EDIT - feature request #10178

Implement fuzzy name matching

10/26/2022 07:42 PM - Andreas Müller

Status:	In Progress	**Start date:**	
Priority:	New	**Due date:**	
Assignee:	Belen Escobari	**% Done:**	50%
Category:	cdmlib	**Estimated time:**	0:00 hour
Target version:	Release 5.43		
Severity:	normal		

Description

Not only for single name requests but also for long lists of names.

Tony Rees: https://pubmed.ncbi.nlm.nih.gov/25247892/

mentioned at NFDI All hands: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13802

some further sources I found:

https://github.com/charvolant/ala-name-matching-2/blob/master/doc/index.md

https://github.com/charvolant/ala-name-matching-2

https://app.slack.com/client/T03ML9WUTU5/C0441A3RK17

GBIF/COL has an online name matching

Ask Gerhard Ludwig for his algorithm

Global Names: https://github.com/gnames/gnverifier or https://github.com/gnames/gndiff together with UI https://verifier.globalnames.org/ and WS-doc https://apidoc.globalnames.org/gnames

APCalign (Australian R-package): https://www.biorxiv.org/content/10.1101/2024.02.02.578715v1 (Code: https://github.com/traitecoevo/APCalign)

etc.

Open issues:

- infraspecific names are returned when searching for a species (and maybe similar issues)
- names with authors not yet implemented
- webservice for **list** matching not yet implemented
- compare with other existing matching services (maybe we are missing something that they have)
- ...

Associated revisions

Revision 80c3a867 - 04/20/2023 02:28 PM - Belen Escobari

ref #10178: Fuzzy names. comparison among two names

Revision 2948f981 - 04/20/2023 02:55 PM - Belen Escobari

ref #10178: method findingMatchingNames added

Revision abc6518f - 04/25/2023 11:36 AM - Belen Escobari

ref #10178: fuzzy name matching

Revision 523be807 - 05/08/2023 11:39 PM - Andreas Müller

ref #10178 implement distinctGenusOrUninomial

Revision ef509ae6 - 05/08/2023 11:42 PM - Andreas Müller

ref #10178 javadoc

Revision 8e17214d - 05/10/2023 12:22 PM - Belen Escobari

ref #10178: Fuzzy search implementation

Revision 38287f52 - 05/15/2023 03:39 PM - Belen Escobari

ref #10178: Fuzzy name matching

Revision ef37b9fd - 05/16/2023 04:12 PM - Andreas Müller

ref #10178 fix current name matching version and adapt tests

Revision 8ecd7215 - 05/30/2023 03:30 PM - Belen Escobari

ref #10178: new methods for normalization of names

Revision 31fc2ce7 - 06/19/2023 10:23 AM - Belen Escobari

ref #10178: fuzzy name searching without Authors

Revision fffbf62f - 06/19/2023 11:24 AM - Belen Escobari

ref #10178: Tests "not exact matches"

Revision f0b94864 - 06/19/2023 02:55 PM - Belen Escobari

ref #10178: query only genus, returns all species

Revision 5f2f7f86 - 07/04/2023 06:53 PM - Andreas Müller

ref #10178 some cleanup for name matching (not finished yet)

Revision 0be7d592 - 07/05/2023 01:14 PM - Andreas Müller

ref #10178 some cleanup for name matching (cont.)

Revision 8d8629ee - 07/13/2023 05:13 PM - Belen Escobari

ref #10178 Name Matching Service classes implemented

Revision 93398cdf - 08/21/2023 03:44 PM - Belen Escobari

ref #10377: matching names including infrageneric epith

Revision 35f3b533 - 08/23/2023 02:40 PM - Belen Escobari

ref #10377: matching names including infrageneric epith. Tests

Revision 439b8e84 - 09/06/2023 02:11 PM - Belen Escobari

ref #10178: matching names including infraspecific.

Revision 01536be1 - 09/22/2023 11:41 AM - Andreas Müller

ref #10178 some comments and code cleaning for name matching

Revision e6f24ce6 - 09/22/2023 12:06 PM - Andreas Müller

ref #10178 replace DoubleResult by SingleNameMatchingResult

Revision 79768b91 - 10/04/2023 12:37 PM - Andreas Müller

ref #10178 add configurator and matching result to name matching

Revision 8c5c1e90 - 11/23/2023 01:09 PM - Belen Escobari

ref #10178: Author Comparison

Revision 10d222da - 11/24/2023 10:36 AM - Belen Escobari

ref #10178: Author Comparison. Test excluded

Revision fc0a46b9 - 11/27/2023 02:24 PM - Belen Escobari

ref #10178: Ex Author Comparison and Authorship with "and" in between

Revision 06780b1a - 03/04/2024 04:05 PM - Belen Escobari

ref #10178 list of names as input

Revision 4716bb17 - 03/04/2024 04:29 PM - Belen Escobari

ref #10178 exact matches score corrected

Revision fa93f48e - 03/21/2024 10:25 AM - Belen Escobari

ref #10178 namematching Controller

Revision 17b82791 - 04/03/2024 12:35 PM - Andreas Müller

ref #10178 adapt nameMatching controller path and request param

Revision 06d235de - 04/03/2024 12:37 PM - Andreas Müller

ref #10178 remove inline call for taxonNamePartsFromDb

Revision 7415dbeb - 04/03/2024 01:02 PM - Andreas Müller

ref #10178 fix equals for genusOrUninomialWithDistance

Revision b7379459 - 04/03/2024 05:27 PM - Andreas Müller

ref #10178 fix IOOB in replaceGenderEnding

Revision d0525ea6 - 04/03/2024 07:47 PM - Andreas Müller

ref #10178 fix jsonConfigurations, introduce api data objects with adapter and cleanup

