Indexing for Life

Enhanced Cross-Map Service Specification

Workpackage 2

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

27 June 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

This is a revised, extended version of an earlier i4Life internal document, "Preliminary Cross-map Service Specification: Discussion Document and Service Proposal". In the current document we specify the operations to be supported by i4Life cross-map services, and also specify what cross-maps are to contain. This information is supported by a number of sections in which we provide background and context about cross-mapping, and also by a description of preliminary cross-mapping tools, currently under development and test, which include a Web interface to control the generation and querying of cross-maps.

In more detail:

- The *Background* section explains the concept of a "cross-map" and makes reference to some past cross-mapping work that we have done; it explains why we are pursuing a rule-based approach to cross-map generation.
- The *Scenarios of Use ("User Stories")* section describes various ways in which a cross-mapping facility – and the resulting cross-maps – can be used.
- The *Architecture of a Cross-Mapping Facility* section explains that, in fact, there is not to be a single cross-map service, as the cross-map facilities can be interacted with in more than one way. We are starting with a Web interface, as explained below.
- *What to Represent in a Cross-Map* discusses the key information to be held in a cross-map in order to provide effective mapping between taxonomies.
- In the *Reasoning* section we discuss in more detail how automated processes can infer relationships, but in the subsequent *Role of the Human Cross-Mapper* section we explain the complementary role that humans must have in intervening in the cross-mapping process in order to refine the cross-maps generated.
- In the *Data formats and Preparation* section we consider how checklist data is to be provided to the cross-map facility, and prepared for cross-mapping.
- The *Cross-Map Service Operations* section details the operations which a cross-mapping facility (whether used via Web pages, SOAP, REST or other methods) is to provide.
- The *Cross-Map Relationships* section describes how individual taxa from different checklists are to be associated together, and in particular it provides a detailed discussion of the taxon-taxon relationships embodied in a cross-map.
- In the *Web-Based Cross-Mapping Facility* section we describe the current state of software we are developing which provides a Web interface for users to exploit a number of cross-mapping services, such as the generation of cross-maps, and querying and retrieving the resulting cross-maps.
- In the *Conclusions* section we summarise how the cross-mapping services will be enhanced and extended as the i4Life project proceeds.

Background

The purpose of a "cross-map" is to enable taxa (including species) and lists of taxa in one species information system to be related to those in another. 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.

In the LITCHI 1 and LITCHI 2 projects (http://biodiversity.cs.cardiff.ac.uk/litchi/), alternative approaches to reasoning with checklists have been explored. LITCHI 1 used *constraints*, expressed initially by taxonomists in natural language (specifically, English), specifying conditions that a consistent checklist must satisfy. These were then expressed in

the Prolog programming language, and used to identify conflicts in either an individual taxonomic checklist or a composite checklist assembled from more than one source. The goal of LITCHI 1 was to create consistent checklists, rather than to map between checklists, and so the relationships between checklists, although implicitly inferred during the "repair" process, are not explicitly represented. In LITCHI 2, a different approach was taken: a cross-map was created using a Java program in an incremental manner, by searching for relationships between the taxon source checklists and within individual checklists. This was implemented in a procedural manner, unlike in LITCHI 1, where a "declarative", inferential approach was taken.

The declarative approach offers significant advantages, especially in the much more direct mapping that it allows between experts' knowledge and the internal representation. Hence, in the development being undertaken in i4Life, it is not appropriate simply to re-use and enhance LITCHI 2. (It should be noted, however, that LITCHI 2 has a modular architecture which would allow it to be extended if additional work was done to map declarative rules onto the LITCHI 2 procedural approach.) Neither is it appropriate simply to re-use and enhance LITCHI 1, because it was developed for a slightly different (though related) purpose, and it is a stand-alone system with a number of software dependencies which one would not fulfil in the same way today. However, we are retaining the rule-based, declarative approach which was demonstrated in LITCHI 1 to be a promising way of capturing taxonomic knowledge, using this approach in the context of ontology and Semantic Web Tools (see later in this document).

