
**Indexing for Life**

# Deliverable 11.1: Preliminary Cross-map Service & Interactive Tools, v1

Workpackage 11

Andrew C Jones, Chang Liu, Richard J White, Alex Hardisty

31 October 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 |



## Introduction

In this document we report on the preliminary cross-map service and interactive tools delivered in Month 12 of the i4Life project. They provide users with the ability to upload and compare checklists, and browse or download the results of the comparison.

In more detail, the contents of this document are as follows:

- The *Background* section outlines the current status of cross-mapping tool development in the i4Life project, explaining why a PHP/MySQL/PERL-based cross-mapping tool has been developed and delivered alongside a Semantic Web-orientated tool which has been under development and available for considerably longer, and is potentially more flexible, but which has performance and presentation issues that need to be addressed.
- The *Delivered software* section presents a PHP/MySQL/PERL-based cross-mapping tool which has been developed in order to address immediate needs of scalable comparison of checklists, and in order to obtain feedback on user interface design, download format, etc.
- The *Future releases* section describes an enhanced version of the PHP/MySQL/PERL cross-mapping tool which is currently available in prototype form but has not been released for general use. It also identifies further developments to increase the effectiveness of this tool. Lastly it describes how we anticipate using elements of the PHP/MySQL/PERL tool in the other cross-mapping tool developed (a Semantic Web-orientated tool, described in D2.2 and the report on MS7), and how we are addressing the scalability issues in this latter tool.
- The *Conclusions* section summarises this deliverable and sources of further information.

## Background

A key aspect of the i4Life project is to achieve cross-mapping between the global partners' taxonomies. A tool for cross-mapping has been developed using Semantic Web technologies, and version 1 of this tool was described in D2.2 and the report on MS7. Version 2 was discussed at the i4Life WP2 meeting in Berlin in September 2011. Although it detects two important "taxonomic" relationships and implements a set of SOAP download-related services – and is capable of being extended to detect a wide range of taxonomic relationships – there are scalability issues associated with this approach that require further work in order to address; also some time will need to be spent working on the user interface in order to bring it into conformance with the "house style" being developed for i4Life tools.

In view of this, and in view of the need for the global partners to be able to use a cross-mapping facility for large checklists as soon as possible, it was agreed in Berlin and (in more detail) at subsequent meetings with WP2 in Cardiff and Reading, to pursue an alternative, less flexible approach which would clearly provide a scalable solution to the problem quite rapidly. It was further agreed that, given:

- the aim expressed in the Description of Work of "exploring the full extent of species diversity",
- the need for a mechanism to detect taxa in a checklist which do not occur in the Catalogue of Life, and
- the convergence among the global partners upon Darwin Core Archive as a standard,

the most important functional requirements for the delivered version (v1) of this software were to:

- accept checklists in Darwin-Core Archive format (specifically the format adopted by the Catalogue of Life),
- detect additional taxa in a checklist when compared with the Catalogue of Life (or other checklist), and
- make these additional taxa available in a download format suitable for the i4Life "piping tools" being developed in Work Package 12.

These requirements have been met; indeed the scalability of our solution has been successfully tested by running comparisons of different editions of the Catalogue of Life Annual Checklist against each other, these being somewhat larger than many of the other anticipated data sets. In these tests the execution time is variable, depending on the hardware and other factors. However, it is of the order of minutes for additional taxa detection. It is of the order of tens of minutes for detection of a larger set of relationships, not covered in the *Delivered Software* but covered in a more extensive version of the software using the same techniques, described in the *Other Developments* section. This approach can therefore provide a cross-mapping service which, although not using technology ideally suited to the detection of some of the more subtle relationships we specified in an earlier deliverable (D2.2), makes it possible for the major relationships between taxa in global partners' taxonomies to be identified.

## Delivered software

The delivered software is implemented uses Smarty[1] for template-based control of the Web presentation of the system. Most tasks are performed using the PHP language, interfacing to a MySQL database to store imported checklists, to create cross-maps, and to store operational data such as information about users and checklists. The PHP system interfaces to an enhanced version of the Taxon Matcher software in order to perform one specific task – checklist importation. Taxon Matcher has been developed primarily to support the process of LSID allocation in successive versions of the Catalogue of Life, but it also provides a checklist importation framework which is applicable to formats other than the internal database structure of the Catalogue. Taxon Matcher is implemented using PERL.

