A novel curation system to facilitate data integration across regional citizen science survey programs

Dana L Campbell [Corresp., 1], **Anne E Thessen** [2, 3], **Leslie Ries** [Corresp. 4]

[1] Division of Biological Sciences; School of STEM, University of Washington, Bothell, Washington, United States

[2] The Ronin Institute for Independent Scholarship, Montclair, New Jersey, United States

[3] Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon, United States

[4] Department of Biology, Georgetown University, Washington, DC, United States

Corresponding Authors: Dana L Campbell, Leslie Ries
Email address: danalc@uw.edu, Leslie.Ries@georgetown.edu

Integrative modeling methods can now enable macrosystem-level understandings of biodiversity patterns, such as range changes resulting from shifts in climate or land use, by aggregating species-level data across multiple monitoring sources. This requires ensuring that taxon interpretations match up across different sources. While encouraging checklist standardization is certainly an option, coercing programs to change species lists they have used consistently for decades is rarely successful. Here we demonstrate a novel approach for tracking equivalent names and concepts, applied to a network of 10 regional programs that use the same protocols (so-called "Pollard walks") to monitor butterflies across America north of Mexico. Our system involves, for each monitoring program, associating the taxonomic authority (in this case one of three North American butterfly fauna treatments: Pelham, 2014; NABA, 2016; Opler & Warren, 2003) that shares the most similar overall taxonomic interpretation to the program's working species list. This allows us to define each term on each program's list in the context of the appropriate authority's species concept and curate the term alongside its authoritative concept. We then aligned the names representing equivalent taxonomic concepts among the three authorities. These stepping stones allow us to bridge a species concept from one program's species list to the name of the equivalent in any other program, through the intermediary scaffolding of aligned authoritative taxon concepts. Using a software tool we developed to access our curation system, a user can link equivalent species concepts between data collecting agencies with no specialized knowledge of taxonomic complexities.

1 **A novel curation to facilitate data integration across regional citizen science survey**

2 **programs**

4 Dana L. Campbell[1], Anne E. Thessen[2,3], Leslie Ries[4]

6 [1]Division of Biological Sciences, School of STEM, University of Washington, Bothell, WA,

7 USA

8 [2]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, USA

9 [3]The Ronin Institute for Independent Scholarship, Montclair, NJ, USA

10 [4]Department of Biology, Georgetown University, Washington, DC, USA

12 Corresponding Authors:

13 Dana Campbell[1]

14 Division of Biological Sciences, School of STEM, University of Washington, Bothell, WA, USA

15 Email address: danalc@uw.edu

17 Leslie Ries[4]

18 Department of Biology, Georgetown University, Washington, DC, 20057

19 Email address: leslie.ries@georgetown.edu

Abstract

21   Integrative modeling methods, such as niche modeling, can now enable macrosystem-level

22   understandings of biodiversity patterns, such as range changes resulting from shifts in climate or

23   land use, by aggregating species-level data across multiple monitoring sources. This requires

24   ensuring that taxonomic entities match up across different sources. While checklist

25   standardization is certainly an option, requesting programs to change species lists they have used

26   consistently for decades is rarely successful. Here we demonstrate a novel approach for tracking

27   equivalent names and concepts, applied to a network of 10 regional programs that use the same

28   protocols (so-called "Pollard walks") to monitor butterflies across America north of Mexico. Our

29   system involves, for each monitoring program, identifying the taxonomic authority (in this case

30   one of three North American butterfly fauna treatments: Pelham, 2014; NABA, 2016; Opler &

31   Warren, 2003) with the most similar taxonomy to the program's working species list. This allows

32   us to define each term on each program's list in the context of the appropriate authority's species

33   concept and associate it with its authoritative concept. We then aligned the names representing

34   equivalent taxonomic concepts among the three authorities. These stepping stones allow us to

35   bridge a species concept from one program's species list to the name of the equivalent in any

36   other program, through the intermediary aligned authoritative taxon concepts. Using a software

37   tool we developed to access our curation system, a user can link equivalent species concepts

38   between data collecting agencies with no specialized knowledge of taxonomic complexities.

39   After integration, taxonomically standardized datasets can subsequently be used as input for

40   many kinds of biodiversity analyses.

Introduction

42 There is a long history of piecing together multiple data sets to understand species distributions

43 and patterns at large spatial and temporal extents, including the building of traditional range

44 maps, range changes predicted by niche models (e.g., Graham et al. 2004) and the emerging field

45 of macrosystems ecology that focuses on cross-scale dynamics (e.g., Heffernan et al., 2014).

46 Citizen science-powered monitoring programs now contribute prolific amounts of data towards

47 these ends. Many programs carry out replicated surveys on permanent sites, using formal

48 protocols similar to those in traditional scientific surveys (Kelling et al., 2019). These programs

49 often operate on local scales to collect robust abundance, effort and absence data amenable for

50 developing population indices across space and time, critical metrics for tracking factors such as

51 population declines or shifts in phenology (see Pollock et al., 2002; Schwanghart, Beck & Kuhn,

52 2008; Pautasso & Weisberg, 2008). For interpreting changes occurring across populations, there

53 is great value in integrating data from monitoring programs operating at different spatial scales

54 (Downes et al., 2005). This has motivated rapid development of integrative modeling methods to

55 support large-scale analyses (Pautasso & Weisberg, 2008; Levy et al., 2014; Zipkin & Saunders,

56 2018; Kéry, 2018).

57

58 Despite newly-emerging analytical tools, actual implementations of integrative models are still

59 rare. For example, in Cooper, Shirk & Zuckergerg's (2014) review of 171 macroscale studies of

60 migratory birds, we note that while 72 of the studies (42%) used data from regional or

61 continental-scale citizen scientist programs with protocols, 40% of those are based on a single

62 source of data. Of the remaining studies, more than half used data from banding stations which

63 benefit from centralized taxonomic standards followed by most participating programs.

64   Published analyses from another highly monitored taxonomic group, the butterflies, also

65   typically include only a single data source. Papers on US butterfly species, including studies

66   examining the influence of environmental variables on entire butterfly communities (e.g.,

67   Forister & Shapiro, 2003; Forister et al., 2010; Diamond et al., 2014; Cayton et al., 2015;

68   Thorson et al., 2016, Wepprich et al. 2019) all use data originating from a single program and

69   thus did not require any data integration. In studies including data from more than one source

70   (e.g. DeVictor et al., 2012), the data from each source were analyzed separately (Schmucki et al.,

71   2016). We could find only one published study (Mills et al., 2017) that performed an analysis

72   integrating butterfly species concepts across multiple monitoring schemes.

73

74   Integrating data among multiple sources may be largely constrained by the challenge of

75   consistent taxon matching. The clear trend in data collection is towards observation-based

76   methods (such as citizen science monitoring programs) that do not generate vouchered specimen.

77   Amano, Lamming & Sutherland (2016) documented that only 50% of all records in the Global

78   Biodiversity Information Facility (GBIF) are associated with a vouchered specimen, and that

79   number drops to 10% when looking just over the past decade. Furthermore, we found 98% of the

80   records put into the GBIF-SCAN (Symbiota Collections of Arthropods Network; the main

81   digitization portal for Lepidoptera occurrence records) in the last 10 years report human

82   observation as basis of record (SCAN-GBIF analysis not presented). This means that in most

83   cases names are the primary identifier for connecting comparable observations across diverse

84   data sources. Integration of data stored in digital repositories such as the GBIF requires

85   resolution of 1) names that contain typographical or formatting issues; 2) synonyms, i.e. different

86   names used for the same taxonomic concept; and, most complex, 3) names that follow different

87   authorities, that is, the same name used for different taxonomic concepts (Boyle et al., 2013).

88

89   For most species-level distribution research, standards for aggregating data consist of resolving

90   deviant spellings, grammatical differences (e.g. hyphens and spaces), or synonymous names, but

91   do not necessarily account for different taxon concepts circumscribed by a species name

92   (Kennedy, Kukla & Patterson, 2005; Giangrande, 2003; Pyle, 2004, Patterson et al., 2010; Mora

93   et al., 2011; Patterson et al., 2016; Remsen, 2016). Systematic and phylogenetic research

94   continuously reclassifies the biological entities circumscribed by a name; thus unless different

95   data sources explicitly reference the same (or compatible) authoritative treatments, it cannot be

96   assumed that they monitor the same taxonomic concept even when they use the exact same name

97   (Ytow, Morse, & Roberts, 2001; Pyle, 2004; Kennedy, Kukla & Paterson, 2005; Franz, Peet &

98   Weakley, 2008; Franz & Peet 2009; Boyle, 2013; Cui et al., 2016; Remsen, 2016). This is

99   problematic especially for unstable taxa, for which different taxonomic studies suggest

100   reorganizations and disputed relationships (Ytow, Morse, & Roberts, 2001).

101

102   Franz et al. (2016) illustrated the scope of this problem within the plant *Andropogon glomeratus-*

103   *virginicus* "complex" (Poaceae). Careful examination of this species group revealed that multiple

104   different sources of data collected over 126 years were attributed to a single species, but in fact

105   the name under which these data were collected encompassed 12 different taxonomic concepts,

106   because the sources collecting the data used different classifications of the taxa within this

107   complex. Franz et al. (2016) found that names were reliable identifiers of taxonomic concepts in

108   only 60% of the 12 pairwise alignments, indicating frequent mismatch between the taxonomic

109   entities and nomenclature. Another study, based on the American Ornithological Union (AOU)'s

110   North American checklist, shows that in the past 127 years 25% of North American bird names

111   represent taxon concepts that were either split from or combined with another taxon, and of those

112   names, almost half were then subsequently revised at least once (Vaidya, Lepage & Guralnick,

113   2018). An identical nomenclature used in gathering data before and after a revision may

114   represent non-combinable records for non-identical taxonomic entities. We highlight an example

115   of these issues for butterflies, the *Celastrina ladon* complex, in which, because of a history of

116   multiple taxonomic interpretations, observations collected under the exact same name often do

117   not actually represent the same species (Fig. 1). We also identify other problematic butterfly

118   complexes encountered by North American monitoring programs.