Scenarios of Use ("User Stories")

In this section we describe a number of contexts in which a task is to be performed which a cross-map can support. In the following section we shall explain the various ways of interacting with a cross-map which these user stories imply are needed. Our suggested user stories are as follows:

Using a Cross-Map

- A user wishes to perform an analysis which involves using information from a number of species information systems. It is desired, starting from one or more scientific names, to retrieve the relevant information from these systems. The user's scientific name is associated with some specific taxonomy (either the Catalogue of Life or one of the global partners' – e.g. the NCBI taxonomy). He or she is using an analytic tool which is able to retrieve and process information from the various systems in a suitable manner, but which uses the cross-map in order to assist, seamlessly, in overcoming the semantic heterogeneity which the differing underlying taxonomies present.
- A user accustomed to working with a given taxonomy knows which scientific name is used for a species of current interest to him or her, and wishes to discover which names would be used in other systems. She or he uses a Web-based system which allows submission of such names, returning the corresponding name(s) in the target system.
- An extension of the above user stories is that, because the user wishes his or her work to be reproducible, it is necessary to make use of a specific cross-map edition (the cross-map being subject to change in later editions), which refers in turn to specific checklist editions.
- Similar stories might apply to common names.

Creating a Cross-Map

- The curator of a given species information system wishes to establish the relationships between the taxa (for example, the species) in his or her system and those in another system. He or she provides the taxonomies for these systems to a cross-mapping system, and automated inference occurs in order to determine possible relationships between the taxa. The automated inference will include:
  - Analysis of the names in the taxonomies being compared, including any synonymy information available in either of the taxonomies, looking for evidence either of exact match (perhaps with some variation of spelling, or of the authority string) or of possible relatedness (e.g. same specific epithet; higher taxa that share many of the same children; ...).
  - Inference of possible explanations for a large number of individual relationships identified, e.g. a higher taxon in one taxonomy might correspond to two or more in the other taxonomy, or there may just be an overlap.
  - Analysis of other related information, e.g. if two taxa are identified as possibly being the same, other information such as similarity of geographical and descriptive information, common names and molecular data could be used to assess the likelihood of this being a genuine match.
- The curator has received a draft cross-map following the above procedure. (S)he is presented with a cross-map which includes relationships which, according to the inference performed, appear to be definite, but also some where a possible relationship has been found, and others where it is clear that some relationship exists, but the precise relationship needs to be confirmed. (As an example of the latter, perhaps there is a clear relationship between a genus in one taxonomy and a different genus in another taxonomy. The expert needs to indicate whether the genera are, in fact, identical in concept, or whether perhaps one of the genera is a broader concept than the other, etc.

Maintaining a Cross-Map

- A classification has been updated. A user wishes to update a cross-map in order to accommodate these changes. A system compares the existing cross-map to the updated taxonomy in order to establish where possible what has changed and how. A human taxonomic editor performs further refinement of the cross-map.
- The Catalogue of Life classification, or other evidence that has been used in creating a cross-map, has been updated. A user wishes to update a cross-map between two taxonomies (neither of which is necessarily the Catalogue of Life) in order to make use of this revised evidence. A system compares the existing cross-map to the updated evidence in order to establish, where possible, what should be changed and how. A human taxonomic editor performs further refinement of the cross-map.

The above user stories come from our past experience and interactions with scientists – both in the context of i4Life and in other contexts – who have an interest in cross-mapping and the uses to which cross-maps can be put. Further user stories will emerge as the i4Life project proceeds, particularly from the intended creators and users of cross-maps, but we believe that the above examples are sufficiently representative to be taken as users' requirements for our design and for what is being implemented.

