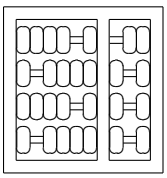

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Tradução taxonômica: o caso do SinBiota

Cleber Mira *Pedro Feijão* *João Meidanis*
Tiago Duque-Estrada *Carlos Joly*

Technical Report - IC-13-09 - Relatório Técnico

March - 2013 - Março

Tradução taxonômica: o caso do SinBiota

Cleber Mira[1], Pedro Feijão[1,2], João Meidanis[1,2], Tiago Duque-Estrada[3,4] and Carlos Joly[3,4]

[1]Scylla Bioinformatics
[2]Institute of Computing, Unicamp
[3]BIOTA/FAPESP
[4]Institute of Biology, Unicamp

Abstract

O Sistema de Informação Ambiental (SinBiota 1.0), desenvolvido para integrar as informações sobre a biodiversidade do estado de São Paulo geradas pelo programa BIOTA/FAPESP, possui por volta de 18 mil nomes científicos adicionados por diversos pesquisadores ao longo de um período de mais de dez anos. Esses termos podem apresentar diversos tipos de problemas, como erros de digitação, ou a espécie pode ter mudado de nome oficial, tornando-se o termo um sinônimo do nome corrente.

Nesse trabalho apresentamos uma estratégia de resolução desses problemas para identificar os nomes científicos válidos no SinBiota, por meio de uma tradução dos termos atualmente cadastrados com base na classificação taxonômica suportada por uma iniciativa internacional.

1 Introdução

Dados taxonômicos e, mais especificamente, nomenclatura científica, é crucial para estudos de biodiversidade. Uma combinação de fatores, tais como o volume enorme de dados existente, a alta taxa de criação de novos nomes científicos, o processo demorado de validação e publicação de nomes certificados, causam a ocorrência de discrepâncias taxonômicas [7], como diferenças hierárquicas, erros de digitação e homônimos. Além disso, o próprio processo de atribuir um nome científico a um taxon pode incorrer na criação de sinônimos e no uso do mesmo nome para taxa diferentes. Dicrepâncias taxonômicas de um taxon em particular podem ser espalhadas em diversos banco de dados. Page [11] propõe um ambiente que combine LSIDs [10] e RDF [13] para

modelar nomes taxonômicos em bases de dados que sejam capazes de lidar com discrepâncias. Uma fonte comum de discrepâncias entre bases de dados taxonômicas são os *binômios*.

Um binômio (gênero + espécie) é utilizado por taxonomistas como um identificador único de uma espécie. Não há duas espécies com o mesmo binômio, mesmo em ramos distantes de uma árvore taxonômica. Portanto, dado um binômio, é possível obter a taxonomia completa de uma espécie usando uma base de dados.

Com o passar dos anos, o binômio de uma espécie pode sofrer atualizações. Por exemplo, o avanço de técnicas de classificação filogenética pode contribuir para encontrar classificações mais adequadas que influenciem no gênero ou espécie do binômio. Outra situação comum decorre da classificação de uma espécie por meio de binômios distintos elaborados por taxonomistas trabalhando de maneira independente. No caso de uma atualização de um binômio, o nome antigo passa a ser um sinônimo do binômio mais recente. Note que o binômio de uma espécie pode ser atualizado várias vezes. Dessa forma, podem existir vários sinônimos para o binômio corrente de uma espécie.

O Sistema de Informação Ambiental [14] (SinBiota 1.0) foi desenvolvido para integrar as informações geradas pelos pesquisadores associados ao programa científico BIOTA/FAPESP e para prover mecanismos de difusão da informação a respeito a biodiversidade do estado de São Paulo, no Brasil. Uma das informações cruciais armazenadas no sistema é classificação taxonômica das espécies coletadas ou observadas em coletas.

O Sinbiota possui uma lista de binômios, cadastrada manualmente ao longo de vários anos, de 1997 a 2009, aproximadamente. Estima-se que haja por volta de 18 mil binômios no SinBiota, adicionados por diversos pesquisadores durante este tempo. Para alguns destes nomes, não há informação completa. Por exemplo, há situações nas quais apenas o gênero é conhecido. Em outros casos, apenas a família ou outro taxon de nível superior na árvore taxonômica é conhecido. Ocorrem ainda no SinBiota binômios contendo erros de digitação ou abreviações em latim que indicam que a espécie não é exatamente aquela, mas é um parente próximo, por exemplo.