### *Visual appearance*

The visual appearance is designed to be consistent with the "house style" adopted by i4Life, and its implementation makes use of Smarty templates which, where appropriate, have been derived from the templates used for the Piping Tools interface implemented in WP12. The "home page" is illustrated in Figure 1.

### *Importing checklists*

The preferred format for checklist import is Darwin Core Archive format (specifically the profile used by the Catalogue of Life download service). Taxon Matcher has been extended in order to support this format. In addition, Taxon Matcher supports data import from the IUCN Red List, which uses a slightly different format and naming conventions for its Darwin Core Archive files, and from NCBI Taxonomy checklist data downloaded from the NCBI Web download facility. As mentioned above, editions of the Catalogue of Life Annual Checklist can also be imported directly from a copy of the database for the edition in

[1] http://www.smarty.net/

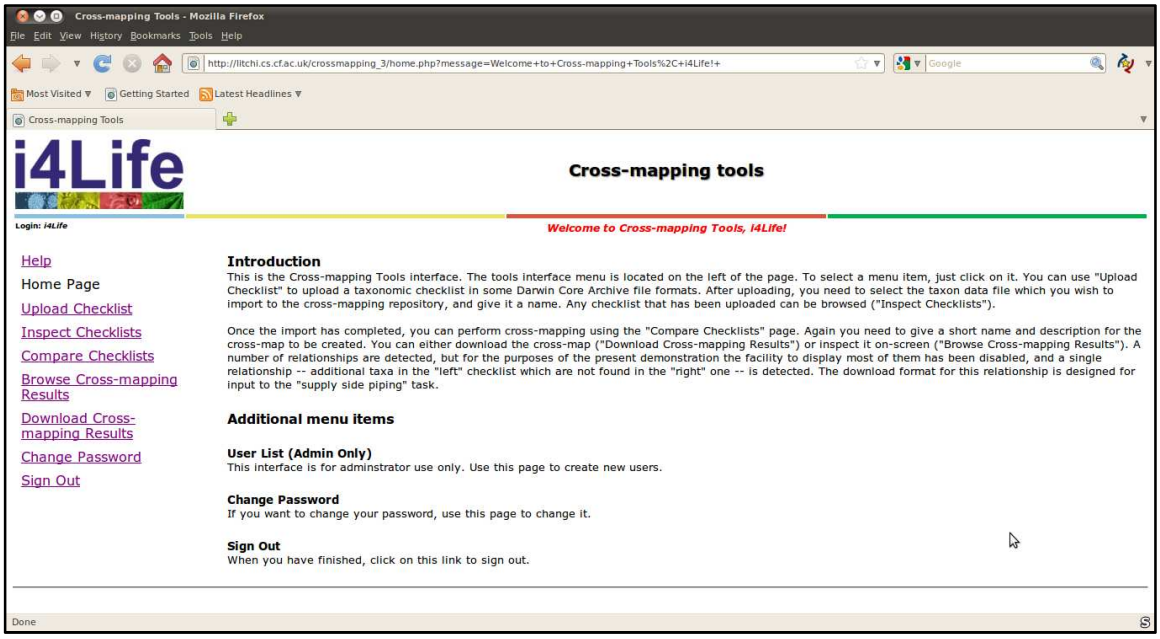



Figure 1: Cross-mapping home page

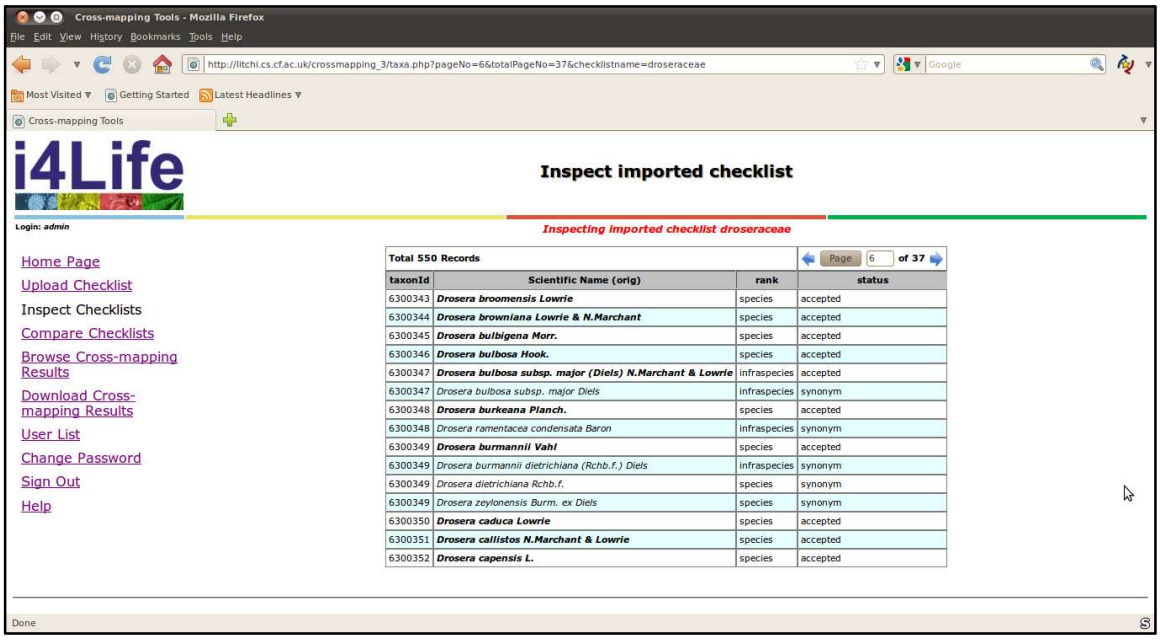



Figure 2: Checklist browsing

question. Taxon Matcher accommodates the changes in Catalogue of Life schema which have been made in successive years.

Using the PHP front-end, the user can upload a checklist, invoke the import routine and then select a checklist which has been imported and browse its contents. (Figure 2 illustrates this last stage.)

## *Cross-mapping*

The Cross-mapping tasks can be performed efficiently using a relational database management system such as MySQL, since it has been designed to support rapid processing of queries using large data sets. In the delivered software, one specific relationship is

