

Paying it forward: Crowdsourcing of taxonomic harmonization and linking of biodiversity identifiers

Brandon K. B. Seah (brandon.seah@thuenen.de)

Thünen Institute for Biodiversity, 38116 Braunschweig, Germany

Abstract

Linking records for the same taxa between different databases is an essential step when working with biodiversity data. However, name-matching alone is error-prone, because of issues such as homonyms (unrelated taxa with the same name) and synonyms (same taxon under different names). Therefore, most projects will require some degree of curation to ensure that taxon identifiers are correctly linked. Unfortunately, formal guidance on such curation is uncommon, and these steps are often ad hoc and poorly documented, which hinders transparency and reproducibility, yet the task requires specialist knowledge and cannot be easily automated without careful validation. Here we present a case study on linking identifiers between the GBIF and NCBI taxonomies for a species checklist dataset. This represents a common usage scenario: finding publicly available sequencing data (available from NCBI) for species chosen by their occurrence or geographical distribution (from GBIF). Wikidata, a publicly editable knowledge base of structured data, can serve as an additional information source for identifier linking. We suggest a software toolkit for taxon name matching and data cleaning, describe common issues encountered during curation, and propose concrete steps to address them. For example, about 2.8% of the taxa in our dataset had wrong identifiers linked on Wikidata because of errors in name matching caused by homonyms. By correcting such errors during data cleaning, either directly (through editing Wikidata) or indirectly (by reporting errors in GBIF or NCBI), we crowdsource the curation and contribute to improvement of community resources, thereby improving the quality of downstream analyses.

25 Introduction

26 Biodiversity science has seen a proliferation of databases and checklists (Feng et al. 2022). While it is
27 clear that taxonomic experts are best-placed to curate data for their respective taxa of expertise, there
28 are drawbacks to group-specific specialized databases: they may not be maintained in the long term,
29 may not be interoperable with other databases, and may duplicate efforts when different projects have
30 overlapping coverage or aims (Schellenberger Costa et al. 2023). Similar observations have been made
31 about the software developed for working with them (Grenié et al. 2021). As a result, users face the
32 challenge of integrating different databases by linking or harmonizing taxon names and database-
33 specific identifiers, before they can take advantage of the domain-specific information contained in
34 them.

35 End-users can match taxa either by their names or taxon identifiers. This task is a subset of data
36 reconciliation or data matching (Christen 2012), a dynamic field with evolving standards (Delpuech et
37 al. 2023). Some databases, particularly those that themselves aggregate multiple sources (“data
38 aggregators”), may incorporate cross-references to other databases, but end-users are ultimately
39 responsible for curating the data they wish to use, and often have to rely on name matching. The
40 Linnaean system has been in use for almost three centuries, which attests to its utility and robustness,
41 but names are human artefacts and hence inherently prone to variants (e.g. in orthography) and errors
42 (Patterson et al. 2016). Additionally, a given name may also embody different taxon concepts (cf.
43 (International Commission on Zoological Nomenclature 1999) Article 61.3; (Turland et al. 2018)
44 Glossary).

45 How can we avoid duplicated effort in data curation? Ideally, users of taxonomic data would share in
46 building and improving community resources, as they are often also the subject-matter experts.
47 Building yet another database is clearly not the answer. Nonetheless, large aggregator projects such as
48 WoRMS and ITIS tend to be centrally organized by design, and may not have a formal avenue for
49 handling user contributions. Wikidata (<https://www.wikidata.org/>) (Vrandečić and Krötzsch 2014)
50 presents an alternative model for how data curation can be crowdsourced. Like Wikipedia, its better-
51 known cousin, Wikidata is freely accessible and editable by online users, and is actually the backend
52 for automatically generated information boxes displayed in Wikipedia articles, e.g. for biological taxa
53 (<https://en.wikipedia.org/wiki/Template:Taxonbar>). The Wikidata project aims to build a general
54 knowledge graph, comprising items (which may be entities or objects of any kind, including abstract
55 concepts) linked together by statements about how these items are related to each other. Each
56 statement comprises a subject and an object (items) linked by a predicate (a property). For example,
57 the item “*Coffea arabica*” ([Q47685](#)) is linked to the item “coffee bean” ([Q153697](#)) by the property
58 “this taxon is source of” ([P1672](#)). Biological taxa are modeled as instances of ([P31](#)) taxon ([Q16521](#)),
59 and typically have properties like taxon name ([P225](#)), authors ([P405](#)), and rank ([P105](#)). Taxon

60 identifiers in other databases can simply be represented through additional statements, e.g. “*Coffea*
61 *arabica*” ([Q47685](#)) has a property “GBIF taxon ID” ([P846](#)) with the value “2895345”.

62 Page (2022) has argued that it is ultimately more productive and sustainable to contribute to an
63 existing project already supported by an active community, such as Wikidata, than to start a new
64 domain-specific project, where such a user base would have to be built up from scratch. Wikidata is
65 already used in the life sciences for purposes such as crowdsourcing biological ontologies and data-
66 mining for drug discovery and disease diagnosis (Waagmeester et al. 2020). In biodiversity
67 informatics, it has been proposed as a platform for a “bibliography of life”—a comprehensive linked
68 database of the taxonomic literature (Page 2022), and to disambiguate personal names in collection
69 records (Groom et al. 2022).