Architecture of a Cross-Mapping Facility

The "user stories" described above presuppose a number of different modes of interaction with a cross-map. In particular:

- For creation of a cross-map, the provision of a Web-based tool which allows users to upload complete or partial checklists and other information, use the cross-mapping tool to generate a preliminary cross-map and then refine it interactively, is attractive. It involves no installation of unusual software on the user's machine; it can be designed to provide a persistent environment in which the user can return to the cross-mapping task a number of times over an extended period until the task is "complete"; it allows others, perhaps experts in some of the taxa concerned, to provide feedback, refinement and annotation of the cross-map; and also offers the advantage that the relatively computationally intensive task of automated cross-map generation can be delegated to servers that may be substantially more powerful than the user's own machine.
- The user who wishes interactively to consult the cross-map may well find a Web interface to be suitable, for reasons not dissimilar to those given above.
- On the other hand, a "bespoke" software tool which needs to perform bulk processing of species-related information and uses a cross-map in order to assist with taxonomic interoperation can operate most efficiently if this cross-map is available locally to the tool. A database, structured to support the retrieval of related taxa, etc., will support this requirement. A download service, for retrieval of complete cross-maps, is therefore desirable.
- If currency is particularly important, then a remote cross-mapping service implementing operations which can be carried out on the latest applicable cross-map might be more desirable than very frequent download of the entire cross-map. Such a facility would be suitable for looking up a single taxon, or a set of taxa (submitted as a batch), responding with sets of taxa to which they map in the specified checklist. (The currency issue also relates to the Web-based interactive cross-map consultation facility mentioned above.)
- Embedability within relevant biodiversity informatics architectures, including the new 4D4Life e-2 architecture and the Global Names Architecture, is important, due to the relevance to these initiatives of the cross-mapping functions which we are developing.

Ideally, then, all these different modes of interaction with a cross-map need to be supported. On the other hand, they also need to be prioritised in order that useful cross-map related facilities start becoming available as soon as possible. Due to the potential scope for extending the cross-mapping software as the project proceeds, and the need to make software available for testing as soon as possible, our first priority is the production of an adequate user interface for testing the framework.

What to Represent in a Cross-Map

Clearly 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-map is therefore fundamentally a mapping between suitable immutable taxon identifiers in the various taxonomies.

However, the relationships between names in a cross-map may also be of importance, especially names which have been determined to be identical, except (for example) for minor

changes in author string representation, etc. If anything can be deduced about the relationships of the names in an individual taxon (do they, in fact refer to the same concept or not?) then this may also be useful to store. In some cases, further useful information can come from nomenclators (such as whether they are orthographic variants, or refer to the same basionym). We also envisage that situations may arise where the scientific name, or a corresponding identifier (such as ITIS TSNs), is the closest that we will be able to get to a stable taxon identifier.

A key issue is whether the cross-maps are pair-wise mappings between individual global partners' taxonomies, or whether they are mappings to and from the Catalogue of Life. The latter kind of mapping can be used in some cases, via what might simplistically be thought of as a transitivity relationship, in order to derive the former kinds of mapping. However, this is a simplistic assumption, because it may be that there is not a precise mapping to and from the Catalogue of Life for some taxa, but there is a precise mapping between the global partners' taxonomies for these same taxa. A rather extreme case would be where exactly the same taxon existed, with the same accepted name, etc., in two global partners' taxonomies but it did *not* exist in the Catalogue of Life - perhaps some related taxa were present, but not the same one. A loss of precision would then occur if one followed a cross-map relationship into the Catalogue of Life and then back out again. Both kinds of cross-maps are needed; global partner-to-global partner "pairwise" cross-maps are essential, and not an optional extra. As mentioned later, the Catalogue of Life has an important role to play in establishing the pairwise relationships as well as relationships with itself.

The above discussion provides the context for the creation and use of a cross-map; in a later section (*Cross-Map Relationships*) we specify the relationships that are in fact to be incorporated in these cross-maps.

