprising 140 contributing databases edited by some 3000 scientists around the world. Integration of the GBPs with CoL, and processing via the cross-mapping and piping tools, generated a highly specific 'wants list' of new names for inclusion in CoL for the first time. A test process to ascertain whether GSDs could handle such processed name data within a moderately short time (months rather than years) was established in WP5, where a series of database editors had the opportunity to annotate spread-sheets of unplaced names from GBPs and to incorporate new names into their own databases while rejecting (through annotation) names that were inappropriate or erroneous. The Pilot Placement project process is reported in D3.4 and D5.1. This element of the virtual research community (GSDs) was pre-existing and is a key element to the functioning of the CoL. However, the increased automation of inclusion of new data from GSDs (both new GSD sectors and updates of GSDs) through the tools developed in 4D4Life, and enhancements by addition of further conversion templates in i4Life, has allowed inclusion of additional species at a greatly enhanced rate. It is expected the CoL will expand from the current 1.4 million species to 1.5 million before the December review meeting of this project despite CoL dipping to circa 1.3 million species last year as a result of substantial data cleansing and the withdrawal of one database. Allocation of names to GSD sectors via the piping tool, guided by the higher level (backbone) classification of the CoL, involved elements of human input because many of the unplaced data had incomplete content. However this problem will necessarily decrease as the knowledge base of annotated names (database of names that are rejected for use or included in GSDs) increases. Placement of new names on the scale used in i4Life will only happen again if several other major biodiversity portals join the virtual research community, and this is unlikely to happen because there are relatively few global biodiversity portals. There will be continuing additions from the current GBPs; the IUCN Red List, in particular, is expected to continue to develop delivery of new taxonomies and new name sets as the Red List expands. A more likely scenario is the possible addition of regional hubs such as Atlas of Living Australia, and the cross-mapping and piping placement service would prove vital tools in the handling of such regional data.

Feedback to GBPs

The pilot placement projects generated annotated names lists that are accessible to the GBPs who in turn can use them to cleanse their own data. Wrongly spelled names can be cleansed from databases and replaced with correct ones based on the annotations on the placement lists. This structured feedback after data-review has provided the GBPs with a highly valuable community edited data cleansing system that was not in place before the i4Life project.

Using CoL in the GBPs

GBIF (D6.1), ENA (D7.1), EoL, ECBOLD and Mycobank (Barcoding, D8.1), UICN (IUCN Red List, D9.1), and the Lifewatch EDIT platform (D10.1) now use CoL taxonomy through their portals, in some cases as one of several classification systems. These are fully reported in the relevant deliverables and are completed actions.

Handling Unconventional species

Two workshops were held to discuss the units recognised in DNA barcoding and in environmental genomics and to establish the degree to which these units were appropriate for incorporation into the CoL. The workshops were held back-to-back at EBI Hinxton (Genomics D2.5 & Barcoding D2.6) to make maximum use of overseas expertise for minimum cost by bringing in experts that could contribute to both workshops. After wide-ranging discussions it quickly became evident that DNA barcoding of multicellular species generated results that were highly relevant to CoL. However, the incorporation of BINs (clusters) for microbiological samples would require additional work not funded within the current project and this has been added to the list of ideas for future work but not implemented. The workshop on environmental genomics concluded that the sequence-based groups identified using these techniques were not identical to species and would require special case categories added to the taxonomic hierarchy of CoL if such data were to be included. Again it was decided that this was not the time to add such activity to i4Life but developments in this area would be monitored.

Interaction with the Global Names Architecture

A workshop held in year 1 of the project brought together international participants in the Global Names Architecture (GNA) reported under D2.3 and made a series of recommendations for future activity. Here I will deal with outcomes of the eight higher priority items among CoL and GNA.

1. The first, and highest priority action, was to use nomenclators to build proto-GSDs and this has now been demonstrated for Mollusca under WP3 as an editorial activity.
2. The second item was to link names to nomenclators. Although this has not yet been practical for the CoL overall, some GSDs (including the new set of SCaDS databases established under WP5) have this functionality built into their systems.
3. The ability of taxon specialists to annotate the output of cross maps has been implemented under WP2 by Cardiff and under WP4 by Naturalis.

4. Future funding possibilities and interactions of GNA and CoL have been discussed and Species 2000 has been invited by the GNA group in the USA to write a letter agreeing to offer support services of CoL and grant applications to NSF for further GNA developments. This has been handled by Peter Schalk as Acting Executive Secretary of Species 2000.
5. Tracking of changes in the concept behind the use of a name has proven impractical at several levels although the cross-mapper does highlight changes in use that are evident through changes in synonymy.
6. The value, or not, of persistent identifiers was the subject of considerable debate in our virtual research community and this was agreed to be a low priority item for i4Life given that the cross-mapper could readily compare differing checklists and align names dynamically. However it was agreed that persistent identifiers should remain a feature for incorporation into the CoL at some future point once a system to implement this effectively was established.
7. The subject of Communication channels between GSDs and nomenclators was agreed among i4Life partners to be an issue outside the scope of the project as CoL does not govern the activities of GSDs – these are independent databases run by groups or individuals who collaborate with CoL.
8. Provenance data is already handled at three levels in CoL under the three-level credit agreement with partners that acknowledges the CoL, the GSD, and the taxonomist(s) who scrutinised the name.

The i4Life search tool

To help monitor developments in content of CoL relative to the GBPs we developed a basic meta-search portal allowing all Global Organisations involved to be searched simultaneously as a project demonstrator. This might be used to develop a fuller search tool as part of future project developments. The i4Life search preliminary demonstrator tool is available at http://www.i4life.eu/i4search/. This work is additional to the original project plan and not included in it.

Conclusion

During the period of i4Life, WP2 has therefore facilitated both development of a virtual research community where the dataflow of CoL is open to editing and scrutiny at many stages by different elements on the Catalogue of Life and GBP community, and an actual research community that has linked and embedded discussions of project development over the GBPs and with the GNA actors. This has allowed the CoL taxonomy to be fully embedded in these systems so that plans for future developments can be made.