70 Graphs of database identifiers have been used instead of name-matching to link over a hundred
71 thousand entries in Wikidata with the Global Biotic Interactions Database (GloBI) (Thessen et al.
72 2018). These large numbers are impressive, but rely on the identifiers being up to date and correctly
73 assigned. As a crowdsourced platform, the accuracy of Wikidata depends on smaller, individual
74 contributions. If one is not solely interested in global patterns but also specific cases, then careful
75 curation is necessary. This more modest but ultimately essential “bricklaying” by individual users is
76 the topic of this case study.

77 Here, we describe how we match taxon names and identifiers between the Global Biodiversity
78 Information Facility (GBIF) Backbone Taxonomy (GBIF Secretariat 2022) and the NCBI Taxonomy
79 (Schoch et al. 2020), integrating Wikidata into the workflow both as a source of linked identifiers to
80 speed up data matching, and as a community resource that we contribute to during data curation. GBIF
81 aggregates biodiversity distribution and occurrence data, whereas the main international repositories
82 for molecular sequence data, the International Nucleotide Sequence Database Collaboration (INSDC)
83 of which NCBI is a member, use the NCBI Taxonomy. This represents a common usage scenario of
84 finding biological sequences that belong to a set of taxa. The dataset used is a checklist of vascular
85 plants from Germany (Bundesamt für Naturschutz 2021). As this is a region well-studied by botanists,
86 we expect that virtually all species have been described and that most are well documented with
87 published occurrence and sequence data.

88 Our aims are to identify issues commonly encountered during data matching, in particular the actual
89 impact of homonymy and synonymy on name matching, and to make concrete suggestions for how to
90 troubleshoot and improve community resources as part of the data cleaning process, as a form of
91 crowdsourcing.

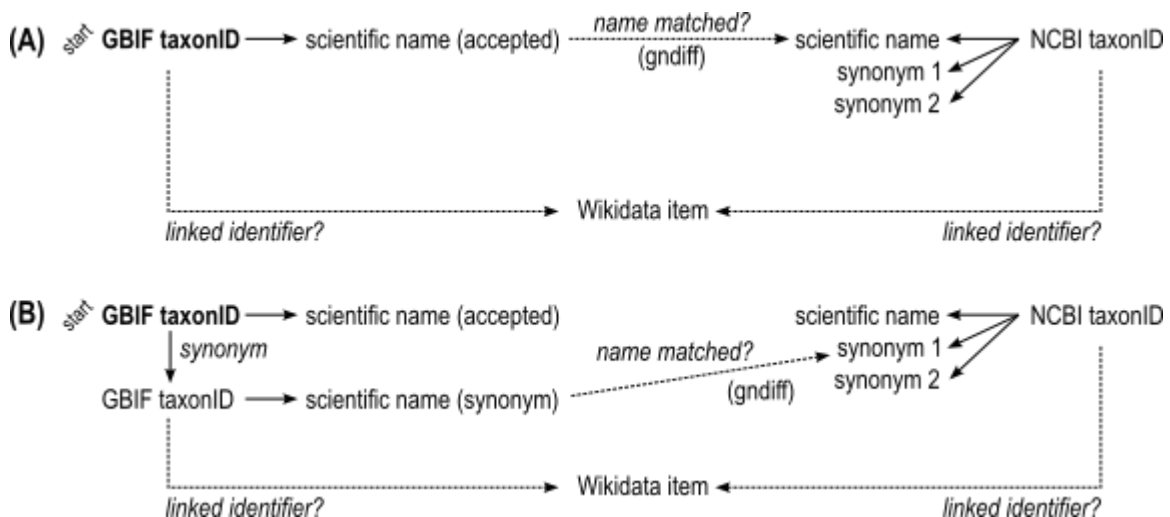
92

93 **Workflow to link identifiers and flag cases for curation**

94 The dataset (<https://doi.org/10.15468/Ofxsox>) comprises 7209 taxon names of vascular plants from
95 Germany (5876 at species rank), and their associated GBIF taxon identifiers (taxonIDs) which we
96 wished to link to equivalent NCBI Taxonomy taxonIDs. The file was downloaded from GBIF as a
97 “species list”, which lists taxa in a tab-separated text file, containing the taxon name as supplied by the
98 data provider, the taxonID for that name, the “accepted” taxon in the GBIF Backbone Taxonomy to
99 which it was matched when the dataset was imported, its taxonomic status and taxon rank, and the
100 names and taxonIDs of the higher taxa to which it belongs (kingdom, phylum, etc.).

101 For reproducibility, we used flatfiles of the latest available versions of the GBIF Backbone Taxonomy
102 (2021-11-26) and the NCBI Taxonomy (2022-12-01) instead of live online queries, so that the analysis
103 could be pinned to a specific version as these databases are continuously updated. For Wikidata, we
104 directly queried the online API instead of downloading a versioned flatfile, because database dump
105 files are large (2023-06-23 over 136 GB) and contain data on all entities, not just biological taxonomy.

106 GBIF taxonIDs in the dataset were matched against the GBIF Backbone Taxonomy to filter out
107 records that have been marked as “doubtful” or problematic, and to find currently accepted names and
108 taxonIDs within the GBIF Backbone Taxonomy, as the latter may have been updated after the dataset
109 was originally imported. This resulted in a table of taxon names (with authors) and taxonIDs of
110 interest. Only taxa of species rank (5721 names) were retained to simplify the search, as the higher
111 taxa can be derived from the list of species. From the NCBI Taxonomy, scientific names (including
112 authors where available) and taxonIDs at species rank classified to Viridiplantae ([NCBI:txid33090](https://ncbi.nlm.nih.gov/Taxonomy/NCBI/Taxa/Linux/Linux?taxid=33090))
113 were retrieved, to reduce the number of names to be searched, and to avoid hemihomonyms.