119

120   Tools have been developed to crosswalk among naming systems, such as Avibase for birds

121   (Lepage, Vaidya & Guralnick, 2014) and the Taxonomic Names Resolution Service (TNRS) for

122   plants (Boyle et al., 2013), and standardized reference data sets are available to use for

123   vertebrates (Zermoglio, Guralnick & Wieczorek, 2016). The development of global taxonomic

124   lists, such as the Integrated Taxonomic Information System (itis.gov) and Catalogue of Life

125   (catalogueoflife.org) is another approach to standardize names across databases. More recently,

126   the Global Names Architecture (GNA) was developed with the goal to find, index, and resolve

127   taxonomic discrepancies among digitized taxonomic resources (Pyle, 2016). While these efforts

128   have greatly increased our ability to link knowledge across platforms focused on biodiversity

129   data (Patterson et al., 2016), they have not solved all the problems related to data integration

130   based on identity by name, even for the taxa for which the tools were created.

131

132   In this paper, we present a curation-based solution for associating comparable taxonomic

133   identities in the context of citizen science-powered monitoring, exemplified by North American

134   butterfly monitoring programs. Throughout this paper we use the term North American

135   butterflies to mean those north of Mexico. A protocol for surveying butterflies was developed

136   over 40 years ago specifically for volunteer monitors (Pollard, 1977). This was well before the

137   current explosion of citizen science monitoring programs (see Pocock et al. 2014 for

138   development in Great Britain), and represents an early example of using structured protocols (i.e.

139   where permanent routes are set up and monitored regularly using a set protocol, producing the

140   most robust of survey data; Kelling et al. 2019). This protocol, referred to as the "Pollard Walk"

141   or "Pollard Transect," involves monitors walking assigned routes at a steady pace and recording

142   butterflies in a restricted observation window. The UK Butterfly Monitoring Scheme adopted the

143   Pollard Walk methodology in 1976 (Schmucki et al., 2016). Other European countries and

144   regions in the US (usually states) then followed throughout the 1980s, 90s and 2000s (Taron &

145   Ries, 2015; Van Swaay et al. 2015; Schmucki et al., 2016). In 2014, a web-enabled data entry

146   platform called PollardBase (pollardbase.org) was launched to capture, manage, share, and

147   archive North American butterfly monitoring data. PollardBase has since attracted the

148   membership of most North American, Pollard-based programs (Taron & Ries, 2015).

149

150   For butterflies, most structured survey monitoring takes place on a regional basis, so any large-

151   scale analyses will necessarily have to integrate data across programs (Schmucki et al., 2016).

152   Any mismatches or confusion about names could lead to profound biases in the data, especially

153   if a mismatch causes spurious zeros to appear (Royle Nichols & Kéry, 2005; Pautasso &

154   Weisberg, 2008). Inflated zeros strongly skew underlying data distributions and would

155   fundamentally bias any analyses (Kéry, 2018). As large-scale analyses of both pattern and

156   processes are becoming more common, especially through the emergence of the field of

157   macrosystems ecology (Heffernan et al., 2014) and the advancement of integrated statistical

158   models (Zipkin & Saunders, 2018), the need to make data interoperable without each end-user

159   knowing the complex and often convoluted taxonomic history of all the species in the regional

160   community will become more and more vital (Levy et al., 2014; Kelling et al., 2019). Further, as

161   integrated modeling schemes combine different types of distribution data, including counts,

162   presence/absence, and presence-only data (Kery, 2018; Zipkin & Saunders 2018), the system we

163   discuss here could potentially help ensure compatibility for curators of many different data types

164   (e.g., iNaturalist, Butterfliesandmoths.org, eButterfly).

165

166   North American butterfly nomenclature is complex and controversial, and significant

167   disagreements have spawned multiple differing lists of species names. As a case in point,

168   estimates of the number of butterfly species in the United States and Canada range from 679

169   (Scott, 1986) to 780 (Opler & Warren, 2003) to 822 (Pelham, 2014). This is in stark contrast

170   with the nomenclature of European butterflies and of North American birds. For birds,

171   nomenclature is actively standardized by the AOU and the North American Classification

172   Committee (NACC), and Avibase is an additional resource in resolving discrepancies (Lepage,

173   Vaidya & Guralnick, 2014; Chesser et al. 2016; Lepage, 2017). Without any central taxonomic

174   authority to reference, North American butterfly monitoring programs follow survey lists that

175   reflect the idiosyncrasies, local expertise, and historical practices of their program. Thus naming

176   discrepancies are much more common among butterfly datasets than bird programs and provide

177   an excellent example for testing the effectiveness of our system.

178

179   To reduce confusion and ensure conformity of observations, it is common for monitoring

180   programs to record data using their own unique and stable species checklist without regularly

181   updating to newly published names or systematic reorganizations (Ytow, Morse & Roberts 2001;

182   Taron & Ries, 2015). Although willing to unite under a shared data management platform,

183   PollardBase programs are disinclined to adopt a single standardized nomenclatural system and

184   there is no way to compel them to do this. Even if it were possible, introducing a standardized

185   nomenclature does nothing to resolve differences in legacy data that have accumulated over 40

186   years of monitoring in the US and Canada. We thus developed our curation system to recognize

187   and automatically match taxonomic entities across PollardBase programs, while allowing each

188   program to retain their own unique checklist. All 21 programs that have joined PollardBase

189   (including the 10 programs in this paper) have agreed to adopt this system. Even in the unlikely

190   case where a group did not wish to cooperate, it would still possible to code their list in a way

191   that adheres to the underlying rules of our system for the purpose of combining data.

192

193   We built our system to resolve alternative names, spellings and taxonomic concepts between any

194   two monitoring programs through the intermediary identification of each program with its

195   underlying authoritative taxonomic treatment. Our system can also address the problem of

196   recording cases where field identification is difficult and a program director chose to have

197   volunteers record certain observations under a name that combines species. Additionally, our

198   system is easily expandable, allowing any monitoring program to opt in at any point. By

199   identifying compatible taxon concepts across programs, our system indicates which data can be

200   combined, thus providing true data interoperability while "requir[ing] little specialized effort on

201   the part of the end user," a major goal in the field of informatics (Wilkinson et al. 2016).

202

203   Development of a novel name curation system

204   To evaluate the usefulness of our system, we show how it works in PollardBase and provide a

205   proof-of-concept that the system can be used to automate the integration of data. Note that the

206   PollardBase platform is designed so that volunteers collecting monitoring data for a particular

207   program can choose only among the official species names included on that program's checklist

208   (which were deliberately specified by the program director). Thus this study, based in the context

209   of 10 PollardBase-registered (for current or future use) programs, allows us to quantify the

210   challenges for data integration present even in this ideal situation where the system controls for

211   data entry level introductions of typographical/spelling errors and for accidental use of an

212   unofficial name (this is the most common source of name discrepancies in data collected outside

213   of platforms like PollardBase). We targeted three specific objectives in developing and testing

214   this new system:

215   • Presentation of a curation system based on an authoritative reference alignment for

216   defining taxonomic concepts

217   • Quantification of the challenges for integrating data across 10 regional butterfly

218   monitoring programs

219   • Development and testing of an automated integration tool to implement our curation

220   system.

221

222   **Objective 1. Presentation of a curation system based on an authoritative reference**

223   **alignment for defining taxonomic concepts.**

224

225   **Step 1: Identify authoritative lists of North American (US and Canada) butterflies and**

226   **assign authority to each program's list.** We identified three research-based, taxonomic

227   compendiums that consider the complete butterfly fauna of the United States and Canada: North

228   American Butterfly Association (NABA) Checklist 2nd edition (version 2.3) and updates (NABA

229   2016; henceforth referred to as NABA); Opler & Warren (2003; henceforth referred to as Opler

230   & Warren); and Pelham (2014; henceforth referred to as Pelham). More information on these

231   authorities is in Supplemental Article S1. We ascertained for the 10 PollardBase programs that

232   we sample in this paper (described in more detail below), that each uses a checklist similar to one

233   of these authoritative treatments. In addition, we also include in our analyses the Integrated

234   Taxonomic Information System (ITIS) species list (which we restricted to include just American

235   butterfly species north of Mexico), because of its wider familiarity across the community of

236   researchers and potential to serve as a link with other taxonomic databases. Table 1 shows the

237   numbers of names each authority lists at different taxonomic levels.

238

239   PollardBase requests that each registered monitoring program provide its species checklist used

240   for data collection in their locale, and, if known, the source of that list. The source could be a

241   regional authority (e.g., a published regional field guide) or a national, research-based systematic

242   authority. For programs that did not specifically base their list on NABA, Opler & Warren or

243   Pelham, PollardBase administrators (which include all three authors of this paper) determined

244   which of these three base authorities it most closely resembled, and assigned that as the

245   program's "base list." None of the 10 programs in this study declare ITIS as an authoritative

246   basis for their own list.

247

248   **Step 2: Describing types of taxonomic deviations between base authority lists and in**

249   **comparisons between programs and their base authority.** We manually aligned the three

250   authoritative lists (NABA, Opler & Warren, Pelham) and the ITIS reference list to identify where

251   nomenclatural differences lie. We used the notes published along with the authoritative lists to

252   help interpret the taxon concepts represented by each list's taxon names and align equivalent

253   concepts across the authorities (Cassie et al., 2001; Opler & Warren 2003; J Pelham, 2016, pers.

