Indexing for Life

Cross-mapping tools: Rules used

Workpackage: 2

Authors:

Andrew C Jones,
Francisco Quevedo, Richard J White, Alex Hardisty

Introduction

In this document we provide a brief non-technical overview of the i4Life cross-mapping software, and describe the rules which the software uses in order to build cross-maps between taxonomic checklists.

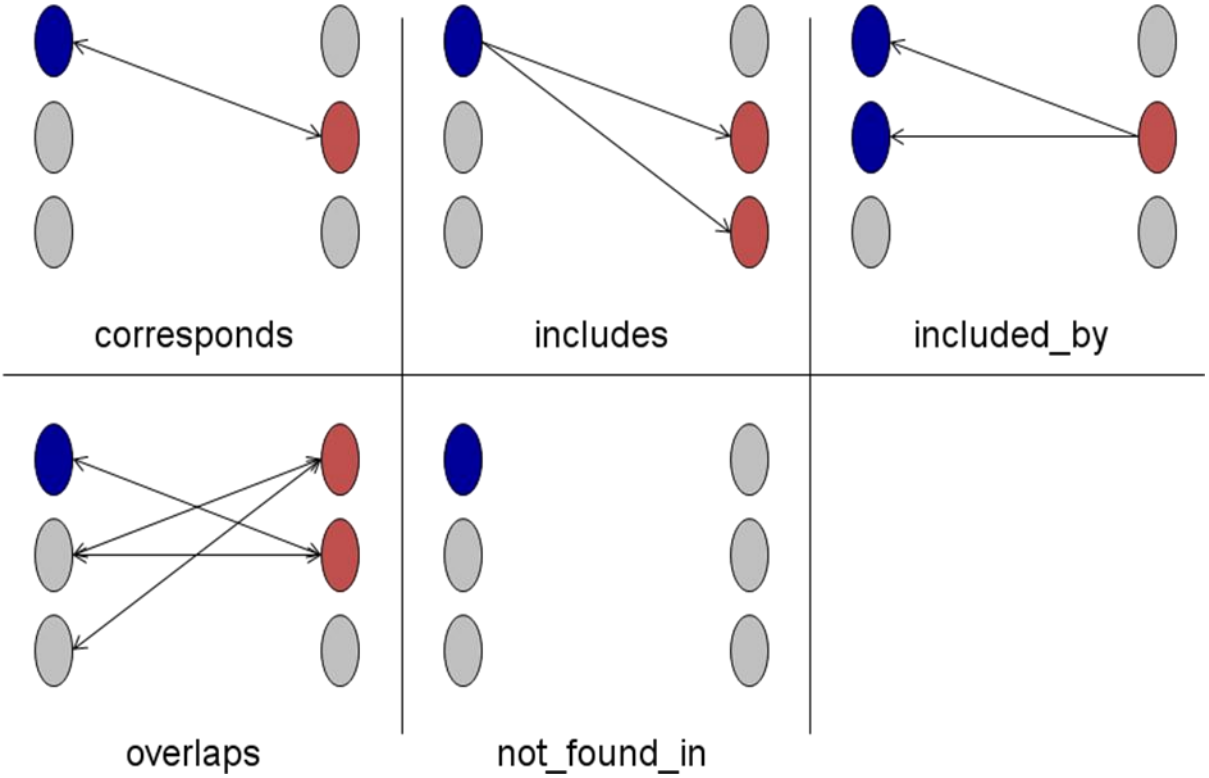
Overview

The i4Life cross-mapping tools are designed to cross-map a pair of taxonomic checklists provided in i4Life Darwin Core Archive-compliant format, identifying relationships between the taxa represented in the checklists by applying a set of rules. The output of the process is a set of three tab-separated-value text files:

- a cross-map, which includes identifiers of the taxa cross-mapped, and the relationships between these taxa;
- a file associating each identifier which occurs in the cross-map with all the scientific names which occur in the corresponding source checklist for that identifier, and
- a file which contains "additional taxa" – all the taxa in the first checklist which do not occur in the second. The additional taxa are output in i4Life Darwin Core Archive-compliant format, and can be used as input to the i4Life Piping Tool.

Cross-map relationships

Five relationships are detected:



- **corresponds** – taxa in the two checklists being cross-mapped correspond (have the same circumscription)
- **includes** – a taxon in the first checklist (checklist on the left-hand side) contains two or more taxa in the second checklist
- **included_by** – a taxon in the first checklist is one of *two or more* taxa included by a taxon in the second checklist

- **overlaps** – a taxon in the first checklist contains parts of the circumscription of at least two taxa in the second checklist *and* at least one of these latter taxa contains part of at least one other taxon in the first checklist.
- **not_found_in** – a taxon in the first checklist does not have any detectable relationship with *any* taxa in the second checklist.

Note that any taxon participating in a "not_found_in" relationship occurs in the cross-map, in the usual way, but also in the "additional taxa" file.