114
115 **Figure 1.** Simplified diagram of identifier linking through name matching. (A) Match accepted taxon names in
116 GBIF against names in NCBI Taxonomy using gndiff, then check if the respective identifiers are also linked in
117 Wikidata. (B) If an accepted name had no matches, retrieve synonyms for a second round of name matching.

118 The GBIF taxon names were matched against the Viridiplantae taxon names from NCBI with Gndiff
119 v0.2.0 (<https://github.com/gnames/gndiff>) (Figure 1A), which matches taxon names while accounting
120 for common orthographic variants, errors, and other issues specific to taxon names. Gndiff uses the
121 same algorithms as Gnverifier (<https://doi.org/10.5281/zenodo.5111542>) and Gnparser (Mozzherin,
122 Myltsev, and Patterson 2017) but can be used offline and without an external database. Gndiff reports
123 three types of matches: “Exact”, “PartialExact”, “Fuzzy”. We excluded “PartialExact” matches
124 because they encompass cases where only the genus name matches. “Fuzzy” matches include potential
125 misspellings, and so were retained. Gndiff parses the author field if present but does not take them into
126 account, so we further classified “Exact” matches into three types based on the author names: “exact”
127 – author names or citations identical, “noauthor” – author names absent from one or both entries
128 (typically from the NCBI record), “author_mismatch” – author names do not match exactly, which
129 includes differences in abbreviation or orthography. The result was a table of GBIF taxonIDs linked to
130 NCBI taxonIDs by name matching.

131 For GBIF names without matches in NCBI Taxonomy, synonyms according to the GBIF Backbone
132 were retrieved, and then used for a second round of name matching (Figure 1B). This was to account
133 for cases where the same taxon has different accepted names in the two databases. The two databases
134 handle synonyms differently. In GBIF, each distinct name has a different taxonID; accepted names vs.
135 synonyms are indicated in the “taxonomicStatus” field, but are maintained as different records. In the
136 NCBI Taxonomy, synonyms are given the same taxonID as the accepted name, and if an existing
137 taxon is deemed to be a synonym then its taxonID is moved to the “merged” list.

138 We queried Wikidata via its SPARQL API (<https://query.wikidata.org/>) for taxon items with the GBIF
139 taxonIDs from our dataset (property [P846](#)). If they were linked to an NCBI taxonID (property [P685](#)),
140 the linked NCBI taxonID was added to our table. If a taxon name was not linked to a Wikidata item
141 via its GBIF taxonID, but the earlier name matching had found an NCBI taxonID, then the NCBI
142 taxonID was used to query Wikidata to find linked Wikidata items and their associated GBIF
143 taxonIDs, if available.

144 The identifier links on Wikidata were then used to categorize the pairs of matched names for further
145 action (Table 1). The aim was to filter out names with no matches (Table 1, curation action “a, nothing
146 more to be done”) or unambiguous matches (Table 1, curation action “b, automatically accepted”)
147 from cases needing additional curation.

148 We identified straightforward cases of missing or outdated information in Wikidata where the
149 necessary updates can be executed through batch edits (Table 1, curation actions d and e). The criteria
150 for these were that GBIF and NCBI names had an exact match (including authorship) and the name
151 was accepted in the GBIF Backbone, but Wikidata either did not have one of the taxonIDs or had a
152 different taxonID from the currently accepted one. Commands for executing these edits in batches
153 with the QuickStatements tool (<https://quickstatements.toolforge.org/>) were generated.

154 **Table 1.** Possible outcomes of data linking steps, further curation steps to be taken, and the number of cases
 155 identified in this example dataset.

| Name match type | GBIF ID linked to Wikidata item? | Wikidata links GBIF to NCBI ID? | NCBI ID from name matching same as on Wikidata? | Wikidata links NCBI to GBIF ID? | Taxonomic status of name on GBIF | Curation action to be taken | Count |
|-----------------|----------------------------------|---------------------------------|---|---------------------------------|----------------------------------|--|-------|
| none | no | - | - | no | - | (a) No matches, including synonyms | 1310 |
| none | no | - | - | yes | - | Other | 8 |
| exact | yes | yes | yes | - | - | (b) Match ok, accept automatically | 3130 |
| exact | yes | yes | no | - | - | (c) Verify and update NCBI taxonID in Wikidata item | 11 |
| exact | yes | no | - | no | - | (d) Batch-add NCBI taxonID to Wikidata item | 177 |
| exact | yes | no | - | yes | - | Other | 52 |
| exact | no | - | - | yes | "accepted" | (e) Batch-update GBIF taxonID in Wikidata item | 245 |
| exact | no | - | - | yes | not "accepted" | (f) Verify if synonym listed in GBIF is valid before linking identifiers | 89 |
| noauthor | yes | yes | yes | - | - | (g) Verify if authorships match before linking identifiers | 211 |
| noauthor | yes | yes/no | no | - | - | (h) Possible homonym, investigate further | 224 |
| author mismatch | yes | yes | yes | - | - | (g) Verify if authorships match before linking identifiers | 271 |
| author mismatch | yes | yes/no | no | - | - | (h) Possible homonym, investigate further | 217 |
| fuzzy | - | - | - | - | - | other | 100 |

156

157 To understand the underlying causes for these erroneous links, we further investigated the cases where
 158 name-matching and Wikidata disagree on the GBIF taxonID (Table 1, curation action e). The current
 159 taxonomic status of the GBIF taxonIDs found in Wikidata was looked up in the GBIF Backbone
 160 Taxonomy. Of the 245 taxonIDs, two cases represented mismatched ranks (one genus and one
 161 subspecies), and another 83 (1.5% of 5721 total) had been removed by GBIF curators and were no
 162 longer listed in the GBIF Backbone, but these updates were not yet propagated to Wikidata. Almost all
 163 the remaining 162 (2.8%) appear to be names wrongly matched when identifiers were added to
 164 Wikidata because of homonymy, because the taxon authors differ.