Reasoning

A very simple, but nevertheless useful, achievement is to identify the location(s) of each individual name in each checklist, and cross-map directly between the taxa that share these names. (This is already done in the present prototype – see later.) This cross-mapping can be further informed by using the Catalogue of Life as a "synonymic index", which might provide the association between names in other taxonomies which would not otherwise be identifiable. A further, relatively straightforward development is to identify taxa that have *potential* relationships. This can facilitate manual cross-mapping by narrowing down the range of taxa that the human cross-map editor needs to consider, and is potentially also a way of reducing the scale of the problem to be dealt with in the full cross-mapping scenario described in the following paragraph.

Rules such as those used in LITCHI can be used to deduce other, more subtle relationships, e.g. that a species has been moved to another taxon. Analysis of the specific epithets and the authority strings may provide evidence for this: for example, "*Trifolium ambiguum* M. Bieb." and "*Amoria ambigua* (M. Bieb.) Sojak" might relate to the same taxon. But note that there is the added complication that the specific epithets are not *identical:* this is an example of where additional processing is needed in order to determine that the difference is (we presume) only one of gender. Such processing is best done as a pre-processing task, normalising or associating such matching names, in order to reduce computation at each stage of the reasoning process. Similarly, there may be variations in the authority string, such as the spacing variation in the example given earlier in this paragraph. As mentioned earlier, other, pre-existing tools can be used to help with this particular task.

More “intelligent” rules can be introduced. The above kinds of reasoning are very “atomised”, detecting relationships between individual taxa. Rules which hypothesise an “explanation” for a large set of differences between checklists are also desirable, especially since the more general the explanations, and (in consequence) the smaller the number of such relationships that need to be presented to an expert for review (and the more evidence is available in each case), the more likely (s)he is to be able to cope with them and assess them objectively. Implementing such inductive rules in an effective way is a challenge which we wish to tackle, but to defer until later in the project.

The above discussion has assumed that the reasoning is primarily about the relationships between scientific names in the checklists being compared. Other forms of reasoning which are desirable include the consideration of associated common names; usage or occurrence data; higher taxa; taxonomic tree topologies, etc. There is the opportunity in the i4Life project for interaction between partners about the precise ways in which such information could be used; later in the project we are hoping to experiment with the enhancement of the rule base in order to accommodate information of this sort.

A further topic that has been raised in discussion with various colleagues is the possibility of introducing a “scoring” facility – reasoning with uncertainty. This would make it possible to set a threshold score above which mapping is performed automatically, without user intervention. This is an attractive possibility, but it presents issues of knowledge representation and reasoning which are likely to have to remain outside the scope of the present project, in which the priority must be to provide a robust, scalable cross-mapping method.

Role of the Human Cross-Mapper

Clearly the cross-maps generated by our system will always have ambiguities. Some ambiguities will be merely the result of limitations in the rules stored by the cross-mapper. More fundamentally, however, many of the ambiguities will result from the fact that there is theoretically more than one possible explanation of the relationship between a set of taxa in one taxonomy and a related set in another taxonomy. These ambiguities can only be resolved by expert intervention, making use of taxon-specific knowledge. Acquiring this information may require consultation of the literature, talking with a taxonomist who has specialised in the groups in question, etc. Hence it is necessary for the human expert to be able to express his or her decisions via the (Web) user interface provided.

The expert is also a source of information about the knowledge to be employed by the cross-mapping tool. Some rules might be applicable to one kingdom but not another, for example. It is desirable, at a minimum, for the expert to be able to select which rules are employed in which circumstances. Moreover, it reduces the opportunity for error if rules are expressed in a declarative form which can automatically be translated into natural language intelligible to the expert. A more long-term goal, towards and beyond the end of the i4Life project, would be to provide an environment in which an expert could express new rules in a form that could then be translated into another form, directly usable by the cross-mapping tool.