254   comm.). Rarely, we also consulted further literature for clarification. Supplemental Table S1

255   shows the alignment of all taxa that have one or more discrepancies among the four lists. This

256   alignment acts as the backbone that ultimately allows us to associate taxonomic concepts

257   between two programs whose checklists are based on different authorities (described in

258   Objective 3).

259

260   Next, we manually compared each program's list with its specified base list to identify all cases

261   where an entry on a program's species list differed from its specified base authority. Programs

262   were contacted to determine if identified deviations were intentional or if the program director

263   might be willing to modify their list to match their authoritative base list. Note that this

264   represented our only attempt to persuade programs to adopt a different nomenclature, as most

265   programs were very clear before joining PollardBase that they would retain control over their

266   species lists (Ries, personal communication with directors from all 10 North American butterfly

267   monitoring programs included in this paper). Program managers agreed to change the name to

268   more fully adhere to their matching authority in about 50% of these cases. Reasons for not

269   changing the name usually stemmed from local custom or the program director's (or local

270   expert's) belief that the name on their list is the appropriate one for them (despite potential

271   conflict with current taxonomic findings). This discussion with the program managers helped

272   PollardBase administrators clarify the precise taxon concept represented by the name on the

273   program list. In cases where we could not identify why the program's list deviated from their

274   base list, we considered the program's discrepancy a discrete difference from its base and

275   assumed the base authority's taxonomic concepts for all other non-divergent taxa. Occasionally

276   ambiguities required us to ask program directors to interpret the implications of their deviation.

277

278   In all, we found six distinct types of deviations which we define, with examples, in Table 2.

279   Some discrepancy types are relevant only between base lists, some only relevant for base

280   list/program list comparisons, and some are relevant in both types of comparisons. We further

281   quantify and characterize these discrepancies in Step 3, below. While many others have noted

282   similar naming differences (for example, Vaidya, Lepage & Guralnick (2018) discuss lumps and

283   splits of taxa that would fall within and explain much of our "Species or subspecies

284   conglomerates" category and issues of extralimital taxa that we categorize into "Unmatched

285   taxa;" likewise Ytow, Morse & Roberts (2001) discuss differently circumscribed taxa), we have

286   not seen as comprehensive a categorization of difference types among names as the one we lay

287   out here.

288

289   **Step 3: Characterizing and quantifying taxonomic deviations among lists.**

290 **Pairwise comparisons between base lists**. Comparisons showed 603 species names that are

291 identical (i.e. representing identical taxon concepts) across the three base authority lists and ITIS.

292 This is 71% of the total number of 849 aligned taxon concepts. Differences in subfamily name

293 were the most common type of deviations among base lists, but only because they were

294 calculated on a per name basis for the entire classification hierarchy, so one change in subfamily

295 name causes as many discrepancies as there are children taxa. Pairwise comparisons of base lists

296 showed between 40 and 257 differences at the subfamily level. Subfamily names generally don't

297 impact the issues of data integration because program lists are based on lower level taxonomy, so

298 hereafter we do not address issues of subfamily deviations but focus instead on pairwise

299 differences at the genus and species level (with some subspecies differences causing species

300 deviations). Table 3 shows the number of each deviation type that occurs between all possible

301 pairwise comparisons of the base lists. ITIS and Opler & Warren lists are the most similar to

302 each other overall, with a total of 5% of names differing. Pelham and NABA, the most different

303 lists, deviate at 23.5% of compared names.

304

305 As we show later, species type deviations include the most difficult discrepancies to resolve

306 (although many are trivial). Species-level deviations among base lists affected up to 12.7% of all

307 species. Genus and spelling discrepancies were rarer, up to 4.4% and 1.6% respectively, but

308 these deviations are generally easily resolvable once detected because in all cases the alternative

309 names align 1:1 with the taxonomic concept they represent. Deviations due to unmatched taxa

310 are also relatively rare (up to 5.1%). We consulted the literature to assure that the missing taxon

311 was actually excluded from the list and not included as a subspecies within another species'

312 concept.

313

314   **Pairwise comparisons between programs and their base authority.** Table 4 shows the types

315   of deviations found between the program lists and their base authorities. Florida's list was the

316   most similar to its base, with 1% of names showing deviations (Table 4). Orange County

317   (California) has the largest number of deviations for its list size (7.6% of names). As in base list

318   comparisons, genus deviations and spelling differences between program and base lists were

319   trivial to resolve. However, other deviation classes can be difficult to disentangle, especially if

320   the person who compiled the list is not available to clarify a choice of nomenclature. Species

321   name deviations and unmatched taxa represent the most difficult type of deviation to resolve

322   because they often arise from fundamental disagreement about taxonomic concepts that are not

323   always articulated in a published source. Species and species groups that have been revised

324   multiple times can have a complex history of splitting/lumping disagreements (as demonstrated

325   for birds in Vaidya, Lepage & Guralnick 2018), creating much confusion when the same species

326   name is used by different programs to refer to a broader or narrower inclusion of subspecies

327   and/or populations. Even if the nomenclature has been taxonomically resolved in the literature,

328   programs retain their own species interpretations. This means that translating accurately from

329   one program to another requires a thorough understanding of each program's interpretation of

330   the nomenclature, complicating the process of establishing the match for each program pair.

331

332   **Objective 2. Quantification of the challenges for integrating data across 10 regional**

333   **butterfly monitoring programs.**

334

335   One of our goals for this project was to quantify the challenges of taxonomic resolution problems

336   across a sample of programs. For this implementation, we present just the simplest use case: that

337   one of the 10 monitoring programs (the receiver) requests to receive data from one of the other

338   monitoring programs (the donor) using the receiver's taxonomic interpretation to compare

339   compatibility between the two programs. In reality, a more likely scenario is a third party

340   requesting data from multiple programs. We do not consider this more complex case here, which

341   requires the third party to provide their own taxonomic preferences for a final list, but consider

342   this an important feature for future iterations of our tool.

343

344   The 10 regional butterfly programs included in this analysis cover a total of 489 distinct,

345   comparable names, which we aligned manually (see Supplemental Table S2). Of these, 82%

346   either matched exactly across all programs in which they occurred (182 names), or were present

347   in only one program list (208 names). In our alignment we took into account the taxon concept

348   each of the 93 names (18%) that did not match exactly across programs. These discrepancies are

349   categorized and summarized in Table 5, with more detail provided on the differences in

350   Supplemental Table S3. Most deviation instances were simple issues (category 1 in Table 5), in

351   which programs names differed by simple spelling or synonym variants, but aligned 1:1 with

352   taxon concept. More difficult were subspecies issues (category 2 in Table 5), where some

353   programs listed names that did not align 1:1 across comparisons. The third category of deviations

354   in Table 5 shows the most difficult discrepancies, including conglomerates and complexes,

355   which present much more difficult alignment issues.

356

357   Our alignment of program names guided our development of a set of species concept matrices

358   that specify the comparable taxonomic unit for every possible pairwise combination in which

359   one of the 10 programs (receiver) requests data from another program (donor), based on the

360   receiver's taxonomic perspective. We use our set of matrices to quantify the challenges of

361   integration and also as a validation check for our translation tool (Objective 3).

362

363   As an illustration of our process to determine comparable taxonomic units across programs, which

364   starts by identifying taxa that do not exactly align and then translating each case into a taxon concept

365   matrix, we discuss here one of our most difficult taxonomic cases, the *Celastrina ladon* complex. This

366   complex is composed of seven taxa (approximate ranges shown in Fig. 1), which are treated

367   differently by each of the authoritative base lists. The *C. ladon* complex is problematic partly because

368   the taxonomic entities can be difficult to distinguish morphologically. Pelham and Opler & Warren

369   elevate all taxa to full species status (although Opler & Warren does not treat *serotina*), whereas

370   NABA considers all as subtaxa within *C. ladon*. In total, our programs monitor five of these taxa

371   (Table 6). Because of conflicting identities, determination of data compatibility between two programs

372   requires careful consideration of the taxonomic concepts used by each program, even when the

373   recorded names are identical.

374

375   For instance, as shown in Figs. 1 and 2, both *ladon* and *neglecta* occur in the area monitored by the

376   Iowa program. Iowa, following NABA's taxonomic interpretation, considers these taxa as subspecies,

377   and collects observations for both together under the species name "*C. ladon*." Conversely, the Florida

378   program, basing its perspective on Opler & Warren's taxonomy, takes observations of *C. ladon* and *C.*

379   *neglecta* separately, as distinct species. Thus, despite sharing the same name, Iowa's "*C. ladon*"

380 observations are not compatible with "*C. ladon*" observations from Florida. Iowa might consider data

381 compatible only when Florida's data from both *C. ladon* and *C. neglecta* are included in comparison

382 with Iowa *C. ladon* data, however a researcher wanting to integrate data could easily overlook this

383 compatibility complication. From Florida's perspective, Iowa's data are not compatible at the species

384 level for either *C. ladon* or *C. neglecta* individually. Similarly, the Cascades program, following

385 Pelham's taxonomy, collects distinct observations for *C. lucia* and *C. echo* as separate species. Thus,

386 from Cascades' taxonomic interpretation, there are no data from the NABA-based Iowa program that

387 is compatible with either *C. lucia* or *C. echo*. On the other hand, Iowa interprets combination as valid

388 as long as both *echo* and *lucia* data from Cascades are included with Iowa's *C. ladon* data. These

389 examples show how the taxonomic interpretation of the monitoring programs fundamentally affect

390 compatibility of the data they collect; data may be compatible between programs when one of the

391 programs is the recipient but not when the other program is; that is, requests may not have a

392 symmetrical solution when flipped the other direction. Furthermore, in the case of the *C. ladon*

393 complex, half the monitoring programs deviate from the interpretation of their own base authority (see

394 Table 6), emphasizing the potential complexity of deciphering compatibility between program

395 observations.