Detection of cross-map relationships

Hierarchical and species-level cross-mapping

The cross-mapping software proceeds by inferring a cross-map between the *species* in the checklists being compared initially. It does this by comparing the names associated with each taxon. Having derived a cross-map between the checklist species, the cross-mapping software proceeds to generate a cross-map between the higher taxa. It does this by traversing step by step up the hierarchies of the two checklists in parallel, stopping at each point where a common taxonomic rank is found, and cross-mapping between the taxa at this rank by considering the cross-mapped taxa at the next lower rank, which are contained within the taxa currently being cross-mapped. For example, having completed a cross-map between species, a cross-map between genera might be created by considering the relationships between the species these genera contain; then a cross-map between families might be created using the relationships detected between the genera; etc. In other words, in hierarchical cross-mapping, the "contained" taxa at the next lower rank play a similar role to that played by the scientific names themselves when cross-mapping between species. Which ranks are cross-mapped will depend on which ranks the checklists have in common. For example, sub-genera will be cross-mapped if both checklists contain taxa of that particular rank.

In the descriptions which follow, we assume we are cross-mapping between two checklists "LEFT" and "RIGHT". Taxa from the left checklist are referred to as T_{LEFT} or (if there is more than one) $T_{LEFT,1}$, $T_{LEFT,2}$, etc. Where higher and lower taxa from the left checklist are being referred to, we will use "H" and "L", e.g. $H_{LEFT,1}$, etc. Similarly for taxa from the right checklist (T_{RIGHT}, etc.)

Note that it seems difficult to describe the cross-mapping rules without using some kind of "mathematical" notation. The reader is encouraged to consider the description of the relationships in the preceding section first of all, and then to consider the species-level rules. When seeking to understand the hierarchy rules subsequently, the reader is encouraged to bear in mind that they are essentially the same, conceptually, as the species-level rules, in the sense that the lower taxa already cross-mapped play the same role, during hierarchical cross-mapping, as the species' *names* play in species-level cross-mapping.

Species-level rules
The relationships are detected as follows.

corresponds
A taxon T_{LEFT} ***corresponds*** to a taxon T_{RIGHT} if:

- T_{LEFT} has a name in common with T_{RIGHT}, and
- T_{LEFT} does not have a name in common with any other taxon in the right-hand checklist, and
- T_{RIGHT} does not have a name in common with any other taxon in the left-hand checklist.

Note that T_{LEFT} and T_{RIGHT} might have more than one name in common. It is not required that all their names are common, because one checklist might have additional names not known to the other checklist compiler, and also some names which are essentially the same might fail to match. Similar considerations apply to the other relationships.

includes
A taxon T_{LEFT} ***includes*** a taxon $T_{RIGHT,1}$ if:

- T_{LEFT} has a name in common with $T_{RIGHT,1}$ and
- T_{LEFT} also has a name in common with some other taxon $T_{RIGHT,2}$ and
- $T_{RIGHT,1}$ does not have names in common with any taxon on the left-hand side other than T_{LEFT}

included_by
This is identical to the ***includes*** relationship, but in reverse. A taxon $T_{LEFT,1}$ is ***included_by*** a taxon T_{RIGHT} if:

- $T_{LEFT,1}$ and T_{RIGHT} have a name in common, and
- T_{RIGHT} has a name in common with some other taxon $T_{LEFT,2}$ and
- $T_{LEFT,1}$ does not have a name in common with any taxon on the right-hand side other than T_{RIGHT}

overlaps
A taxon $T_{LEFT,1}$ ***overlaps*** a taxon $T_{RIGHT,1}$ if:

- $T_{LEFT,1}$ and $T_{RIGHT,1}$ have a name in common, and
- $T_{LEFT,1}$ also has a name in common with some other taxon $T_{RIGHT,2}$ and
- $T_{RIGHT,1}$ also has a name in common with some other taxon $T_{LEFT,2}$

not_found_in
A taxon T_{LEFT} is ***not_found_in*** the right-hand list if:

- none of its names occur in any of the taxa in the right-hand list

Hierarchical rules

Having "bootstrapped" the cross-mapping process by creating a species-level cross-map, we then proceed, rank by rank, up the hierarchies of the two checklists, cross-mapping all the ranks which are in common between the checklists in turn by using the cross-map already built up so far.

corresponds

A higher taxon H_{LEFT} ***corresponds*** to a higher taxon H_{RIGHT} if:

- H_{LEFT} has a lower taxon L_{LEFT} which has a relationship with a lower taxon L_{RIGHT} contained within H_{RIGHT} and
- H_{LEFT} does not contain a lower taxon which has a relationship with any other lower taxon in the right-hand checklist, other than those contained within H_{RIGHT} and
- H_{RIGHT} does not contain any lower taxon having a relationship with any other taxon in the left-hand checklist other than taxa contained within H_{LEFT}.