Data Formats and Preparation

For input to the cross-mapping tools, a canonical checklist representation is desirable, in order to minimise the variations that need to be accommodated during cross-map construction. We have developed a database schema (described below) that contains the information needed, from which the cross-mapping tools can take input and perform their reasoning. This schema has been developed in order to present the checklists in a suitably

pre-processed form, and is intended for internal use within the cross-mapping software; we do not anticipate that users would wish to submit checklists in this form. Instead, we note that a number of i4Life partners are preparing to use Darwin Core Archive to transmit and receive checklists, so this is one important type of input that will need to be supported, and converted into our internal format. There is flexibility to accommodate other input formats too, if necessary; the appropriate software to perform each conversion will need to be written. At present, manual conversion routines have been implemented for some checklist formats that are being used for testing purposes.

The process of conversion also requires normalisation (not least to remove deal with mundane phenomena such as extraneous spaces), which is currently done as part of the manual conversion process, but we intend to exploit complementary developments available from other researchers here. For example, Tony Rees' TaxaMatch is a much more mature approach to dealing with orthographical variation, etc., than anything previously developed in the LITCHI projects, and has been developed further by Jerry Cooper (LandCare Research, NZ) into a tool for aggregating taxon concepts from multiple checklists. Greg Whitbread (Australian National Herbarium, Centre for Plant Biodiversity Research) has a similar Taxon Match tool.

Schema for Pre-Processed Checklists

The following structures are used to represent taxa and their names in a pre-processed form, for input to the cross-mapping software:

Table	Field	Description
Name	id	Internal identifier for a name
	name	A "tidied" name
NameUse	id	Internal identifier
	edition	Name of the source (e.g. "ac2008")
	origName	Original name, in unprocessed form
	tidyName	Name, possibly amended due to "tidying" operations
	nameId	Reference to occurrence in *Name* table
	uuid	Globally unique identifier (if available)
	rank	Taxonomic rank
	as	Status – accepted, synonym, etc.
	match	Whether original and "tidied" names match
Taxon	edition	Name of the source (e.g. "ac2008")
	id	Identifier of this taxon
	rank	Taxonomic rank of this taxon
	parentId	Identifier of this taxon's parent taxon
	left	Identifiers included for efficiency purposes, to support a "nested set" representation of the tree
	right	

The Cross-map Service Operations

As we have observed earlier in this document, there is more than one way in which clients will want to interact with a cross-map service. The main distinction is between Web interfaces for human users and interfaces for direct, programmatic access (such as REST or SOAP Web Services; use within the 4D4Life e-2 architecture, and within the GNA) that can be used by other software that has embedded support for the cross-map tools to be provided. We envisage the Web interface will be more useful for creators and maintainers of cross-maps, and the direct programmatic access will be more useful for creators of software that

makes use of cross-maps for query enrichment and other such purposes. It should be borne in mind, however, that for some purposes a full cross-map available locally will be more efficient than a long string of cross-map query operations, and this cross-map could be retrieved by either approach.

Many of the operations specified in the following table therefore have more than one realisation: for example, in a Web Interface, an important further step is to define precisely how the interface will be presented to the user; in a SOAP service, it is necessary (among other things) to construct a WSDL file, providing details needed by software clients for invocation of these operations. We have started with the Web Interface, as described later, but in the following table we specify the operations that are to be provided and, for each one, whether it is primarily for use in a Web interface or via an interface designed for programmatic access.

Behind these various forms of external interface to our software, it is desirable to have a common "engine" which remains unchanged. The Service Component Architecture[1] makes it possible to do this in a disciplined manner, and it is the approach being taken in the related 4D4Life project for its new e-2 architecture.