165 The remaining cases were then tabulated for manual curation. This requires some knowledge of
 166 taxonomy and nomenclature rules to be able to evaluate whether two names are equivalent or not, as
 167 well as cross-checking against additional databases.

168 The above workflow is available from <https://github.com/monagrland/taxo-harmo>. The software
169 toolchain required is specified in a definition file for the Conda environment manager, using packages
170 distributed via the open-source conda-forge and bioconda channels (Grüning et al. 2018). The code to
171 reformat the input, perform the initial name-matching with Gndiff, query Wikidata for identifiers, and
172 prepare the tables for manual curation is listed and documented in a Jupyter notebook. The workflow
173 can be applied to other GBIF species list datasets simply by updating the filenames and the target
174 taxon group (if not Viridiplantae). Likewise, the pipeline can be re-run when newer versions of the
175 source databases are available.

176 [Guide to manual curation and improving community resources](#)

177 Here we describe what issues can be found during manual curation, and what concrete action users can
178 take to improve the database resources. In brief: Wikidata can be edited directly to fix errors or add
179 missing information, preferably after creating a user account; issues with the GBIF Backbone
180 Taxonomy can be reported via the website feedback dialog, by email, or via Github; issues with the
181 NCBI Taxonomy should be reported by email.

182 [\(A\) Errors due to name matching](#)

183 Error modes in name matching have been extensively discussed before (Patterson et al. 2016; Remsen
184 2016). In the curation process, homonyms can be quickly recognized by mismatches in authorships;
185 those links can be rejected unless they are simply orthographic differences such as the removal of
186 diacritics (e.g. “Hultén” vs. “Hulten”) or different abbreviation conventions (“Hook.f.” vs.
187 “Hook.fil.”). Typographical errors are to some extent ameliorated by the fuzzy matching in Gndiff.

188 **Example:** Name matching errors may also appear in the source databases. The original dataset lists the
189 genus *Ammophila* Kirby, 1798 (GBIF taxonID [1346141](#)), a genus of wasps, instead of the grass genus
190 *Ammophila* Host (GBIF [2703794](#)). Both names are valid under their respective, independent
191 nomenclatural codes, i.e. they are hemihomonyms. Here the error appears to have occurred during
192 import of the data from the original provider into GBIF.

193 **Action:** Accept or reject the linked identifiers after verification.

194 [\(B\) Errors or information gaps in databases](#)

195 If the results of name matching disagree with database identifiers, it is possible that one or more of the
196 source databases have incomplete or erroneous information.

197 *(1) GBIF taxonID has been deprecated or merged*

198 The GBIF Backbone Taxonomy is continually revised, and taxa may be deleted if they are e.g.
199 doubtful names, orthographic errors, or duplicates. However, the deprecated GBIF taxonIDs may still
200 be linked in Wikidata. In some cases, the accepted taxon in GBIF may also be in error (see point 6
201 below).

202 **Example:** Wikidata record for *Helianthus annuus* ([Q171497](#)) was linked to the GBIF taxonID
203 [3119195](#), which was deleted on 2018-02-01. The currently accepted GBIF record for this species is
204 [9206251](#).

205 **Action:** When unambiguous, edit the Wikidata entry to add the currently accepted GBIF taxon, after
206 checking that it is not a homonym. Record the access date in the reference with the property
207 “retrieved” ([P813](#)), which will help future editors troubleshoot if the GBIF record changes again. See
208 Shafee et al. (2023) for guidance on editing Wikidata.

209 *(2) NCBI taxonID has been deprecated or merged*

210 Unlike GBIF, the NCBI Taxonomy merges synonyms under the same taxonID, which can be
211 problematic if there is disagreement about whether two taxa are truly synonymous.

212 **Example:** *Calamagrostis stricta*, formerly [NCBI:txid497295](#), has been merged as a synonym of
213 *Calamagrostis neglecta* [NCBI:txid395286](#) in the NCBI Taxonomy. Furthermore, the GBIF Backbone
214 accepts *C. stricta* ([2704899](#)) while designating *C. neglecta* ([4104731](#)) as a synonym of *Achnatherum*
215 *calamagrostis* ([4142326](#)).

216 **Action:** Searching the NCBI website for a merged taxonID or entering its URL will auto-redirect to
217 the current accepted one. However, the ENA Taxonomy API
218 (<https://www.ebi.ac.uk/ena/taxonomy/rest/>), which in principle uses the same NCBI Taxonomy,
219 usually returns no result for merged taxonIDs, indicating that merged taxonIDs may cause problems
220 with downstream tools that do not take them into account. The currently accepted NCBI taxonID can
221 be added to the Wikidata entry, but the old taxonID may help disambiguate the record and should not
222 be deleted.

223 *(3) Incorrect species linked on Wikidata*

224 The Wikidata record may be linked to an identifier for a different species. These cases are usually
225 homonyms, which can be recognized by the different taxon author.