396

397 We built a species concept integration matrix for the *C. ladon* complex indicating the

398 compatibility of data collected by each receiver program with data collected by any of the nine

399 potential donor programs (rows and columns, respectively, in Table 7). When the donor program

400 uses the same epithet *and* taxonomic concept as does the receiver, the outcome is coded in the

401 matrix as a Perfect Match (PM). Where programs do not have an exact match, the receiver's

402 request results in one of the following outcomes:

403   • No match (Zero). The taxonomic unit does not occur at the donor program. This means that

404   receivers can assume that abundances are always zero for that species throughout the donor

405   program's range.

406   • Compatible match (CM). Compatible matches connect observations that represent the same

407   taxonomic concept as specified by the receiver, but symbolized by the donor with a different

408   nomenclature. This could be due to the use of a different genus or species epithet, the

409   adoption of alternate spelling, or the differential recognition of species/subspecies taxonomic

410   levels.

411   • Multiple match (MM). When there are differences in lumping and splitting taxonomic

412   entities, one program may pool together, under a single name, a concept that is separated into

413   two species by another program. If the program that lumps under one name requests data

414   from a program that splits the species into two, the tool will return both species, which we

415   designate here a "Multiple Match." In the reverse case, where a program that separates the

416   taxa into two names requests a match for one of the names from a program that lumps the

417   taxa together, we flag the response so that the user knows that donor data are only

418   combinable when data from both receiver taxa are included (and thus may not represent a

419   species-level entity from the receiver's perspective). These cases are indicated in Table 7 as

420   CM+.

421

422   Supplemental Table S4 lists the 39 species concept integration matrices for all other taxa (18%

423   of taxon entities) where there are taxonomic names and/or concepts mismatches between

424   programs.

425

426 **Objective 3. Development and testing of a novel integration tool.**

427

428 Using our alignment of program names in combination with integration matrices, we built a tool

429 for users to answer, on a case by case basis, whether it is appropriate to integrate data associated

430 with any particular taxon listed on any monitoring program and data gathered by any other

431 monitoring program. The tool is currently available through GitHub

432 (https://github.com/diatomsRcool/butterfly) and mybinder

433 (http://mybinder.org/repo/diatomsRcool/butterfly). This proof-of-concept tool is currently a

434 stand-alone tool, but one that we intend to make fully operable for automating data integration

435 within the PollardBase citizen science butterfly monitoring project infrastructure and can also be

436 implemented on other appropriate portals, such as butterfliesandmoths.org. Thus, the user

437 interface is, as yet, underdeveloped. The tool consists of two parts: 1) a file that contains the

438 mapping between each program list and their base list and from the base lists to a master

439 switchboard list (see Figure 2); and 2) an algorithm that takes user input and uses the mapping

440 file to return the requested results. The following text describes in detail the mapping and the

441 algorithm.

442

443 **The mapping.** We prepared a data file that links names in each program's checklist to the

444 corresponding name in the program's base list. This file also links corresponding base list names

445 across authorities in the master switchboard list. These mappings are codified using parent/child

446 relationships conforming to the Darwin Core biodiversity standard (Wieczorek et al., 2012). This

447 initial linking was done manually as described in Steps 2 and 3 in Objective 1 above. Figure 2

448 shows two examples for how these links connect names between two program lists through their

449   associated base lists. Figure 2A shows an example of how we adjusted lists in the mapping file

450   (lower panel, at head of arrow shows we made duplicate *C. ladon* entries), so that they are not

451   exactly true to the original lists (upper panel). This duplication is necessary to account for

452   multiple taxon concept interpretations. Figure 2B shows the mapping of taxa between program

453   lists that have different base lists, using a master switchboard list to 1) disambiguate and manage

454   the multiple concept interpretations of the same name between base lists and 2) allow us to

455   protect the granularity of a program list that may use a less specific base list.

456

457   As is standard for Darwin Core representations of biodiversity classifications, every name on

458   each program and base list included in our mapping file is assigned a unique Taxon ID, Parent

459   ID, and a Source (Wieczorek et al., 2012). Each name in a program list has a parent in the base

460   list. Each name in the base list has a parent in the switchboard. The Taxon ID and the Parent ID

461   are alphanumeric strings. The Source will always be either a program- or a base-list reference. It

462   is this chain of Taxon ID and Parent ID that the algorithm uses to travel from a Start Program to

463   the switchboard and back down to the End Program.

464

465   **The algorithm.** The algorithm takes three pieces of information as input from the user: the

466   recipient program, the taxon of interest, and the donor program. It then uses the mapping file to

467   return the requested result – the equivalent taxon of interest from the donor program. The

468   algorithm searches for the taxon of interest in the recipient program using string matching and

469   then uses the parent identifier to move up the parent/child hierarchy to the switchboard (e.g. Fig.

470   3A). After reaching the switchboard, the algorithm starts traveling back down the parent/child

471   hierarchy by searching for taxa with a specified parent identifier until it finds the appropriate

472   taxon or taxa in the donor program (Fig. 3B). If no compatible taxa can be found, the system

473   returns: "There is no match in this list (species presumed absent)." This is the case for requests

474   between programs that show "Zero" as the outcome in the concept integration matrix.

475

476   **Tool outcomes.** In the cases where the matching is too complicated to capture in an algorithm,

477   the tool returns a flag along with the outcome to explain the confusion for that particular pairwise

478   comparison or to suggest that the outcome requires the user to make an individual determination

479   on whether the two taxon concepts are indeed compatible before combining data. For technical

480   reasons, the tool must give symmetrical returns for both directions (i.e. when donor and receiver

481   roles are exchanged) in a pairwise comparisons of programs. In cases where the species matches

482   are not symmetrical in both directions (gray cells in Table 7), the tool returns a flag.

483   Supplemental Article S2 details these two types of flagged outcomes. Note that the tool flags

484   results of difficulties in cases where there is substantial disagreement, even though in our

485   equivalence matrix the comparison might be designated as Compatible or Multiple Match. The

486   *C. ladon* complex is complicated by a broad diversity of concept interpretations among base- and

487   program-lists, and every pairwise comparison except two (requests between OC and MPG)

488   results in a flag accompanying the taxon outcome (Table 7).

489

490   **Testing the Tool.** We carried out comprehensive testing on pairwise comparisons across all taxa

491   in all programs, shown in Supplemental Table S5. Each comparison was listed with its expected

492   results. Tests were submitted either in bulk on the command line or singly in a Jupyter notebook.

493   A test was successful if the tool gave a result that matched what was expected based on our pre-

494   constructed compatibility matrices. Refinement of the tool occurred iteratively until 100% of

495   comparison tests were successful.

496

497   **Discussion**

498   We have developed a system for the curation of names and associated taxonomic concepts for

499   programs that actively collect data over long time periods. This allows data to be easily

500   integrated between any monitoring programs that adopt this system, even when they use different

501   or even unique nomenclatures. While there are other available systems to resolve taxonomic

502   name conflicts of existing data, they are all retrospective and often rely on assumptions about the

503   collector of the data, who may or may not be available to articulate their concepts when they

504   affixed a name to a data point. Our system is proactive in that it coordinates programs that are

505   continually collecting data by registering their own specified taxonomic names and concepts in

506   relationship to a standardized checklist. By ensuring that PollardBase program directors

507   themselves retain the control of associating taxonomic names between their check list and base

508   lists, we eliminated any disinclination or barriers programs might have in signing on to use our

509   system, and in fact rather than resenting our intrusion, all programs appreciated our help in

510   curating their species lists.

511

512   Combined with PollardBase, which prevents the accidental creation of incorrect name text

513   strings, this system creates a novel taxonomic curation and integration system that maps all,

514   including legacy, concepts between programs while making no demands on the involved

515   monitoring programs and requiring little taxonomic knowledge on the part of the end-user. Our

516   testing shows that our tool can be used to implement our curation system. Using 10 North

517   American monitoring program checklists and three taxonomic authority lists, our tool accurately

518   recovered comparable taxonomic units between monitoring programs when programs identify

519   taxa at the same taxonomic resolution. When species are identified at different taxonomic levels,

520   our tool flags the data as non-combinable and identify where there may be problems of spurious

521   zeros. Developing a taxonomic backbone that could ensure 100% correct integration under every

522   situation is extraordinarily complex. As currently designed our tool does not provide a solution

523   for mismatches, but instead leaves it up to the individual researcher to resolve them. Other

524   systems have been designed to deal with this problem (see Ellingsen et al. 2017 for one example)

525   and there is room for improving specific integration guidance for these more complex situations

526   in future iterations of this tool.

527

528   A key advantage of our system is its scalability. Adding a new monitoring program only requires

529   the program to declare one of the taxonomic authorities as its base, and to register any deviations

530   from it. Once that is done, the tool automatically determines name compatibility between the

531   new and pre-existing programs as described above. Similarly, if a new authority list is published

532   that the community begins to use, we can add it as a new base list simply by encoding its

533   relationships via our switchboard. While the potential for much more complicated relationships

534   between concepts increases with additional programs, the structure of the mapping file makes

535   future interoperability tractable.

536

537   Currently in early stages of development, our tool is a beginning for implementing our curation

538   framework for associating combinable data. Important improvements would include enabling the

539   tool more generally and make it more easily operable by assimilating it with a more developed

540   interface into the PollardBase and other appropriate platforms; expanding its flexibility by

541   allowing users more customizable taxonomic interpretations; extending taxonomic

542   interoperability to include the perspective of independent third parties interested in the data from

543   these programs; and developing the means for automated updating of taxonomic concepts as they

544   arise in the literature. As case in point, we note the recent revision of the *Anthocharis* species

545   complex (Stout, 2018) is not reflected in any of our base authorities (although Pelham's most

546   recent (2019) version of his species catalog does include Stout's revision); we updated by

547   flagging the user to consider potential discrepancies.