[1] http://www.osoa.org/display/Main/Service+Component+Architecture+Home

Operation	Inputs	Outputs	Comments	Web interface?	Programmatic access?
Create project	Project name, project owner (+ access rights)	Success/failure	Each cross-mapping task has its own associated project	✓	(✓)
Upload checklist	Checklist in appropriate form (support for a Darwin Core Archive format being mandatory), or a URL referring to the location of the checklist; project name	Success/failure, internal checklist label	This would be used for CoL and for other checklists	✓	(✓)
Create cross-map	Checklists to cross-map, additional evidence (such as CoL) to be used in cross-mapping; project name	Success/failure, internal reference for current cross-map + version	Cross-map will include "unresolved issues" at this stage	✓	(✓)
Update cross-map	Revised checklists to cross-map	Success/failure, internal reference for current cross-map + version	This is for the situation where revised versions of checklists have been uploaded, and the cross-map is to be updated accordingly, as far as it is possible for this to be done automatically	✓	(✓)
Retrieve cross-map	Project name, Cross-map reference	Success/failure, Cross map	Retrieve cross-map (orcross-map subset) for local use	✓	✓
Refine cross-map	Project name, Cross-map reference, updates	Success/failure, internal reference for current cross-map + (new) version	Updates are things like "connect taxon X in checklist A to taxa Y and Z in checklist B with relationship R"	✓	(✓)

Operation	Inputs	Outputs	Comments	Web interface?	Programmatic access?
Retrieve cross-mapped taxon	Globally Unique Identifier for source taxon, or scientific name + source taxon checklist name; destination checklist	Success/failure, List of globally unique identifiers for corresponding taxa in destination checklist + metadata regarding their relationships to the source taxon, OR list of taxa (scientific names)	This operation may later be supplemented by several more specific operations	(✓)	✓
Retrieve attached data from external service X	Source taxon, service specification for service X, SOAP (or other) specification of message to send to service X information about taxonomy supported by service X	Success/failure, result received from service X, taxonomic metadata qualifying the result	The idea is of an "umbrella service" that can act as a client for species information services, and supplement the data retrieved with taxonomic metadata. The taxonomic metadata is (for example) whether the result set relates to a broader or narrower concept than the concept associated with the source taxon	(✓)	✓

Cross-Map Relationships

The i4Life cross-map makes some basic assumptions:

1. It is possible to treat all relationships of any significance as being one-to-one or one-to-many. Although a taxonomic revision may involve significant rearrangement of the concepts involved in a particular treatment, this will mean that all one might be able to be certain of is that there are some relationships between higher taxa in the original and revised treatments (a *subtaxa-regrouped* relationship – see the table of relationships below).
2. Therefore all relationships of any significance can be expressed as triples of the form:

   <taxon1> <relationship> <taxon2>

   where <taxon1> and <taxon2> are identifiers of two taxa involved in the relationship. For one-to-many relationships there will be more than one triple. For example, if a source taxon t_s has relationship r with destination taxa t_{d1} , t_{d2} , etc., then the triples would be:

   t_s r t_{d1}
   t_s r t_{d2}
   etc.
3. A *closed-world assumption* is needed – in this context, the assumption that if no relationship is given in a cross-map between two taxa, they are unrelated. Otherwise, one has to indicate explicitly that almost all taxa are unrelated to almost all taxa, and the scale of the reasoning needed increases substantially.
4. A hierarchy of relationships is needed, for at least two reasons:
   a. An automated process may discover that there is *some* relationship between a given set of taxa; a human refinement process may lead to assertion of more precisely what this relationship is.
   b. This gives us a way of extending our relationships as new categories of relationship become apparent without breaking software that uses an earlier set of relationships: the software can discover that a newly-introduced relationship type is a specialisation of some relationship type that it is able to accommodate.
5. An important category of relationship is relationships between circumscriptions, which can be expressed as set relationships (includes, overlaps, etc.). However, another important category of relationship is one which denotes that there is some causal explanation (split, merged, etc.). We therefore include these in the set of relationships that can be modelled.

The following table specifies the relationships that a cross-map can define between taxa. As mentioned above, it is designed so that it can be extended if necessary in due course. We define four levels in the hierarchy of relationship types, but there is no reason why specialisations that introduce one or more further levels cannot be introduced when extending the set of relationship types. Note that it will sometimes be necessary for more than one relationship to be defined between taxa – especially in the case of attached-data-changed because this might imply some change of concept, but does not limit us to one particular kind of change such as the circumscription having been made narrower.