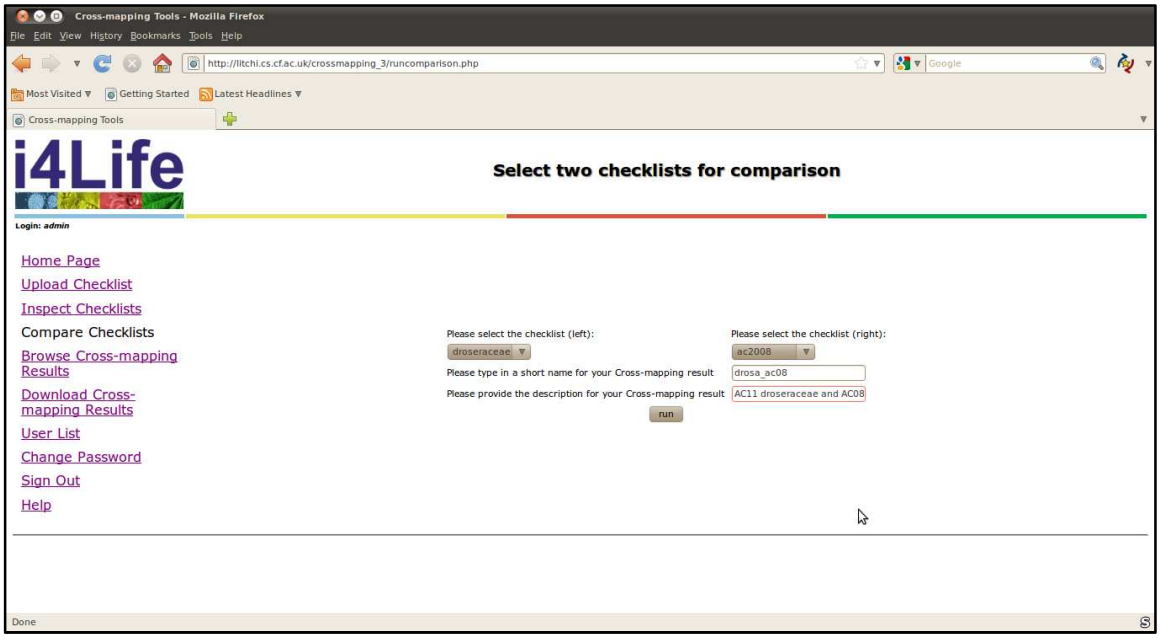



Figure 3: Creating a cross-map