226 **Example:** The Wikidata record for *Rubus gracilis* C.Presl & J.S.Presl ([Q17248013](#)) was previously
227 linked to identifiers for the homonymous *Rubus gracilis* Roxb. in GBIF ([2990660](#)) as well as another
228 database, GRIN-Global ([32332](#), explicitly annotated as “non J.S.Presl & C.Presl 1822”).

229 **Action:** When unambiguous, edit the Wikidata entry to remove the incorrect statement, or point to the
230 correct identifier, if available. Record the access date using the Wikidata property “retrieved” ([P813](#)).
231 Different Wikidata items for homonymous taxa can be disambiguated with the property “different
232 from” ([P1889](#)).

233 *(4) Ambiguous entity in Wikidata - conflicting taxon authors*

234 Some cases may require taxonomic/nomenclatural expertise or additional information to resolve.

235 **Example:** The Wikidata record for *Willemetia stipitata* ([Q1362051](#)) states that the taxon author
236 (property [P405](#)) is Karl Wilhelm von Dalla Torre ([Q79155](#)). The linked GBIF entry ([5389300](#)) for *W.*
237 *stipitata* (Jacq.) Dalla Torre is annotated as “doubtful” in GBIF. Furthermore, the linked NCBI entry
238 ([NCBI:txid519273](#)) represents the homonym *W. stipitata* Cass. Linked records in other Wikis are also
239 inconsistent: German-language Wikipedia – *W. stipitata* (Jacq.) Dalla Torre
240 (<https://de.wikipedia.org/wiki/Kronenlattich>); Wikispecies – *W. stipitata* Cass.
241 (https://species.wikimedia.org/wiki/Willemetia_stipitata).

242 **Action:** The Wikidata entity may need to be split into separate entities for each homonym. Start a
243 thread on the corresponding discussion/talk page in Wikidata or Wikispecies to alert other users to the
244 issue. For one’s own research, make a judgement call and document it.

245 *(5) Ambiguous entity in Wikidata - no taxon author*

246 Some taxon names on Wikidata may lack the “taxon author” ([P405](#)) or “taxon author citation” ([P6507](#))
247 properties.

248 **Action:** As above. These should probably be split into separate entities if they are indeed homonyms,
249 but it would then be unclear how the linked identifiers should be distributed between them.

250 *(6) Error in accepted taxon in GBIF Backbone Taxonomy*

251 These can often be traced back to errors in the source datasets used to populate the GBIF Backbone.
252 The following example was found because the Wikidata entry was linked to both GBIF and NCBI
253 taxonIDs and agreed with the name-matching with Gndiff, but the author names conflicted.

254 **Example:** “*Primula matthioli* K.Richt.” is an accepted taxon in the GBIF Backbone Taxonomy
255 ([5640570](#)); GBIF’s source dataset for this name is “Synonymic checklists of the vascular plants of the
256 world” (Hassler 2022). However, the International Plant Names Index (IPNI), a nomenclatural
257 database for botanical names, only reports “*Primula matthioli* (L.) V.A.Richt.”
258 (<https://www.ipni.org/n/702251-1>). Wikidata records the same author as IPNI for *Primula matthioli*
259 ([Q50859720](#)), namely Vincenz Aladár Richter ([Q6163148](#)). GBIF annotates “*Primula matthioli* (L.)
260 V.A.Richt.” ([9764749](#)) as a “homotypic synonym”, and additionally has a record for “*Primula*
261 *matthioli* (L.) J.A.Richt. 1894” ([9781637](#)), also listed as a “homotypic synonym”.

262 Given the corroboration from IPNI, the author names in GBIF records [5640570](#) (“K.Richt.”) and
263 [9781637](#) (“(L.) J.A.Richt.”) are likely to be typographical errors for [9764749](#) (“(L.) V.A.Richt.”).

264 **Action:** Report errors or issues via the feedback system on the GBIF website (must be logged in with a
265 GBIF user account). Feedback reports are handled via the issue tracker on GitHub, and can also be
266 submitted directly there or by email. The issue opened for the above example is here:
267 <https://github.com/gbif/portal-feedback/issues/4673>. If their data curators can trace the issue to an

268 upstream data source, the report is passed upwards. Curators can also apply “patches” to the GBIF
269 Backbone Taxonomy, where the upstream source cannot be updated in a timely manner.

270 (7) *Error in accepted taxon in NCBI Taxonomy*

271 **Example:** *Carex binervis* Sm. (Wikidata [Q160245](#)) is an accepted taxon in the GBIF Taxonomy
272 ([2723521](#)), but the NCBI record had different authors “Gren. & Godr.” ([NCBI:txid372257](#)).

273 IPNI lists four homonyms for the name *Carex binervis*, but none with “Gren. & Godr.” as authors.
274 Only *C. binervis* Sm. is validly published (<https://www.ipni.org/?q=carex%20binervis>). The remainder
275 are either nom. inval., *C. binervis* Wahlenb. ex Kunth, or nom. illeg., *C. binervis* Willd. ex Kunth and
276 *C. binervis* Dewey, the latter according to Plants of the World Online
277 (<https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:77237975-1>).

278 “*Carex binervis* Gren. & Godr.” turned out to be a chresonym, where the authors after the binomen are
279 not the authors of the name itself, but a reference to a usage of the name in some other publication.
280 The Tropicos database has an entry for “*C. binervis* Gren. & Godr.” with a citation to the publication
281 *Flore de France* by Grenier & Godron (1855) (<http://legacy.tropicos.org/Name/9900008>). This
282 allowed us to find a digital copy online (<https://bibdigital.rjb.csic.es/idviewer/10272/430>) where the
283 name “*C. binervis* Sm.” is listed, showing that this was indeed the intended name.