Revision 01582801 - 04/03/2024 08:33 PM - Andreas Müller

ref #10178 extract adapter methods into own class

History

#1 - 10/27/2022 10:11 AM - Belen Escobari

- The algorithm Taxamatch described by Rees (2014) employs a Modified Damerau-Levenshtein Distance in addition to two phonetic algorithm (Rees near match 2007 phonetic algorithm and Soundex) which seem to outperform all other tested algorithms. Taxamatch is written in Oracle PL/SQL programming language and can be downloaded from https://confluence.csiro.au/public/taxamatch/downloads.
- The paper by Grenie et al. (2022) mentioned at the NFDI conference uses a R package available from https://mgrenie.shinyapps.io/taxharmonizexplorer/
- The Damerau–Levenshtein distance is employed for Taxamatching in GBIF and in the REST-API as suggested by Gerhard Ludwig. It can be downloaded from https://github.com/crwohlfeil/damerau-levenshtein

#2 - 10/28/2022 05:54 PM - Andreas Müller

- Target version changed from Unassigned CDM tickets to Release 5.44

#3 - 04/20/2023 02:59 PM - Belen Escobari

- % Done changed from 0 to 20

new methods were included in the classes NameServiceImpl and INameService.

a method calculates the distance among two strings using the levenshtein distance. A second method retrieves the best matching name in database based on the distances LS in regard to the input name

#4 - 05/10/2023 12:47 PM - Belen Escobari

- File file.png added

The query is parsed and each part (genus, epithet) will be searched in a database one after each other.

A initial list is made with all the genera in the database starting with the first character of the input genus (alternative, all genera in the database could be listed independently of the first character match).

1. The input genus is compared agains each element of the initial list and all (near)coincidences are added to a map. The map includes all TaxonNameParts corresponding to the genus name and the distance scored. The default threshold value of similarity is 70%

2. The input epithet is compared agains epithets in the map built in the previous step and distances are added (distance Genus + distance epithet). Need to check how this is made in the TaxaMatch algorithm

3. Sort the map according distances and return the first x best matches (let the user decide)

Further documentation:

SQL code written by Rees: http://www.cmar.csiro.au/datacentre/downloads/taxamatch/taxamatch1.sql
Workflow: https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0107510.g002

#5 - 05/22/2023 04:41 PM - Belen Escobari

- File Rees-Taxamatching.docx added

- Status changed from New to In Progress

- Assignee changed from Belen Escobari to Andreas Müller

I attach a word file with the summary of the Taxamatching algorithm by Rees.

There are some points that need to be discussed from my point of view.

#6 - 05/23/2023 02:06 PM - Belen Escobari

- Assignee changed from Andreas Müller to Belen Escobari

#7 - 06/19/2023 03:43 PM - Belen Escobari

- % Done changed from 20 to 50

All points described by Rees for fuzzy name matching are included except Authors (to be discussed). The algorithm uses a max total distance of 4 changes between the names.

#8 - 08/28/2023 10:38 AM - Belen Escobari

The algorithm compares monomials and binomial names of the kind genus, genus + infrageneric epithet and genus + epithet. The default distance for monomial names is set to 2 and 4 for binomial names. Authors comparison needs discussion

#9 - 09/06/2023 02:35 PM - Belen Escobari

The algorithm compares monomial names (Genus), binomial names (Genus + infrageneric epithet / Genus + specific epithet), and trinomial names (Genus + epithet + infraspecific epithet).

#10 - 09/22/2023 12:08 PM - Andreas Müller

- Description updated

#11 - 10/18/2023 10:42 AM - Andreas Müller

- Description updated

#12 - 11/23/2023 01:25 PM - Belen Escobari

The algorithm includes comparison of authorities (cache), nevertheless the method that is used to parse the authority getAuthorshipCache() does not parse authorities as Teams but as single authors. The basionym should be excluded from the comparison.

It should be discussed if the authorities comparison distance should be added to the total distance (species names distances). In this publication https://www.mdpi.com/2223-7747/10/5/974, the total score is calculated like: ReturnedScore = (TaxonScore × 0.9 + AuthorScore × 0.1) × 100

#13 - 11/27/2023 10:31 AM - Belen Escobari

a new parser is needed for input names containing "and" instead of "&" in the authorship

#14 - 11/27/2023 02:48 PM - Belen Escobari

as temporary solution for names containing "and" in the authorship if the input name contains "and" it is replaced by "&" and only after the name will be parsed. This should work for all input names as the input names dont include publications.
Ex Authors are excluded from the authorship and only the last author is compared.

#15 - 02/12/2024 06:34 PM - Andreas Müller

- Description updated

#16 - 02/12/2024 06:38 PM - Andreas Müller

- Description updated

#17 - 02/14/2024 11:25 PM - Andreas Müller

- Target version changed from Release 5.44 to Release 5.43

#18 - 02/20/2024 11:15 AM - Belen Escobari

Find matching names using a list of names as input: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-16
UI: https://tnrs.biendata.org/

#19 - 03/04/2024 04:39 PM - Belen Escobari

new methods are implemented to use a list of names as input

#20 - 04/03/2024 08:46 PM - Andreas Müller

- Description updated

A first working webservice version for a single name matching exists now. Open issues are handled in the ticket description.

Files

file.png	2.6 MB	05/10/2023	Belen Escobari
Rees-Taxamatching.docx	15.5 KB	05/22/2023	Belen Escobari