548

549   Although developed for and tested on the PollardBase network of butterfly monitoring programs,

550   there is nothing about this system that is particular to butterflies. The utility of our system could

551   be implemented to help solve the significant stumbling block of inaccurate taxonomic integration

552   across structured monitoring programs focused on any taxonomic group for which: 1) a

553   taxonomic authority can be identified as the basis for each program's species checklist; 2) any

554   differences among these authorities can be mapped to build an infrastructure to translate between

555   programs; 3) programs can state any deviations between their checklist and their adopted

556   baselist. Regional structured surveying of taxa is becoming increasingly popular (Kelling et al.,

557   2019), and taxa such as odonates and amphibians are particularly amenable to surveys using

558   Pollard-like protocols. Indeed, Pollardbase began accepting odonate and amphibian survey

559   program participants in 2018 and 2019 respectively.

560

561   Citizen science monitoring observations are one of the only sources of large-scale

562   spatiotemporally-replicated biodiversity data available to answer our most important global

563 change questions. Increasingly complex and powerful models are available to integrate data

564 among regional monitoring programs but accurate results depend on consistent nomenclature to

565 prevent systemic mismatches. Our system is an innovative approach to identify and resolve

566 regionally employed taxon concepts by curating them alongside published authoritative

567 treatments and will guide researchers in tracking North American butterfly populations, while

568 preventing erroneous ecological conclusions drawn from incorrectly associated taxa.

569

570 ## Acknowledgements

571 The authors would like to thank the directors of each of the participating butterfly monitoring

572 programs. Doug Taron especially helped us to develop a system that would be tractable to

573 implement. Program directors at the time we implemented these programs include Doug Taron

574 (Illinois), Jerome Wiedmann (Ohio), Ashley Cole-Wick (Michigan), Sarah Garret (Colorado),

575 Jaret Daniels (Florida), Nathan Brockman (Iowa), Jutta Burger (Orange County), Jeff Pippen

576 (MPG Ranch) and Steve McGaffin (Tennessee). We thank Rob Guralnick and Jeff Pippen for

577 helpful feedback on the project and Jeff Pippen for helping to implement the curation system for

578 new users of PollardBase.

579

580 ## References

581 Amano, T., Lamming, J.D. and Sutherland, W.J., 2016. Spatial gaps in global biodiversity

582 information and the role of citizen science. *Bioscience*, *66*(5), pp.393-400.

583

584 Chesser, R.T., Burns, K.J., Cicero, C., Dunn, J.L., Kratter, A.W., Lovette, I.J., Rasmussen, P.C.,

585 Remsen, J.V., Rising, J.D., Stotz, D.F. and Winker, K., 2016. Fifty-seventh Supplement to the

586 American Ornithologists' Union Check-list of North American Birds. *The Auk*, *133*(3), pp.544-

587 561.

588

589   Boyle, B., Hopkins, N., Lu, Z., Garay, J.A.R., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L.,

590   Piel, W.H., Mckay, S.J. and Lowry, S., 2013. The taxonomic name resolution service: an online

591   tool for automated standardization of plant names. *BMC bioinformatics*, *14*(1), p.16.

592

593   Cassie, B., Glassberg, J., Swengel, A. and Tudor, G., 2001. North American Butterfly

594   Association (NABA) Checklist & English Names of North American Butterflies.

595   https://www.naba.org/pubs/enames2_3.html

596

597   Cayton, H.L., Haddad, N.M., Gross, K., Diamond, S.E. and Ries, L., 2015. Do growing degree

598   days predict phenology across butterfly species? *Ecology*, *96*(6), pp.1473-1479.

599

600   Cooper, C.B., Shirk, J. and Zuckerberg, B., 2014. The invisible prevalence of citizen science in

601   global research: migratory birds and climate change. *PloS one*, *9*(9), p.e106508.

602   https://doi.org/10.1371/journal.pone.0106508

603

604   Cui, H., Xu, D., Chong, S.S., Ramirez, M., Rodenhausen, T., Macklin, J.A., Ludäscher, B.,

605   Morris, R.A., Soto, E.M. and Koch, N.M., 2016. Introducing Explorer of Taxon Concepts with a

606   case study on spider measurement matrix building. *BMC bioinformatics*, *17*(1), p.471.

607

608   Devictor, V., Van Swaay, C., Brereton, T., Brotons, L., Chamberlain, D., Heliölä, J., Herrando,

609   S., Julliard, R., Kuussaari, M., Lindström, Å. and Reif, J., 2012. Differences in the climatic debts

610   of birds and butterflies at a continental scale. *Nature climate change*, *2*(2), p.121.

611

612   Diamond, S.E., Cayton, H., Wepprich, T., Jenkins, C.N., Dunn, R.R., Haddad, N.M. and Ries,

613   L., 2014. Unexpected phenological responses of butterflies to the interaction of urbanization and

614   geographic temperature. *Ecology*, *95*(9), pp.2613-2621.

615

616   Downes, C.M., Bart, J., Collins, B.T., Craig, B., Dale, B., Dunn, E.H., Francis, C.M., Woodley,

617   S. and Zorn, P., 2005. Small-scale monitoring-can it be integrated with large-scale programs? *In:*

618   *Ralph, C. John; Rich, Terrell D., editors 2005. Bird Conservation Implementation and*

619   *Integration in the Americas: Proceedings of the Third International Partners in Flight*

620   *Conference. 2002 March 20-24; Asilomar, California, Volume 2 Gen. Tech. Rep. PSW-GTR-191.*

621   *Albany, CA: US Dept. of Agriculture, Forest Service, Pacific Southwest Research Station: p.*

622   *993-996* (Vol. 191).

623   https://www.fs.fed.us/psw/publications/documents/psw_gtr191/psw_gtr191_0993-

624   0996_downes.pdf.

625

626   Ellingsen, K.E., Yoccoz, N.G., Tveraa, T., Hewitt, J.E. and Thrush, S.F. 2017. Long-term

627   environmental monitoring for assessment of change: measurement inconsistencies over time and

628   potential solutions. *Environmental monitoring and assessment*, *189*(11), pp.595.

629

630   Forister, M.L. and Shapiro, A.M., 2003. Climatic trends and advancing spring flight of

631   butterflies in lowland California. *Global change biology*, *9*(7), pp.1130-1135.

632

633   Forister, M.L., McCall, A.C., Sanders, N.J., Fordyce, J.A., Thorne, J.H., O'Brien, J., Waetjen,

634   D.P. and Shapiro, A.M., 2010. Compounded effects of climate change and habitat alteration shift

635   patterns of butterfly diversity. *Proceedings of the National Academy of Sciences*, *107*(5),

636   pp.2088-2092.

637

638   Franz, N.M., Chen, M., Kianmajd, P., Yu, S., Bowers, S., Weakley, A.S. and Ludäscher, B.,

639   2016. Names are not good enough: reasoning over taxonomic change in the Andropogon

640   complex 1. *Semantic Web*, *7*(6), pp.645-667. https://content.iospress.com/articles/semantic-

641   web/sw220

642

643   Franz, N.M., Peet, R.K. and Weakley, A.S., 2008. On the use of taxonomic concepts in support

644   of biodiversity research and taxonomy, in: The New Taxonomy, Q.D. Wheeler, ed., *Systematics*

645   *Association Special Volume*, *76*: Taylor & Francis, Boca Raton, (2008):63–86.

646   doi:10.1201/9781420008562.ch5

647

648   Franz, N.M. and Peet, R.K., 2009. Towards a language for mapping relationships among

649   taxonomic concepts. *Systematics and Biodiversity*, *7*(1), pp.5-20.

650

651 Giangrande, A., 2003. Biodiversity, conservation, and the 'Taxonomic impediment'. *Aquatic*

652 *Conservation: Marine and Freshwater Ecosystems*, *13*(5), pp.451-459.

653

654 Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A.T., 2004. New developments

655 in museum-based informatics and applications in biodiversity analysis. *Trends in ecology &*

656 *evolution*, *19*(9), pp.497-503.

657

658 Heffernan, J.B., Soranno, P.A., Angilletta Jr, M.J., Buckley, L.B., Gruner, D.S., Keitt, T.H.,

659 Kellner, J.R., Kominoski, J.S., Rocha, A.V., Xiao, J. and Harms, T.K., 2014. Macrosystems

660 ecology: understanding ecological patterns and processes at continental scales. *Frontiers in*

661 *Ecology and the Environment*, *12*(1), pp.5-14. doi:10.1890/130017.

662

663 Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M.,

664 Hochachka, W.M., Julliard, R., Kraemer, R. and Guralnick, R., 2019. Using Semistructured

665 Surveys to Improve Citizen Science Data for Monitoring Biodiversity. *BioScience*, *69*(3),

666 pp.170-179.

667

668 Kennedy, J.B., Kukla, R. and Paterson, T., 2005, July. Scientific names are ambiguous as

669 identifiers for biological taxa: their context and definition are required for accurate data

670 integration. In *International Workshop on Data Integration in the Life Sciences* (pp. 80-95).

671 Springer, Berlin, Heidelberg.

672

673 Kéry, M., 2018. Identifiability in N-mixture models: A large-scale screening test with bird data.

674 *Ecology*, *99*(2), pp.281-288.

675

676 Lepage, D., Vaidya, G. and Guralnick, R., 2014. Avibase–a database system for managing and

677 organizing taxonomic concepts. *ZooKeys*, (420), p.117.

678

679 Lepage, D. 2017. Avibase – the world bird database. Retrieved from https://avibase.bsc-

680 eoc.org/avibase.jsp?lang=EN&pg=home

681

682 Levy, O., Ball, B.A., Bond-Lamberty, B., Cheruvelil, K.S., Finley, A.O., Lottig, N.R.,

683 Punyasena, S.W., Xiao, J., Zhou, J., Buckley, L.B., Filstrup, C.T., Keitt, T.H., Kellner, J.R.,

684 Knapp, A.K., Richardson, A.D., Tcheng, D., Toomey, M., Vargas, R., Voordeckers, J.W.,

685 Wagner, T. and Williams, J.W., 2014. Approaches to advance scientific understanding of

686 macrosystems ecology. *Frontiers in Ecology and the Environment*, *12*(1), pp.15-23.