284 Where did the NCBI Taxonomy find this chresonym “*Carex binervis* Gren. & Godr.”? The NCBI web
285 interface lists two references: Monocot Checklist (<http://www.kew.org/wcsp/home.do>, accessed 2010-
286 11-01), and a research paper (Villaverde et al. 2020). However, the former website is defunct and
287 redirects to Plants of the World Online, while cites only *Carex binervis* Sm. (Supplementary Table
288 S10). The incorrect taxon authors were presumably sourced from Tropicos or another database which
289 has since been updated or taken offline. Chresonyms look like legitimate taxon names with
290 authorships, and cannot be easily detected without cross-checking or conferring original sources, so
291 are especially prone to being propagated across aggregators.

292 **Action:** Report errors and updates to the NCBI helpdesk by email (Schoch et al. 2020). The example
293 above was reported and has already been corrected.

294 (8) *Disagreements in taxon concepts between databases*

295 The “same” taxon may appear under different names, classifications, or even be split or lumped into
296 different taxa, depending on the source consulted. Which names are accepted as valid, and which as
297 synonyms, are points of legitimate scientific disagreement; one name may represent different
298 taxonomic concepts. When data aggregators designate accepted names or use a particular
299 classification, they gloss over potentially valid taxonomic conflicts (Franz and Sterner 2018).

300 **Example:** The species *Rosa inodora* Fr. (GBIF taxonID [3002258](#), Wikidata [Q15844731](#)) in our
301 dataset does not have an NCBI taxonID, i.e. no sequence data is available. However, *Rosa elliptica*

302 Tausch (GBIF taxonID [3003248](#), Wikidata [Q9325795](#)), listed as a synonym of *Rosa inodora* by GBIF,
303 does have an NCBI taxonID ([NCBI:txid323240](#)).

304 **Action:** For the purposes of our own data analyses, we may choose to accept a taxonomic opinion and
305 link these taxa that are designated as synonyms by GBIF or NCBI. However, it would be inappropriate
306 to link the Wikidata item for *Rosa inodora* to the NCBI taxonID of *Rosa elliptica*, because they are
307 heterotypic synonyms that represent a taxonomic theory which is subject to potential disagreement and
308 future revision. Therefore, the original name of interest, accepted names, and synonyms are kept in
309 separate data columns in our workflow. In Wikidata, synonymous taxa can be represented by the
310 “taxon synonym” property ([P1420](#)), whereas homonyms can be disambiguated with the “different
311 from” property ([P1889](#)).

312 Discussion

313 The state of biodiversity identifier linking is patchy, even across well-resourced, heavily used
314 databases, and for well-studied sets of species like the German vascular plant flora. As expected, naive
315 name matching alone is problematic and can cause linking errors, affecting at least 2.8% of Wikidata
316 entries for the species names in the dataset examined here. Ironically, better studied groups and more
317 comprehensive databases may contain more historical names and homonyms that need to be accounted
318 for. Most of such linking errors are easily caught by using author names and higher taxa to
319 disambiguate taxa, allowing us to focus manual curation efforts on the most challenging cases.

320 Existing recommendations and workflows for taxon name harmonization (Grenié et al. 2021; Jin and
321 Yang 2020) recognize the same pitfalls of name matching and the limitations of source databases, such
322 as different accepted synonyms, inconsistent classifications, and lack of taxon author citations in some
323 datasets. Dealing with the name matching problem is by no means straightforward, as evidenced by
324 the infrastructure and numerous tools built by the Global Names Architecture (Gnames) project (Pyle
325 2016; Thessen et al. 2022; Mozzherin, Myltsev, and Patterson 2017), including the Gndiff tool used in
326 this workflow.

327 Generally, though, databases are presented as resources to be accepted as-is, over which the user has
328 no influence. Apart from simply filtering out problematic records, what more can be done? We
329 therefore suggest the following additional recommendations for users to be active participants and
330 help “pay it forward” in the community:

- 331 • Pay attention to potential synonyms and other taxonomic or nomenclatural issues when
332 designing a workflow, and choose software tools that can handle them, e.g. taxadb (Norman,
333 Chamberlain, and Boettiger 2020) or tools from Gnames.
- 334 • When publishing your own checklists, do not omit taxon authors and higher classification,
335 even when these details appear to be obvious from context.

- 336 • Report errors in source databases, as described in the examples above. Both GBIF and NCBI
337 have workflows for dealing with such reports and have been responsive to constructive
338 feedback, in our experience.
- 339 • Publish validated, linked identifiers on Wikidata. Each user will of course need to check for
340 themselves, but it helps subsequent users filter cases during data linking to focus manual
341 curation on the more problematic records. The Wikidata data model is highly extensible so it
342 is possible to perform sophisticated queries and integrate information about taxa with other
343 domains.

344 Why take the trouble to edit Wikidata and send feedback? Curation of biodiversity data is labor
345 intensive and requires a highly specialized skill set, so updating community resources will reduce
346 duplicated effort and have a positive, compounding effect (“virtue propagation”). Wikidata in
347 particular is increasingly integrated into the biodiversity informatics infrastructure, de facto
348 recognition of its practical usefulness: the database cross-references displayed on species pages on the
349 GBIF website (<https://www.gbif.org/species/search>) are sourced from Wikidata, and the iNaturalist
350 citizen science app uses Wikidata to link species pages to their respective Wikipedia articles in various
351 languages (Waagmeester et al. 2019). Applications beyond biodiversity show its versatility.