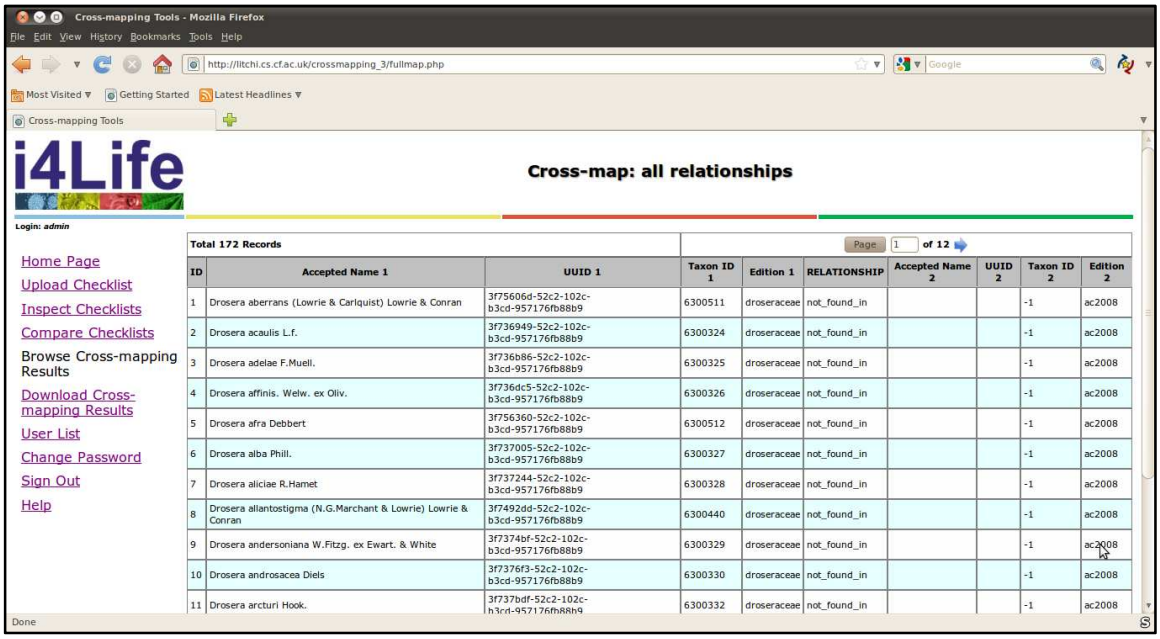
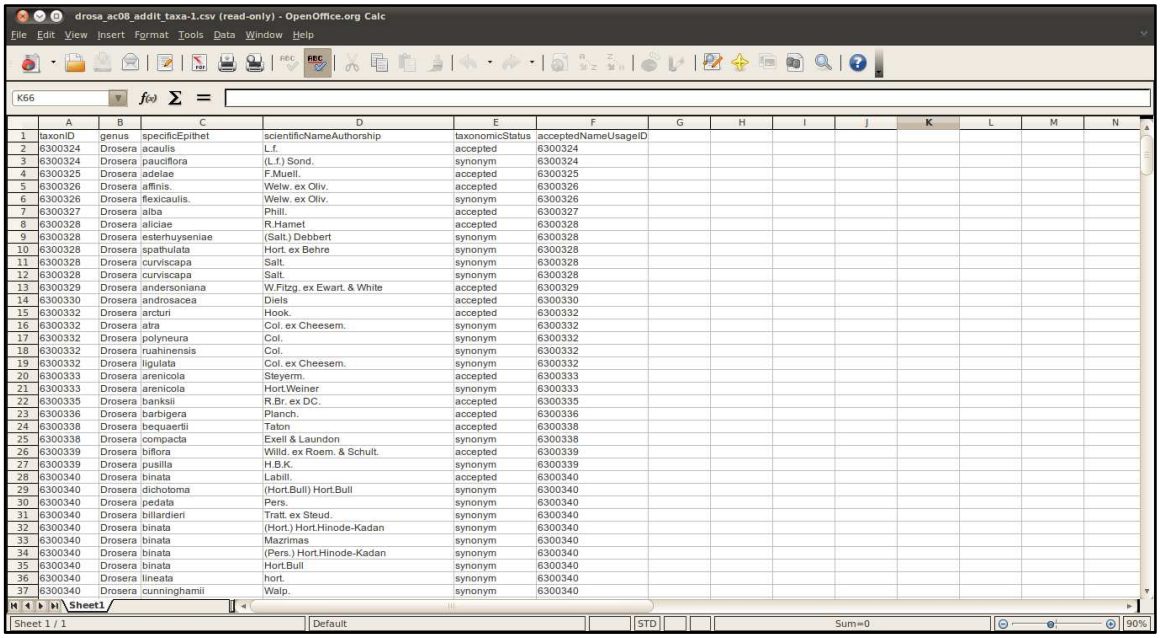