687 doi:10.1890/130019.

688

689 Mills, S.C., Oliver, T.H., Bradbury, R.B., Gregory, R.D., Brereton, T., Kühn, E., Kuussaari, M.,

690 Musche, M., Roy, D.B., Schmucki, R. and Stefanescu, C., 2017. European butterfly populations

691 vary in sensitivity to weather across their geographical ranges. *Global ecology and*

692 *biogeography*, *26*(12), pp.1374-1385.

693

694 Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G. and Worm, B., 2011. How many species are

695 there on Earth and in the ocean? *PLoS biology*, *9*(8), p.e1001127.

696

697 NABA Board of directors, Fall/Winter 2006. NABA Butterfly count column 2. American

698 Butterflies, p. 56-57. Available at

699 http://www.naba.org/pubs/ab143_144/ab143_4count_column_program_changes.pdf

700

701 NABA Names Committee. 2015. Interim Report of the NABA Names Committee American

702 Butterflies 22(4): 26-39.

703 https://www.naba.org/ftp/2015a_Interim_Report_NABA_Names_Committee.pdf.

704

705 NABA Names Committee. 2016. Second Interim Report of the NABA Names Committee

706 American Butterflies 23(3/4): 26-45.

707 http://www.naba.org/ftp/2016_Interim_Report_NABA_Names_Committee.pdf

708

709 North American Butterfly Association, Inc. 2001-2016. Checklist of North American Butterflies

710 Occurring North of Mexico – Edition 2.3 http://www.naba.org/pubs/enames2_3.html.

711

712   Opler, P.A. and A.D. Warren, 2003. Lepidoptera of North America. 4. Scientific Names List for

713   Butterfly Species of North America, north of Mexico. Contributions of the C. P. Gillette

714   Museum of Arthropod Diversity, Colorado State University. 79pp.

715

716   Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L. and Remsen, D.P., 2010. Names are key to the

717   big new biology. *Trends in ecology & evolution*, *25*(12), pp.686-691.

718

719   Patterson, D., Mozzherin, D., Shorthouse, D.P. and Thessen, A., 2016. Challenges with using

720   names to link digital biodiversity information. *Biodiversity Data Journal*, (4):e8080. doi:

721   10.3897/BDJ.4.e8080.

722

723   Pautasso, M. and Weisberg, P.J., 2008. Negative density–area relationships: the importance of

724   the zeros. *Global Ecology and Biogeography*, *17*(2), pp.203-210.

725

726   Wright, D.M. and Pavulaan, H., 1999. Celastrina idella (Lycaenidae: Polyommatinae): a new

727   butterfly species from the Atlantic coastal plain. *The Taxonomic Report*, *1*(9), pp.1-11.

728

729   Pelham, J.P. 2008. A catalogue of the butterflies of the United States and Canada with a

730   complete bibliography of the descriptive and systematic literature. Journal of Research on the

731   Lepidoptera, vol. 40. xiv + 658. Revised 22 June 2014 and 18 April 2019.

732

733   Pocock, M.J., Roy, H.E., Preston, C.D. and Roy, D.B., 2015. The Biological Records Centre: a

734   pioneer of citizen science. Biological Journal of the Linnean Society, 115(3), pp.475-493).

735

736   Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L. and Sauer, J.R., 2002.

737   Large scale wildlife monitoring studies: statistical methods for design and analysis.

738   *Environmetrics: The official journal of the International Environmetrics Society*, *13*(2), pp.105-

739   119.

740

741   Pyle, R.L., 2004. Taxonomer: a relational data model for managing information relevant to

742   taxonomic research. *PhyloInformatics*, *1*, pp.1-54.

743

744   Pyle, P. and DeSante, D.F., 2006. Updates to four-letter and six-letter alpha codes based on

745   Revisions by the American Ornithologists' Union. *North American Bird Bander*, *34(3)*, pp.109-

746   110.

747   http://www.birdpop.org/docs/pubs/Pyle_and_DeSante_2009b_Updates_to_Four_and_Six_Letter

748   _Alpha_Codes.pdf

749

750   Pyle, R.L., 2016. Towards a Global Names Architecture: The future of indexing scientific

751   names. *ZooKeys*, (550), p.261.

752

753   Remsen, D., 2016. The use and limits of scientific names in biological informatics. *ZooKeys*,

754   (550), p.207.

755

756   Ries, L. and Oberhauser, K., 2015. A citizen army for science: quantifying the contributions of

757   citizen scientists to our understanding of monarch butterfly biology. *BioScience*, *65*(4), pp.419-

758   430. https://doi.org/10.1093/biosci/biv011.

759

760   Royle, J.A., Nichols, J.D. and Kéry, M., 2005. Modelling occurrence and abundance of species

761   when detection is imperfect. *Oikos*, *110*(2), pp.353-359.

762

763   Schmidt, B.C. and Layberry, R.A., 2016. What Azure blues occur in Canada? A re-assessment of

764   Celastrina Tutt species (Lepidoptera, Lycaenidae). *ZooKeys*, (584), p.135-164. doi:

765   10.3897/zookeys.584.7882.

766

767   Schmucki, R., Pe'er, G., Roy, D.B., Stefanescu, C., Van Swaay, C.A., Oliver, T.H., Kuussaari,

768   M., Van Strien, A.J., Ries, L., Settele, J., Musche, M., Carnicer, J., Schweiger, O., Brereton,

769   T.M., Harpke, A., Heliola, J., Kuhn, E., and Julliard, R. 2016. A regionally informed abundance

770   index for supporting integrative analyses across butterfly monitoring schemes. British Ecological

771   Society Special Feature: Demography Beyond the Population. *Journal of Applied Ecology*,

772   *53*(2), pp.501-510. doi: 10.1111/1365-2664.12561.

773

774   Schwanghart, W., Beck, J. and Kuhn, N., 2008. Measuring population densities in a

775   heterogeneous world. *Global Ecology and Biogeography*, *17*(4), pp.566-568.

776   doi:10.1111/j.1466-8238.2008.00390.x

777

778   Scott, J.A. 1986. Butterflies of North America: A Natural History and Field Guide. Stanford

779   University Press. 583 pp.

780

781   Stout, T.L. 2018. A review of three species-level taxa of the Anthocharis sara complex

782   (Lepidoptera: Pieridae: Pierinae: Anthocharidini). *Insecta Mundi.* 1133.

783   https://digitalcommons.unl.edu/insectamundi/1133

784

785   Taron, D. and Ries, L., 2015. Butterfly monitoring for Conservation. In *Butterfly Conservation in*

786   *North America* (pp. 35-57). Springer, Dordrecht.

787

788   Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C. and Zipkin,

789   E.F., 2016. Joint dynamic species distribution models: a tool for community ordination and

790   spatio-temporal monitoring. *Global Ecology and Biogeography*, *25*(9), pp.1144-1158.

791

792   Van Swaay, C., Regan, E., Ling, M., Bozhinovska, E., Fernandez, M., Marini-Filho, O.J.,

793   Huertas, B., Phon, C.-K., Korösi, A., Meerman, J., Pe'er, G., Uehara-Prado, M., Sáfián, S., Sam,

794   L., Shuey, J., Taron, D., Terblanche, R., and Underhill, L. 2015. Guidelines for Standardised

795   Global Butterfly Monitoring. *Group on Earth Observations Biodiversity Observation Network,*

796   *Leipzig, Germany. GEO BON Technical Series,* 1, p.32.

797

798   Vaidya, G., Lepage, D., and Guralnick, R. 2018. The tempo and mode of the taxonomic

799   correction process: How taxonomists have corrected and recorrected North American bird

800   species over the last 127 years. *PLoS One* 13(4): e0195736.

801   https://doi.org/10.1371/journal.pone.0195736

802

803   Wepprich, T., Adrion, J., Ries, L., Wiedmann, J. and Haddad, N., 2019. Butterfly abundance

804   declines over 20 years of systematic monitoring in Ohio, USA. *BioRxiv*, p.613786.

805

806 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. and

807 Vieglais, D., 2012. Darwin Core: an evolving community-developed biodiversity data standard.

808 *PloS one*, *7*(1), p.e29715.

809

810 Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A.,

811 Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The

812 FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*.

813

814 Ytow, N., Morse, D.R. and Roberts, D.M., 2001. Nomencurator: a nomenclatural history model

815 to handle multiple taxonomic views. *Biological journal of the Linnean Society*, *73*(1), pp.81-98.

816

817 Zermoglio, P.F., Guralnick, R.P. and Wieczorek, J.R., 2016. A standardized reference data set

818 for vertebrate taxon name resolution. *PloS one*, *11*(1), p.e0146894.

819

820 Zipkin, E.F. and Saunders, S.P., 2018. Synthesizing multiple data types for biological

821 conservation using integrated population models. *Biological Conservation*, *217*, pp.240-250.

Numbers of taxa recognized by each base list and ITIS

Base List	Sub families	Tribes	Sub tribes	Genera	Sub genera	Species groups	Species	Sub species	Total
Pelham	23	46	37	247	38	36	820	1570	2817
NABA	24	0	0	221	0	0	725	112	1082
OW	22	2	0	231	0	0	784	2	1041
ITIS	25	44	3	235	0	0	804	1541	2652

Types of discrepancies observed in pairwise comparisons between authoritative base lists (NABA, O&W and Pelham) and between program checklists and their base authority.

Note: some comparisons showed more than one type of discrepancy.