352 Communities can be built on top of Wikidata to curate specific knowledge domains such as gene
353 annotations (Putman et al. 2017); alternatively, existing wiki-type projects can be imported and
354 interlinked with Wikidata to foster data integration (Martens et al. 2021).

355 The workflow presented here still relies on ad hoc scripting, which is to some extent unavoidable
356 because the point of manual curation is to handle what automation cannot deal with, but it is desirable
357 to minimize this to improve reproducibility as well as the reusability of code. A promising alternative
358 is OpenRefine (<https://openrefine.org/>), a dedicated tool for data reconciliation, which records all data
359 cleaning steps in a given project, allowing them to be shared and re-run on new data. It also supports
360 querying and editing Wikidata within the software, as well as URL-based queries (e.g. calls to the
361 GBIF name parser API).

362 Routine sharing of curation workflows by researchers, coupled with the transparent handling of issue
363 reports by database maintainers, will foster more community buy-in and faster adoption of useful
364 practices, improving the quality of downstream analyses.

365 Acknowledgements

366 I thank C. Levers and W. Sickel for feedback on a draft of this manuscript, D. Mozzherin for help with
367 gndiff, and GBIF and NCBI Taxonomy curators for their responses to my queries and feedback.

368 References

369 Bundesamt für Naturschutz. 2021. “Flora von Deutschland (Phanerogamen).” Bundesamt für
370 Naturschutz / Netzwerk Phytodiversität Deutschland. <https://doi.org/10.15468/0FXSOX>.

371 Christen, Peter. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity*
372 *Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer Berlin Heidelberg.
373 <https://doi.org/10.1007/978-3-642-31164-2>.

374 Delpuch, Antonin, Adrian Pohl, Fabian Steeg, and Thad Guidry Sr. 2023. “Reconciliation Service
375 API v0.1 : A Protocol for Data Matching on the Web.” W3C Community Group Final Report.
376 Entity Reconciliation Community Group.
377 <https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.1-20230321/>.

378 Feng, Xiao, Brian J. Enquist, Daniel S. Park, Brad Boyle, David D. Breshears, Rachael V. Gallagher,
379 Aaron Lien, et al. 2022. “A Review of the Heterogeneous Landscape of Biodiversity
380 Databases: Opportunities and Challenges for a Synthesized Biodiversity Knowledge Base.”
381 *Global Ecology and Biogeography* 31 (7): 1242–60. <https://doi.org/10.1111/geb.13497>.

382 Franz, Nico M, and Beckett W Sterner. 2018. “To Increase Trust, Change the Social Design behind
383 Aggregated Biodiversity Data.” *Database* 2018 (January): bax100.
384 <https://doi.org/10.1093/database/bax100>.

385 GBIF Secretariat. 2022. “GBIF Backbone Taxonomy. Checklist Dataset.”
386 <https://doi.org/10.15468/39omei>.

387 Grenié, Matthias, Emilio Berti, Juan Carvajal-Quintero, Gala Mona Louise Dädlow, Alban Sagouis,
388 and Marten Winter. 2021. “Harmonizing Taxon Names in Biodiversity Data: A Review of
389 Tools, Databases and Best Practices.” *Methods in Ecology and Evolution* 14 (1): 12–25.
390 <https://doi.org/10.1111/2041-210X.13802>.

391 Groom, Quentin, Christian Bräuchler, Robert Cubey, Mathias Dillen, Pieter Huybrechts, Nicole
392 Kearney, Niels Klazenga, et al. 2022. “The Disambiguation of People Names in Biological
393 Collections.” *Biodiversity Data Journal* 10 (October): e86089.
394 <https://doi.org/10.3897/BDJ.10.e86089>.

395 Grüning, Björn, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H.
396 Tomkins-Tinch, Renan Valieris, and Johannes Köster. 2018. “Bioconda: Sustainable and
397 Comprehensive Software Distribution for the Life Sciences.” *Nature Methods* 15 (7): 475–76.
398 <https://doi.org/10.1038/s41592-018-0046-7>.

399 Hassler, Michael. 2022. “Synonymic Checklists of the Vascular Plants of the World.” *Catalogue of*
400 *Life*. <https://doi.org/10.48580/DFQT-3DD>.

401 International Commission on Zoological Nomenclature. 1999. *International Code of Zoological*
402 *Nomenclature*. Edited by W. D. L. Ride, H. G. Cogger, C. Dupuis, O. Kraus, A. Minelli, F. C.
403 Thompson, and P. K. Tubbs. 4. ed. London: International Trust for Zoological Nomenclature.
404 <https://www.iczn.org/the-code/the-code-online/>.

405 Jin, Jing, and Jun Yang. 2020. “BDcleaner: A Workflow for Cleaning Taxonomic and Geographic
406 Errors in Occurrence Data Archived in Biodiversity Databases.” *Global Ecology and*
407 *Conservation* 21 (March): e00852. <https://doi.org/10.1016/j.gecco.2019.e00852>.

408 Martens, Marvin, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina
409 Hanspers, Ryan A. Miller, et al. 2021. “WikiPathways: Connecting Communities.” *Nucleic*
410 *Acids Research* 49 (D1): D613–21. <https://doi.org/10.1093/nar/gkaa1024>.

411 Mozzherin, Dmitry Y., Alexander A. Myltsev, and David J. Patterson. 2017. “‘Gnparser’: A Powerful
412 Parser for Scientific Names Based on Parsing Expression Grammar.” *BMC Bioinformatics* 18
413 (1): 279. <https://doi.org/10.1186/s12859-017-1663-3>.