Figure 4: Browsing a cross-map

detected, namely the taxa in the “left checklist” which are not present in the “right checklist”. The criterion for this is that none of the names (accepted name or synonyms) in a particular species in the “left checklist” appear anywhere in the “right checklist”.

The user specifies which checklists (s)he wishes to compare (see Figure 3), and, the comparison having been completed, can inspect the resultant cross-map (see Figure 4). In this case, the Droseraceae in the current Catalogue of Life are being compared with those in the 2008 Catalogue of Life. (There are many differences, because the provider of data for this particular family changed in 2009.) A cross-map of additional taxa does not have any taxa on the right hand side corresponding to those on the left hand side, so an invalid taxon id of -1 is




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | taxonID | genus | specificEpithet | scientificNameAuthorship | taxonomicStatus | acceptedNameUsageID |
| 2 | 6300324 | Drosera | acaulis | L.f. | accepted | 6300324 |
| 3 | 6300324 | Drosera | pauciflora | (L.f.) Sond. | synonym | 6300324 |
| 4 | 6300325 | Drosera | adelae | F.Muell. | accepted | 6300325 |
| 5 | 6300326 | Drosera | affinis. | Welw. ex Oliv. | accepted | 6300326 |
| 6 | 6300326 | Drosera | flexicaulis. | Welw. ex Oliv. | synonym | 6300326 |
| 7 | 6300327 | Drosera | alba | Phill. | accepted | 6300327 |
| 8 | 6300328 | Drosera | aliciae | R.Hamet | accepted | 6300328 |
| 9 | 6300328 | Drosera | esterhuyseniae | (Salt.) Debbert | synonym | 6300328 |
| 10 | 6300328 | Drosera | spathulata | Hort. ex Behre | synonym | 6300328 |
| 11 | 6300328 | Drosera | curviscapa | Salt. | synonym | 6300328 |
| 12 | 6300328 | Drosera | curviscapa | Salt. | synonym | 6300328 |
| 13 | 6300329 | Drosera | andersoniana | W.Fitzg. ex Ewart. & White | accepted | 6300329 |
| 14 | 6300330 | Drosera | androsacea | Diels | accepted | 6300330 |
| 15 | 6300332 | Drosera | arcturi | Hook. | accepted | 6300332 |
| 16 | 6300332 | Drosera | atra | Col. ex Cheesem. | synonym | 6300332 |
| 17 | 6300332 | Drosera | polyneura | Col. | synonym | 6300332 |
| 18 | 6300332 | Drosera | ruahinensis | Col. | synonym | 6300332 |
| 19 | 6300332 | Drosera | ligulata | Col. ex Cheesem. | synonym | 6300332 |
| 20 | 6300333 | Drosera | arenicola | Steyerm. | accepted | 6300333 |
| 21 | 6300333 | Drosera | arenicola | Hort.Weiner | synonym | 6300333 |
| 22 | 6300335 | Drosera | banksii | R.Br. ex DC. | accepted | 6300335 |
| 23 | 6300336 | Drosera | barbigera | Planch. | accepted | 6300336 |
| 24 | 6300338 | Drosera | bequaertii | Taton | accepted | 6300338 |
| 25 | 6300338 | Drosera | compacta | Exell & Laundon | synonym | 6300338 |
| 26 | 6300339 | Drosera | biflora | Willd. ex Roem. & Schult. | accepted | 6300339 |
| 27 | 6300339 | Drosera | pusilla | H.B.K. | synonym | 6300339 |
| 28 | 6300340 | Drosera | binata | Labill. | accepted | 6300340 |
| 29 | 6300340 | Drosera | dichotoma | (Hort.Bull) Hort.Bull | synonym | 6300340 |
| 30 | 6300340 | Drosera | pedata | Pers. | synonym | 6300340 |
| 31 | 6300340 | Drosera | billardieri | Tratt. ex Steud. | synonym | 6300340 |
| 32 | 6300340 | Drosera | binata | (Hort.) Hort.Hinode-Kadan | synonym | 6300340 |
| 33 | 6300340 | Drosera | binata | Mazrimas | synonym | 6300340 |
| 34 | 6300340 | Drosera | binata | (Pers.) Hort.Hinode-Kadan | synonym | 6300340 |
| 35 | 6300340 | Drosera | binata | Hort.Bull | synonym | 6300340 |
| 36 | 6300340 | Drosera | lineata | hort. | synonym | 6300340 |
| 37 | 6300340 | Drosera | cunninghamii | Walp. | synonym | 6300340 |