Type of discrepancy (abbreviation)	Definition	Examples
Subfamily name deviation (F)	Lists use different subfamily names *Definition relevant only for comparisons between base authority lists.*	O&W don't recognize subfamily Polyommatinae; all taxa designated as Polyommaninae by NABA, Pelham and ITIS are considered Lycaenidae by O&W.
Genus name deviation (G)	Lists use different genus names (usually species epithet is the same). *Definition relevant for base list comparisons and base/project list comparisons.*	1) ITIS, Pelham, O&W base lists distinguish genus *Zerene* from genus *Colias*, while NABA does not. 2) Illinois list deviates from NABA base in using *Plebejus melissa* instead of *Lycaeides melissa*
Species epithet deviation (S)	Lists use different specific epithets and/or subspecies usage, due to: 1) differences in recognition of rank level for subspecies (subspecies promotion). 2) errors of determination. 3) different synonym is used. 4) differences in subspecies use - One list extends to subspecies, the other does not. Note: "Species" definition refers only to cases in which different species names are used purposefully; differences due to spelling are treated separately below. *Definition relevant for base list comparisons and base/project list comparisons.*	1) Subspecies promotion: NABA recognizes subspecies *Hesperia comma colorado*, while Pelham and O&W promote this to *H. colorado*. Note: this means a) different species names exist for the "*colorado*" taxon concept; and b) the term "*H. comma*" represents more than one concept. 2) Misdetermination: Hawaiian *Erionota torus* is listed as *E. thrax* by all base taxonomies except O&W, who note it was originally misdetermined as *E. thrax* by Linnaeus (1767). 3) Different synonyms: ITIS, Pelham, O&W use Piruna aea, NABA uses P. cingo. 4) Subspecies use deviations. In these cases, how/ whether names can be integrated across lists may depend on geography or concept: MPG lists *Speyeria cybele leto*, while all other monitoring groups list *S. cybele*.
Spelling/historical differences in scientific names (L)	Lists use alternate spellings. *Definition relevant for base list comparisons and base/project list comparisons. As mentioned in text, project list spellings are sanctioned, not introduced by individual monitors.*	NABA list spells *Atrytonopsis edwardsi* whereas Pelham and O&W spell *A. edwardsii*.
Species or subspecies conglomerates (C)	One project list combines two or more species or subspecies together (that are separated in base lists). *Definition relevant only for base/project comparisons.*	In every case encountered in our study, this divergence type occurred due to unreliability of distinguishing the taxa in the field. E.g. Illinois lumps *Colias eurytheme* and *C. philodice* (as *C. eurytheme/philodice*) although Illinois' base list (NABA) treats them as distinct species.
Unmatched Taxon (U)	Base or program list includes a species not on another base list, even under another name. *Definition relevant for base list comparisons and base/project list comparisons.*	This deviation usually happens when a list includes a taxon that rarely/never occurs in North America. E.g. Pelham includes *Kisutam syllis* (a rare stray to southern Texas).

Numbers of deviations by discrepancy type seen in pairwise comparisons of authority
lists (including ITIS)

We calculate percentage for each discrepancy type as a function of the maximum number of
species for each list pair (maximums shown in Table 1; e.g., for Pelham/NABA comparison,
species number is 820).

	Genus Deviation (G)	Species Deviation (S)	Spelling (L)	Unmatched Taxon (U)	Total deviations
Pelham/NABA	35 (4.3%)	104 (12.7%)	12 (1.5%)	42 (5.1%)	193 (23.5%)
Pelham/OW	17 (2.1%)	34 (4.1%)	3 (0.4%)	32 (3.9%)	86 (10.5%)
OW/NABA	26 (3.2%)	85 (10.8%)	11 (1.4%)	18 (2.3%)	140 (17.9%)
ITIS/Pelham	15 (1.8%)	12 (1.5%)	1 (0.1%)	20 (2.4%)	48 (5.9%)
ITIS/NABA	29 (3.6%)	94 (11.7%)	13 (1.6%)	25 (3.1%)	161 (20.0%)
ITIS/OW	2 (4.4%)	20 (2.5%)	2 (0.2%)	16 (2.0%)	40 (5.0%)

Deviations between each project list and its base list, by discrepancy type.

See Table 2 for definitions of discrepancy types.

Project	Project abbreviation	Base List	Total taxa	Total deviations	Deviations per # taxa (%)	Genus Name (G)	Species Name (S)	Spelling (L)	Unmatched Taxon (U)	Species
Cascades	CAS	Pelham	151	8	5.3	4	2	1	1	0
Colorado	CO	O&W	245	10	4.1	0	6	2	2	0
Florida	FL	O&W	192	2	1	0	0	0	2	0
Illinois	IL	NABA	142	5	3.5	1	0	2	0	2
Iowa	IA	NABA	118	2	1.7	2	0	0	0	0
Orange County (CA)	OC	O&W	79	6	7.6	0	2	0	4	0
Michigan	MI	NABA	135	3	2.2	0	1	1	0	1
MPG Ranch (MT)	MPG	Pelham	116	4	3.5	3	0	2	0	4
Ohio	OH	NABA	165	7	4.2	1	1	4	1	0
Tennessee	TN	NABA	135	9	6.7	5	0	2	0	2

Summary of discrepancies across monitoring programs

For more detail, see Supplemental Table S3. Note: issues here are those that pertain only to our 10 program lists. Authority list comparisons have taxonomic conflicts between taxa that are not sampled by these program lists.

Discrepancy types	Number of incidences	Number of names used	Number of concepts	Notes
1. Simple issues (names align 1:1 with taxon concept)				
Genus deviations: different genus names used	8	16	8	Combinable.
Species deviations: alternate synonym or spelling used	8	19	8	Combinable. For one taxon (*Phyciodes selenis*), programs used four name deviations that we consider (combinable) synonyms, however users should be aware that other researchers may recognize taxon concept distinctions among these terms
2. Subspecies Issues				
Some programs specify taxon to subspecies level (names align 1:1 with concepts at all comparisons)	10	20	20	Combinable at species- and/or subspecies-level comparisons when appropriate subspecies data is parsed at both programs. User should ensure level of analysis is clear.
Some programs list more than one subspecies (subspecies names do not align 1:1 with species-level concept)	2	6	6	Combinable at species- and/or appropriate subspecies-level comparisons. Both subspecies must be included in species-level comparisons with other programs. User combine with care.
3. Taxonomic issues: identification, alternate species names, taxonomic revisions, subspecies promotion interpretations (Names do not align 1:1 with taxonomic concept)				
Conglomerates (species level ID difficult to distinguish in field so some programs collect combined data)	2	6	6	For some program combinations, data must be analyzed at levels higher than species concept. A third taxon conglomerate occurs in *C. ladon* species complex; numbers of names and concepts for this conglomerate are included in complexes category below.
Complexes (different taxon concept interpretations)	7	19	18	Includes issues of subspecies promotion, taxonomic revisions and different interpretations of species concepts. Tool interprets compatibility based on the taxonomic interpretation of program receiving data and geographic range when appropriate.
Celastrina ladon species complex	1	7	8	Within the *C. ladon* species complex some programs also conglomerate taxon collection for *ladon* and *neglecta*

The taxonomic entities from the *C. ladon* complex occurring in each program range, and nomenclature used by each program.

(D) indicates that the program's interpretation deviates from that of its declared authoritative base. Dorsal (left wing) and ventral (right wing) images show wing variation among taxa (although much variation also exists within each taxon, a source of the taxonomic confusion). No dorsal image available for *humulus*. Photo credits: *C. ladon*, *C. neglecta*, *C. lucia*, Schmidt & Layberry, 2016; *C. echo*, CBG Photography Group, Centre for Biodiversity Genomics, CreativeCommons BY-NC-SA; *C. humulus*, Jeffrey Pippen.

		Taxonomic entity and nomenclature used by each program				
		ladon	*neglecta*	*lucia*	*echo*	*humulus*
Program	**Base authority**					
Illinois	NABA (D)	*C. ladon/neglecta*				
Iowa	NABA	*C. ladon*				
Ohio	NABA (D)	*C. ladon*	*C. neglecta*			
Michigan	NABA (D)	*C. ladon/neglecta*				
Tennessee	NABA (D)	*C. ladon*				
Florida	O&W	*C. ladon*	*C. neglecta*			
Cascades	Pelham			*C. lucia*	*C. echo*	
MPG Ranch, MT	Pelham				*C. echo*	
Colorado	O&W	*C. ladon*				*C. humulus*
Orange Co, CA	O&W (D)				*C. ladon echo*	

Data integration matrix for the *Celastrina ladon* complex showing matches between each pairwise combination of monitoring programs

In total, our ten monitoring programs use five taxonomic names (ladon, neglecta, echo, lucia, humulus) to represent multiple taxon concepts in the C. ladon complex. No program contains more than two taxa within its range. Each cell in the matrix indicates the outcome of a request for equivalent data from the "Receiver program" (R) in rows, to the "Donor program" (D), in columns. Where there is no match in the donor program, the cell is marked "Zero." Types of match (CM, PM, MM) are defined in text. A plus symbol (+) next to the match type indicates that the receiver program must include data from more than one of its surveyed taxa for compatibility with the donor's match, thus the receiver may consider the donor's content as not combinable at their species-level interpretation. A pair of programs may have compatibility in one direction but not in the other (that is, if donor and receiver roles are reversed). Gray cells indicate these situations where the match is not symmetrical in both directions of request. Comparisons between programs that share NABA as their base list (top left quadrant, boarded with thick lines) have less conflict than comparisons between programs whereby one program has a NABA base and the other program has Pelham or OW base list.

Figure 1

Map of North America showing locations of monitoring programs and *C. ladon* taxa.