includes

A higher taxon H_{LEFT} ***includes*** a higher taxon $H_{RIGHT,1}$ if:

- H_{LEFT} has a lower taxon L_{LEFT} which has a relationship with a lower taxon L_{RIGHT} contained within $H_{RIGHT,1}$ and
- H_{LEFT} also has a lower taxon which has a relationship with some other taxon $H_{RIGHT,2}$ and
- $H_{RIGHT,1}$ does not have lower taxa related to any lower taxa on the left-hand side other than those contained within H_{LEFT}

included_by

As in the case of species-level mapping, this relationship is identical to the ***includes*** relationship, but in reverse. A higher taxon $H_{LEFT,1}$ is ***included_by*** a higher taxon H_{RIGHT} if:

- $H_{LEFT,1}$ has a lower taxon $L_{LEFT,1}$ with some relationship to a lower taxon L_{RIGHT} contained within H_{RIGHT}, and
- H_{RIGHT} has a lower taxon which is related to some taxon contained within *another* higher taxon $H_{LEFT,2}$ and
- $H_{LEFT,1}$ does not contain a lower taxon with a relationship to any lower taxon on the right hand side, other than those contained within T_{RIGHT}

overlaps

A higher taxon $H_{LEFT,1}$ ***overlaps*** a higher taxon $H_{RIGHT,1}$ if:

- $H_{LEFT,1}$ contains a lower taxon $L_{LEFT,1}$ with some relationship to a lower taxon $L_{RIGHT,1}$ contained within $H_{RIGHT,1}$ and
- $H_{LEFT,1}$ also contains a lower taxon $L'_{LEFT,1}$ with some relationship to a lower taxon $L_{RIGHT,2}$ contained within some *other* higher taxon $H_{RIGHT,2}$ and
- $H_{RIGHT,1}$ also contains a lower taxon $L'_{RIGHT,1}$ with some relationship to a lower taxon $L_{LEFT,2}$ contained within some other higher taxon $H_{LEFT,2}$

not_found_in
A taxon H_{LEFT} is ***not_found_in*** the right-hand list if:

- none of its lower taxa are related to any of the lower taxa in the right-hand list

NOTES
1) We do not distinguish between different kinds of relationships between the lower taxa being compared in this process. Our hypothesis is that typically this does not matter. Comments on whether there are situations where it will affect the accuracy of cross-mapping would be welcome.
2) Three additional details, omitted in the above description, for simplicity, but included here, for completeness, are as follows:
   a) When initiating a cross-mapping task, the *highest rank to be compared* must be specified. All taxa of rank higher than this are ignored in the cross-mapping process. It is possible to specify "kingdom" as the highest rank to compare, thereby causing the entire hierarchy to be cross-mapped.
   b) An additional case where a taxon in the left-hand checklist is determined to be ***not_found_in*** the other checklist is where no taxon of the same rank exists in the other checklist. If, for this (or any other) reason, the current higher taxon cannot be cross-mapped, the taxa it contains are passed to the parent taxon for cross-mapping at the next higher rank.
   c) At present, infraspecific species are not cross-mapped. One development planned in the immediate future is to pass them through as ***not_found_in*** taxa. In the more distant future (possibly there will not be time to do it in i4Life), we intend to cross-map infraspecific taxa "properly", using rules similar (but not identical) to those used for species.

Matching modes and detection of additional potential relationships
In the above description, we have not been specific about how names are matched. We support various modes of operation, as follows:

1) Two modes of operation are supported:

   a) "Strict" match between names (where name strings, including authors, are matched for equality, as in previous versions)
   b) "Without author" match between names (where names are compared ignoring authors, e.g. "*Vicia faba* L." and "*Vicia faba* Linnaeus" would be considered to be the same)

2) The "Strict" match is extended to provide three options for dealing with names in taxa that the previous rules were unable to match:

   a) "None": behave as in previous versions, doing no further processing of these names.
   b) "Names only": for each name belonging to a taxon "not_found_in" the other list, search for a "without author" match with some name in that list. If such a match is found, replace the "not_found_in" relationship by a "poss_name_match" relationship.
   c) "Generic transfer": in addition to the comparison described in 2b, also search for evidence of generic transfer, where the specific epithet is the same, but the genus is different, and where the author string of one of the names is non-empty but contained in the other one. For example, "*Adenopa anglica* (Huds.) Raf." is a synonym of

"*Drosera anglica* Huds.", and it will be noted that the epithet *("anglica")* is the same, and the basionym author ("Huds") occurs within the author string for *Adenopa anglica.* If such a match is found, replace the "not_found_in" relationship by a "poss_gen_trnfr" relationship.

Taxa that are labelled with one of these two new relationships (poss_name_match; poss_gen_trnfr) are not included in the "additional taxa" list included in the download format.

The recently-introduced facility for refining a cross-map provides a means to restrict oneself to consideration of these "possible" relationships when seeking to identify equivalent names in the checklists being cross-mapped.