Figure 5: Inspecting downloaded "Additional Taxa"

given to indicate this. When displaying a cross-map for browsing, the accepted names of the taxa involved are displayed.

### *Download*

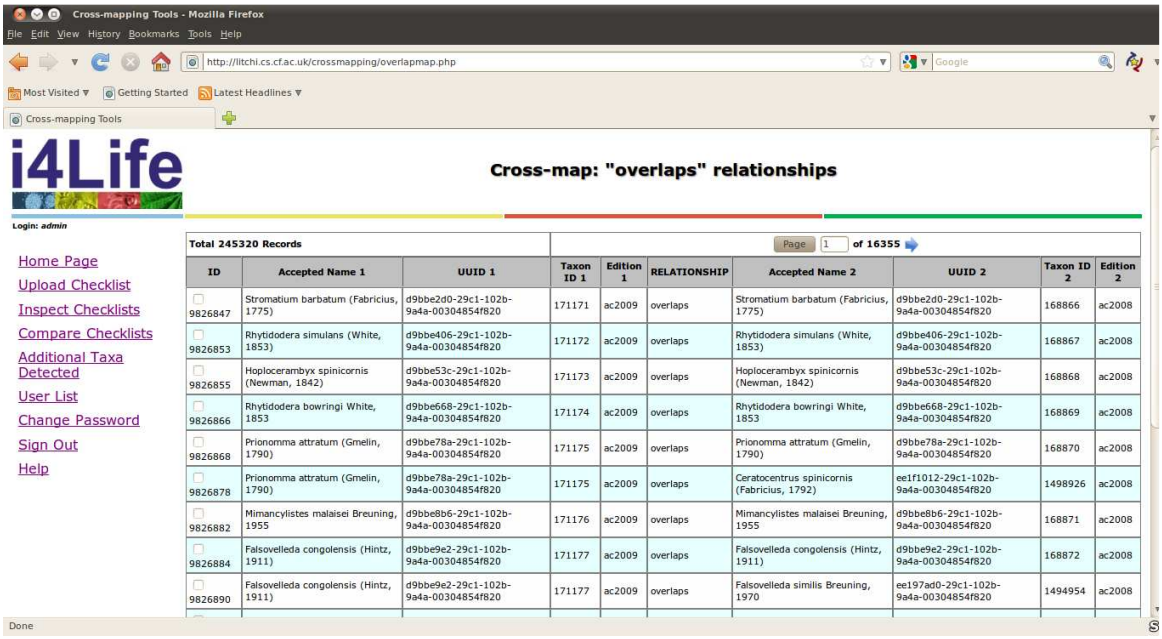
The additional taxa can be downloaded in a form suitable for input to the i4Life "Piping Tools", as a tab-separated file with fields *taxonID*, *genus*, *specificEpithet*, *scientificNameAuthorship*, *taxonomicStatus* and *acceptedNameUsageId*. All scientific names for additional taxa are included (accepted name, synonyms and other associated names for which the status is "unknown"). Figure 5 illustrates the file downloaded for the example given in earlier figures being inspected in a spreadsheet.