Monitoring areas of the ten programs included in this paper are shaded in gray. Program abbreviations are given in Table 4. Approximate geographical distributions of the seven taxa included in the *C. ladon* complex (colored circles) are shown as described in the original base lists (more recent taxonomic updates may conflict with these distributions). Note that *serotina* and *idella* do not occur within the monitoring areas of any of these programs.

lucia

Cascades

MPG

humulus

echo

IA

MI

IL

OH

serotina

OC

CO

TN

ladon

neglecta

idella

FL

*Northern Indiana
is monitored by
both IL and MI
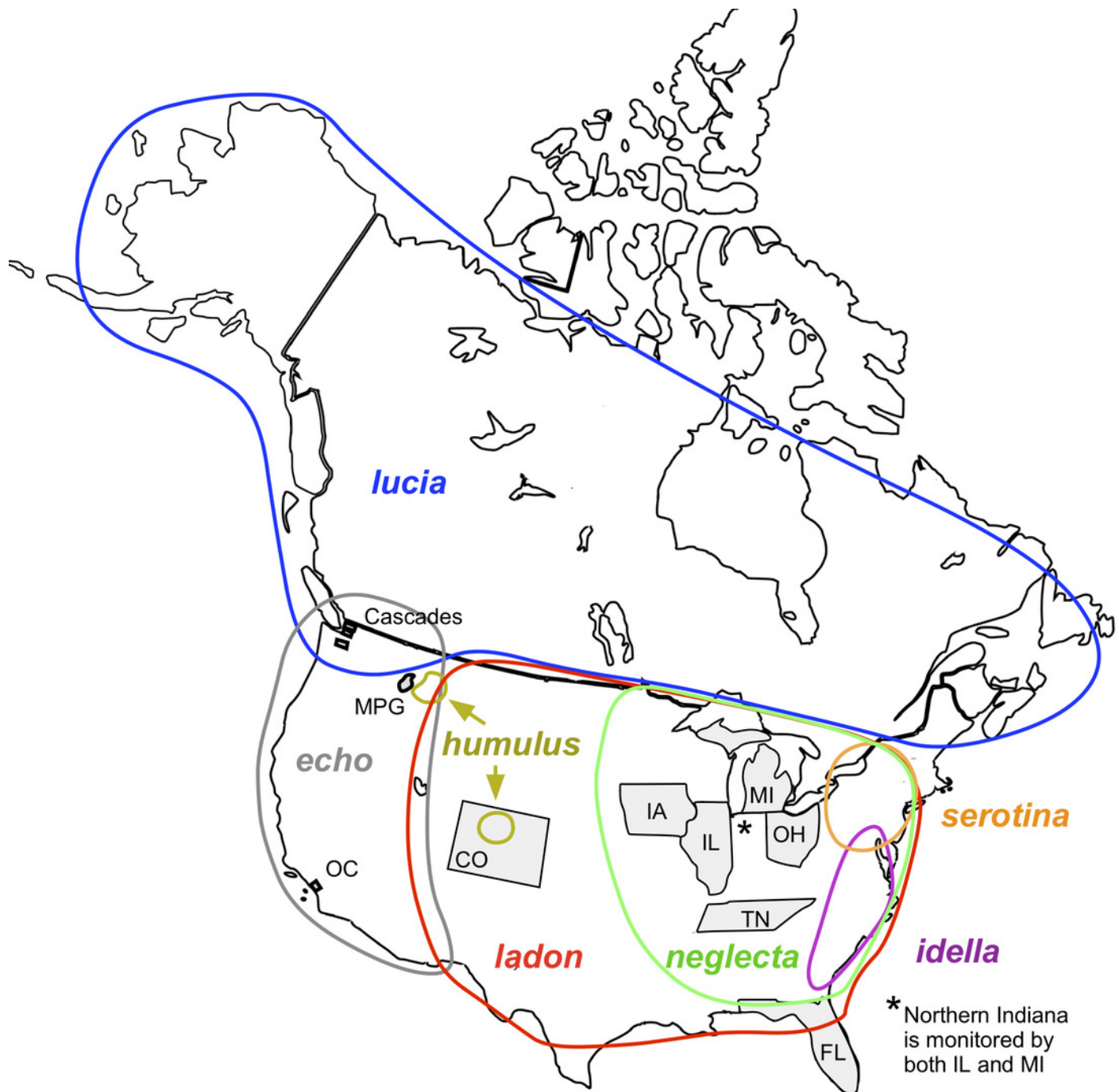
Figure 2

Examples of taxa translation across programs in the tool's name-mapping data file.

In some cases, a project list deviates from its base list such that multiple taxa on the project list match to a single species on the corresponding base. For example, Ohio's list includes *C. ladon* and *C. neglecta*. (A) Top panel illustrates that if we were to directly translate at the species level from Ohio's taxa to its (NABA) base equivalent, we would, in the process, eliminate Ohio's differentiation of *C. ladon* and *C. neglecta*, since NABA considers these subspecies of *C. ladon*. To retain granularity and manage relationships that are not differentiated by name string, we duplicate the name in the base list, allowing mapping of each project concept to a distinct base list holder (second panel, end of gray arrow). (B) When two programs have different base lists, we translate between them by introducing a "switchboard" that manages relationships and granularity between taxa in the base lists. Our algorithm (demonstrated in Fig. 5) uses this structure to find equivalent names between programs.

a.

b.
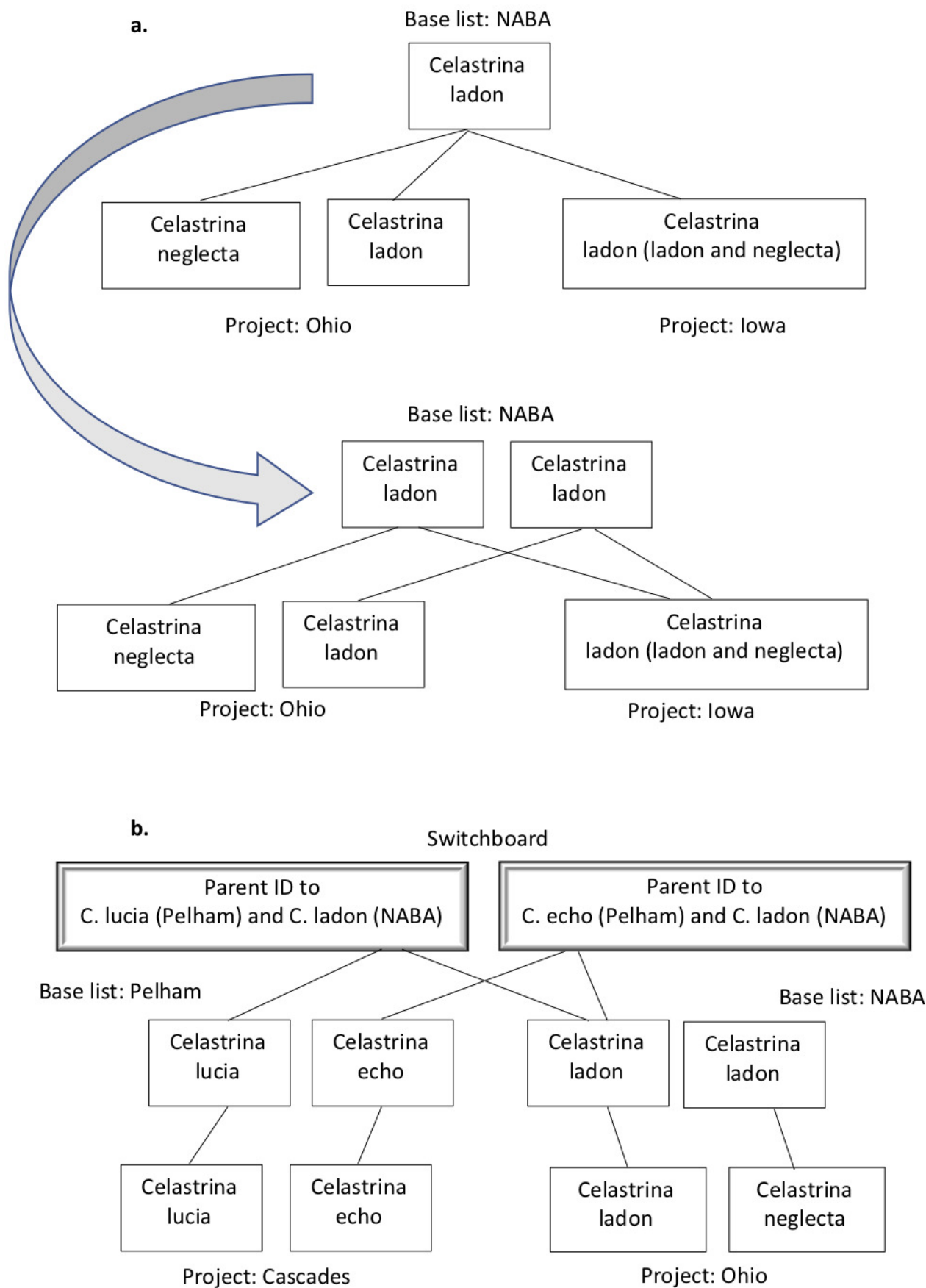
Figure 3

Tool algorithm for translating between two program lists through the switchboard - example query

This query uses the mapping file described in Figure 2b. The user has entered three pieces of information: a taxon name (*C. lucia*); the "Recipient program" (Cascades Butterfly Project); and the "Donor program" (Ohio Butterfly Monitoring Network). A. The algorithm first works its way up the hierarchy to the switchboard, starting with the input taxon name *C. lucia* (1) and "Recipient" program (Cascades Butterfly project) to find the appropriate data row in the mapping file. The parent identifier (2) for *C. lucia* at Cascades (T2936) is used to find the row for the corresponding taxon in the Pelham base list (3). The parent identifier for the base list (4; T1267) acts as the taxon ID at the switchboard (5, and dashed red line). B. The algorithm searches for rows that have the switchboard taxon identifier as the parent identifier (1'). These results are filtered, because the user identified a "Donor program," (Ohio) which has a defined base list (i.e., the algorithm knows to look only for a row where NABA is the source). Once the appropriate base list taxa have been found, its taxon identifier (2'; T1267) is used to find its child in the (Ohio) project list (3'; T3784). The name associated with the appropriate parent identifier (4' T6762) is returned as the result (5'; *Celastrina ladon*).