414 Norman, Kari E. A., Scott Chamberlain, and Carl Boettiger. 2020. “Taxadb: A High-Performance
415 Local Taxonomic Database Interface.” *Methods in Ecology and Evolution* 11 (9): 1153–59.
416 <https://doi.org/10.1111/2041-210X.13440>.

417 Page, Roderic D. M. 2022. “Wikidata and the Bibliography of Life.” *PeerJ* 10 (July): e13712.
418 <https://doi.org/10.7717/peerj.13712>.

419 Patterson, David, Dmitry Mozzherin, David Shorthouse, and Anne Thessen. 2016. “Challenges with
420 Using Names to Link Digital Biodiversity Information.” *Biodiversity Data Journal* 4 (May):
421 e8080. <https://doi.org/10.3897/BDJ.4.e8080>.

422 Putman, Tim E., Sebastien Lelong, Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Colin
423 Diesh, Nathan Dunn, Monica Munoz-Torres, et al. 2017. “WikiGenomes: An Open Web
424 Application for Community Consumption and Curation of Gene Annotation Data in
425 Wikidata.” *Database: The Journal of Biological Databases and Curation* 2017 (1): bax025.
426 <https://doi.org/10.1093/database/bax025>.

427 Pyle, Richard. 2016. "Towards a Global Names Architecture: The Future of Indexing Scientific
428 Names." *ZooKeys* 550 (July): 261–81. <https://doi.org/10.3897/zookeys.550.10009>.

429 Remsen, David. 2016. "The Use and Limits of Scientific Names in Biological Informatics." *ZooKeys*
430 550 (July): 207–23. <https://doi.org/10.3897/zookeys.550.9546>.

431 Schellenberger Costa, David, Gerhard Boehnisch, Martin Freiberg, Rafaël Govaerts, Matthias Grenié,
432 Michael Hassler, Jens Kattge, et al. 2023. "The Big Four of Plant Taxonomy – a Comparison
433 of Global Checklists of Vascular Plant Names." *New Phytologist*, May.
434 <https://doi.org/10.1111/nph.18961>.

435 Schoch, Conrad L, Stacy Ciuffo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda
436 Khovanskaya, Detlef Leipe, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on
437 Curation, Resources and Tools." *Database: The Journal of Biological Databases and*
438 *Curation* 2020 (August): baaa062. <https://doi.org/10.1093/database/baaa062>.

439 Shafee, Thomas, Daniel Mietchen, Tiago Lubiana, Dariusz Jemielniak, and Andra Waagmeester.
440 2023. "Ten Quick Tips for Editing Wikidata." *PLOS Computational Biology* 19 (7):
441 e1011235. <https://doi.org/10.1371/journal.pcbi.1011235>.

442 Thessen, Anne E., Dmitry Mozzherin, David Peter Shorthouse, and David J. Patterson. 2022.
443 "Improving the Discoverability of Biodiversity Data Using the Global Names Finder."
444 *Biodiversity Information Science and Standards* 6 (December): e90026.
445 <https://doi.org/10.3897/biss.6.90026>.

446 Thessen, Anne E., Jorrit H. Poelen, Matthew Collins, and Jen Hammock. 2018. "20 GB in 10 Minutes:
447 A Case for Linking Major Biodiversity Databases Using an Open Socio-Technical
448 Infrastructure and a Pragmatic, Cross-Institutional Collaboration." *PeerJ Computer Science* 4
449 (September): e164. <https://doi.org/10.7717/peerj-cs.164>.

450 Turland, Nicholas J., John H. Wiersema, Fred R. Barrie, Werner Greuter, D. L. Hawksworth, Patrick
451 S. Herendeen, Sandra Knapp, et al., eds. 2018. *International Code of Nomenclature for Algae,*
452 *Fungi, and Plants (Shenzhen Code): Adopted by the Nineteenth International Botanical*
453 *Congress Shenzhen, China, July 2017*. Regnum Vegetabile 159. Glashütten: Koeltz Scientific
454 Books. <https://www.iapt-taxon.org/nomen/main.php>.

455 Villaverde, Tamara, Pedro Jiménez-Mejías, Modesto Luceño, Marcia J Waterway, Sangtae Kim, Bora
456 Lee, Mario Rincón-Barrado, et al. 2020. "A New Classification of *Carex* (Cyperaceae)
457 Subgenera Supported by a HybSeq Backbone Phylogenetic Tree." *Botanical Journal of the*
458 *Linnean Society* 194 (2): 141–63. <https://doi.org/10.1093/botlinnean/boaa042>.

459 Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase."
460 *Communications of the ACM* 57 (10): 78–85. <https://doi.org/10.1145/2629489>.

461 Waagmeester, Andra, Daniel Mietchen, Siobhan Leachman, and Quentin Groom. 2019. "Using
462 Crowd-Curation to Improve Taxon Annotations on the Wikimedia Infrastructure."
463 *Biodiversity Information Science and Standards* 3 (June): e35216.
464 <https://doi.org/10.3897/biss.3.35216>.

465 Waagmeester, Andra, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good,
466 Malachi Griffith, Obi L Griffith, Kristina Hanspers, et al. 2020. "Wikidata as a Knowledge
467 Graph for the Life Sciences." Edited by Peter Rodgers and Chris Mungall. *ELife* 9 (March):
468 e52614. <https://doi.org/10.7554/eLife.52614>.

469