## Future releases

### *Enhanced prototype*

An enhanced version of the PHP/MySQL/PERL cross-mapping tool is currently available in prototype, but has not been released for general use, pending some modifications in order to accommodate a revised database schema and completion of the cross-map download facility. It identifies four relationships between pairs of taxa from different checklists (which we designate checklist A and checklist B in the following explanation):

1. *not_found_in* – the "additional taxa" which the delivered software detects.
2. *corresponds* – a one-to-one correspondence between a taxon in checklist A and a taxon in checklist B. The MySQL query identifies taxa as corresponding if the taxa have some names in common, and there is no other taxon in checklist B containing any of the names in the checklist A taxon, and there is no other taxon in checklist A containing any of the names in the checklist B taxon.
3. *includes* – a pair of taxa which participate in a one-to-many relationship between checklist A and checklist B. The MySQL query identifies a taxon in checklist A as including a taxon in checklist B if the taxa have some names in common, and there is no other taxon in checklist A containing any of the names in the checklist B taxon, but




| ID | Accepted Name 1 | UUID 1 | Taxon ID 1 | Edition 1 | RELATIONSHIP | Accepted Name 2 | UUID 2 | Taxon ID 2 | Edition 2 |
|---|---|---|---|---|---|---|---|---|---|
| 9826847 | Stromatium barbatum (Fabricius, 1775) | d9bbe2d0-29c1-102b-9a4a-00304854f820 | 171171 | ac2009 | overlaps | Stromatium barbatum (Fabricius, 1775) | d9bbe2d0-29c1-102b-9a4a-00304854f820 | 168866 | ac2008 |
| 9826853 | Rhytidodera simulans (White, 1853) | d9bbe406-29c1-102b-9a4a-00304854f820 | 171172 | ac2009 | overlaps | Rhytidodera simulans (White, 1853) | d9bbe406-29c1-102b-9a4a-00304854f820 | 168867 | ac2008 |
| 9826855 | Hoplocerambyx spinicornis (Newman, 1842) | d9bbe53c-29c1-102b-9a4a-00304854f820 | 171173 | ac2009 | overlaps | Hoplocerambyx spinicornis (Newman, 1842) | d9bbe53c-29c1-102b-9a4a-00304854f820 | 168868 | ac2008 |
| 9826866 | Rhytidodera bowringi White, 1853 | d9bbe668-29c1-102b-9a4a-00304854f820 | 171174 | ac2009 | overlaps | Rhytidodera bowringi White, 1853 | d9bbe668-29c1-102b-9a4a-00304854f820 | 168869 | ac2008 |
| 9826868 | Prionomma attratum (Gmelin, 1790) | d9bbe78a-29c1-102b-9a4a-00304854f820 | 171175 | ac2009 | overlaps | Prionomma attratum (Gmelin, 1790) | d9bbe78a-29c1-102b-9a4a-00304854f820 | 168870 | ac2008 |
| 9826878 | Prionomma attratum (Gmelin, 1790) | d9bbe78a-29c1-102b-9a4a-00304854f820 | 171175 | ac2009 | overlaps | Ceratocentrus spinicornis (Fabricius, 1792) | ee1f1012-29c1-102b-9a4a-00304854f820 | 1498926 | ac2008 |
| 9826882 | Mimancylistes malaisei Breuning, 1955 | d9bbe8b6-29c1-102b-9a4a-00304854f820 | 171176 | ac2009 | overlaps | Mimancylistes malaisei Breuning, 1955 | d9bbe8b6-29c1-102b-9a4a-00304854f820 | 168871 | ac2008 |
| 9826884 | Falsovelleda congolensis (Hintz, 1911) | d9bbe9e2-29c1-102b-9a4a-00304854f820 | 171177 | ac2009 | overlaps | Falsovelleda congolensis (Hintz, 1911) | d9bbe9e2-29c1-102b-9a4a-00304854f820 | 168872 | ac2008 |
| 9826890 | Falsovelleda congolensis (Hintz, 1911) | d9bbe9e2-29c1-102b-9a4a-00304854f820 | 171177 | ac2009 | overlaps | Falsovelleda similis Breuning, 1970 | ee197ad0-29c1-102b-9a4a-00304854f820 | 1494954 | ac2008 |

Figure 6: Cross-mapping prototype - inspection of "overlaps" relationships

there is some other taxon in checklist B which contains some of the names in the checklist A taxon.