Level 1	Level 2	Level 3	Level 4	Relationship type	Description
unrelated				General	We do not envisage explicitly including this relationship in cross-maps, because of the closed-world assumption described above
possibly-related				General	This allows us to assert that there may be a relationship between taxa, so that the closed-world assumption is not applied, deducing that they are definitely unrelated. Typically such a relationship would be replaced by some other relationship during refinement of a cross-map (or removed altogether, indicating that the taxa are unrelated after all)
related				General	Some relationship exists between taxa, but nothing is (yet) known about its nature. Typically it would be replaced by a more specialised relationship during refinement of a cross-map
	congruent			Set	The two taxa have the same circumscription
		transfer		Causal	The taxon has been moved to a different higher taxon, without change of circumscription
			generic-transfer	Causal	Taxon transferred to different genus, and so generic name and possibly specific epithet will have changed
			rank-change	Causal	A taxon is unchanged in circumscription but has moved to a new higher or lower rank (and its name – and even the form (binomial, trinomial, etc.) – may have changed in consequence)
		regrouped			The two taxa are the same concept, but their subtaxa have been regrouped
		nomenclature			A taxon is unchanged in circumscription but has had its name changed by a nomenclatural act
	includes			Set	The second taxon is considered part of the first
		merged-from		Causal	The second taxon is one of a set of taxa that have been merged into the first one
		split-into		Causal	The first taxon has been split into two or more taxa, of which the second taxon is one
	overlaps			Set	Part of the first taxon is the same as part of the second, but each taxon has parts which are not in the other one

Level 1	Level 2	Level 3	Level 4	Relationship type	Description
	disjoint			Set	Taxa do not overlap (but have some apparent commonality, e.g. related by homonym or misapplied name, so the fact that they do not overlap may need explicitly to be asserted)
	attached-data-changed			Causal	Some attached data has been changed. This may be associated with some change of circumscription, but the change could be of more than one kind (see discussion before this table). The sub-types of this particular category are likely to be subject to being added to in due course
		distribution-changed		Causal	Change in distribution data
		reference-changed		Causal	Change in bibliography

Web-based Cross-Mapping Facility

A cross-mapping tool is under development at Cardiff; the current version is available for test and use at:

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

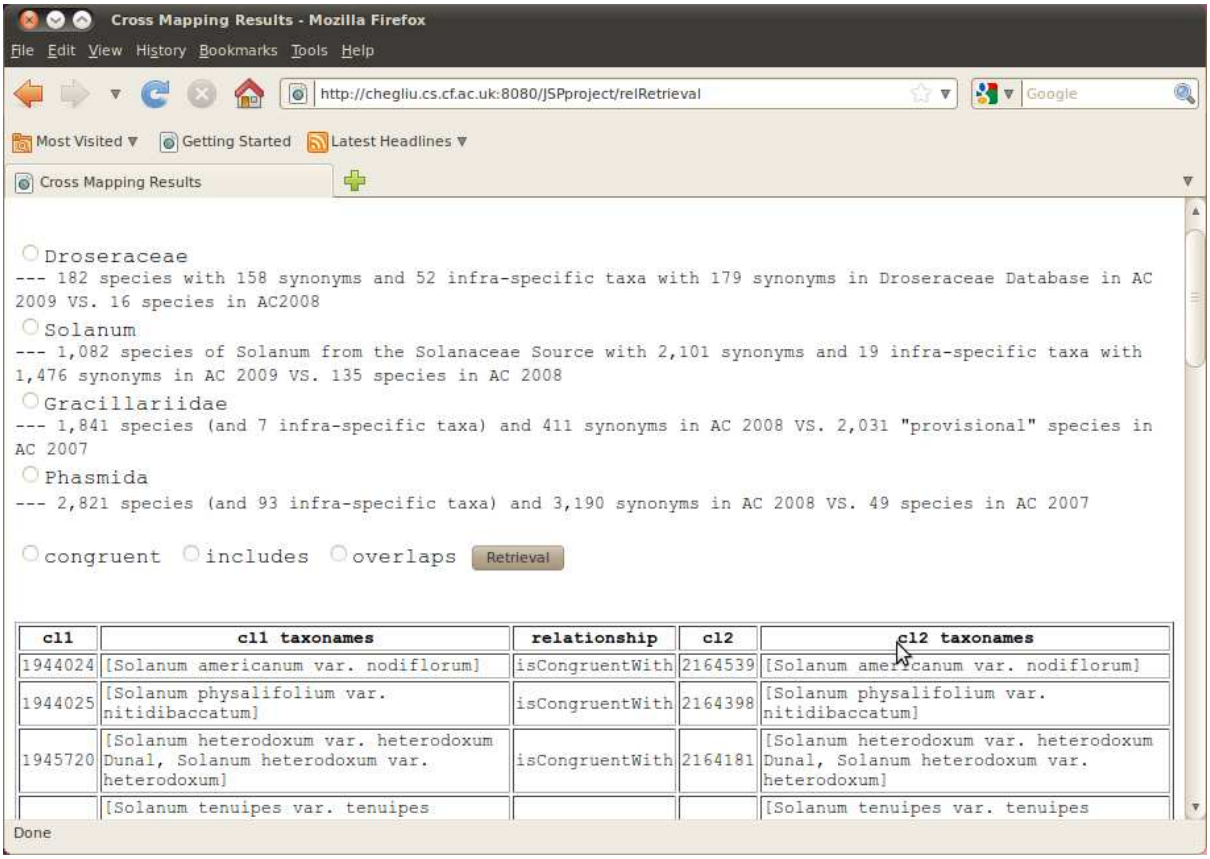
In addition to the latest version being available at this site throughout the i4Life project, access to previous versions of importance will also be maintained, as necessary. Information such as the underlying ontology used, etc., will also be maintained at that site.

This tool makes use of Semantic Web technologies, although it does not currently publish its data as Linked Data. Data published in Resource Description Framework (RDF) and Web Ontology Language (OWL) has the ability to be readily interlinked, which is a fundamental feature of the Semantic Web, and this is attractive in the context of inter-linking species-related data. In our application of the Semantic Web approach, an ontology written in OWL is used to model and represent the individual taxa and their relationships. Initial interlinking is discovered by running SPARQL queries against the taxa imported into our ontology as instances. Based on the interlinks generated, new information or knowledge such as the relationships among taxa is inferred using rules expressed in SWRL (Semantic Web Rule Language).

At the time of writing, import of checklists into a form that can be used by our cross-mapping tools is a manual process, as described earlier (*Data Formats and Preparation*); the cross-mapping tool itself provides the following features:

1. Select one of the pre-imported checklist sets (at the time of writing there is a choice of four) and invoke the SPARQL and SWRL reasoning phases, generating a cross-map.
2. Select a cross-map and a relationship type, and display all relationships of that type that have been inferred. (At the time of writing, three relationships can be inferred: *includes*, *overlaps* and *congruent*.)
3. Type in a scientific name and retrieve the taxa in a selected cross-map that pertain to that name.

The figure on the following page illustrates a typical example of a cross-map, as displayed by our prototype. It should be noted that although we are adopting a particular internal representation, this representation does not dictate the nature of the input formats, the cross-maps generated by the system, or the services available (although some are more effectively provided by some technologies than by others).



Conclusions

This document has specified the operations and facilities that i4Life cross-mapping services are to support. A fundamental issue is the *quality* of the automated cross-mapping that these services perform. This is dependent on the rules implemented, which at the time of writing are fairly simplistic. As explained in this document, we have experience of implementing more sophisticated rules in past projects. Nevertheless, as the cross-mapping prototypes evolve, this is a key area where experimentation, consultation and further elicitation of partners' understanding of the taxonomic process will be needed.