4. *overlaps* – a pair of taxa which participate in a many-to-many relationship between checklist A and checklist B. The MySQL query identifies a taxon in checklist A as overlapping a taxon in checklist B if the taxa have some names in common, but there is some other taxon in checklist B containing some of the names in the checklist A taxon, and there is some other taxon in checklist A containing some of the names in the checklist B taxon.

Figure 6 shows a cross-map created by this prototype being browsed. When browsing the cross-map, the accepted names are shown. Clearly in the download format it is necessary to be able to recover all names associated with a taxon, not just its accepted name. A download facility is under construction which splits the download into two distinct parts:

- A cross-map between taxa, each row being of the form taxonId1-checklistId1-relationship-taxonId2-checklistId2
- A set of all names for each of the taxa involved in the checklist, each row being of the form taxonId-checklistId-name-genus-epithet-infraspecificEpithet-authority-rank-status. (infraspecificEpithet is defined to be empty where not applicable; similarly, for higher taxa, epithet is empty.)

The “additional taxa” download format will continue to be supported, due to its key role in the i4Life “pipelines”, but it will be extended to include information about the higher taxa above each species listed.

## *Further development of PHP/MySQL/PERL cross-mapping tool*

Further developments planned for the PHP/MySQL/PERL cross-mapping tool include:

- "Fuzzy" name matching (the current software regards names as being the same only if they are identical, and so does not identify corresponding names in a pair of checklists in cases where there are minor variations in the authority string, etc.).
- Greater decoupling between the PHP software which controls the user's interaction with the system and the cross-mapping process.
- SOAP services to control the upload, cross-mapping and download processes programmatically.

### *Further development of "Semantic Web" cross-mapping tool*

As mentioned above, another cross-mapping tool has been developed which uses Semantic Web technologies, including SPARQL and SWRL, in order to represent and reason with taxonomic checklists. This version currently supports two taxonomic relationships – the *includes* and *corresponds* relationships – and provides a Web front end which enables users to create and browse cross-maps. SOAP services have also been implemented to allow users to download cross-maps. At present there are scalability problems with this system, and so it would not be possible to compare two very large checklists against each other in a reasonable period of time. We are currently investigating the use of a triple store in order to provide us with the scalability we require. With appropriate performance, the ability to represent knowledge declaratively regarding detection of relationships between taxa (for example, using SWRL) means that such knowledge can be encoded in a way that corresponds more nearly to the ways in which it has been formulated by human experts.

Some significant elements of the PHP/MySQL/PERL software can be re-used to enhance the Semantic Web tool, particularly the import routines and the routines which prepare intermediate tables – for example, a table containing all pairs of taxa from the two checklists being compared which have at least one name in common. The intention is to continue to use a relational database to create these tables, and then to import the resultant data into a triple store for the extended reasoning that is subsequently to be performed.

A final consideration is that it is intended to enhance the cross-mapping tools so that they are compatible with the new e-2 architecture developed in the related 4D4Life project.

## Conclusions

In this document we have described the cross-mapping software delivered as i4Life D11.1, using which checklists provided in Darwin Core Archive form can be uploaded into a checklist repository and compared, detecting additional taxa, which can then be browsed or downloaded for subsequent use. We have also described an enhanced version currently in prototype form, using which a wider range of checklist relationships can be detected, and another prototype, using Semantic Web technology, which is intended to provide a more declarative, flexible way of describing rules for detecting relationships between taxa in checklists.

The delivered software can be accessed via the following URL:

http://i4life.eu/crossmapping.php

Further information about cross-mapping in i4Life, including links to the other prototypes mentioned in the present document, can be accessed via the following URL:

http://litchi.cs.cf.ac.uk/