Para que os binômios cadastrados no SinBiota possam ser utilizados como chaves de identificação confiáveis de uma base de dados é necessário haver um meio de distinguir entre a lista de binômios válidos, ou seja, que são identificadores de espécies e a lista de binômios que apresentam algum dos problemas comentados anteriormente para que recebam um tratamento adequado à sua situação.

Apresentamos nesse trabalho uma estratégia de resolução de binômios para o caso do sistema SinBiota com o objetivo de obter a lista de binômios reais de sua base de dados, por meio da correção dos casos de sinonímia, erros de digitação, entre outros problemas. Essa estratégia consiste em uma tradução taxonômica de entradas de binômios no SinBiota em binômios ou sinônimos encontrados em bases de dados de classificação taxonômica suportadas por iniciativas internacionais.

Na Seção 2 descrevemos o problema de resolução de binômios do SinBiota. A Seção 3 apresenta a estratégia proposta para resolver os binômios do SinBiota por meio de um procedimento de tradução taxonômica. A Seção 4 mostra como os binômios considerados problemáticos segundo diversos critérios foram tratados para conseguirmos a sua resolução. Apresentamos os resultados da aplicação da estratégia de resolução de binômios na Seção 5. Sumarizamos esse trabalho na Seção 6.

2 Problema de resolução de Binômios

Nós chamamos de *nome corrigido* o nome de um binômio que passou por um processo de verificação e correção de erros, como por exemplo erros de digitação. Um *nome atual* de uma espécie é o valor aceito como binômio que identifica a espécie segundo uma classificação taxonômica confiável.

Dado um binômio cadastrado no SinBiota, o problema de resolução de binômios consiste em encontrar o nome atual que identifica a espécie a qual seria identificada pelo binômio original do SinBiota. Quando o binômio proveniente do SinBiota identifica uma certa espécie, então o seu estado de binômio é mantido, caso contrário, ele pode ser corrigido e atualizado para o estado de um sinônimo do nome atual da espécie. Mais especificamente, o problema de resolução de binômios pode ser subdividido em dois objetivos:

1. Identificar todos os binômios da lista original do SinBiota com binômios em bases de dados internacionais de confiança, corrigindo eventuais erros de digitação.

2. Dados os binômios corrigidos, obter o nome atual de cada espécie, caso o binômio obtido seja um sinônimo, e associá-los.

O atingimento desses objetivos exige a resolução de algumas questões. Um dos problemas que surgem no objetivo 1 é determinar o nível de similaridade entre nomes que aceitaremos para realizar uma correção. Determinar um nível de similaridade é necessário para que não ocorram falsos-positivos. Outro problema é determinar de quais locais é possível extrair a informação sobre nomes e sinônimos. Como essa é uma informação que está constantemente sendo atualizada, é preciso também determinar uma forma de atualizar a lista periodicamente, preferencialmente de um modo automático.

Uma tentativa em atender a esses objetivos já foi feita no passado, utilizando classificação taxonômica do Catalogue of Life, com bom índice de sucesso [9]. Porém, acreditamos que, utilizando as informações dos sinônimos e de bases de dados complementares, podemos obter resultados ainda melhores, chegando a quase 100% de cobertura.

3 Estratégia de tradução taxonômica

Uma maneira de solucionar o problema de identificar binômios poderia ser a realização de buscas por meio dos serviços web fornecidos pelas bases de dados de iniciativas internacionais. Se adotarmos essa estratégia, devemos definir um limite na realização de consultas individuais, por exemplo, estipulando um limite de uma consulta a cada dez segundos, para evitar a sobrecarga dos servidores que provém os serviços web. Tendo em vista que a consulta de todos os nomes que se encontram no SinBiota deve demorar várias horas (estimamos cerca de 50 horas para os 18 mil binômios), seria bom considerar testes iniciais com apenas uma pequena amostra dos dados.

A abordagem que adotamos para atacar o problema consiste em uma estratégia incremental: primeiro fazer uma busca exata pelos binômios originais do SinBiota, o que cobriria a grande maioria dos dados. Depois, buscar pelos binômios que restam, aumentando gradualmente a diferença aceita para erros tipográficos, e finalmente tratar com outras estratégias alguns poucos casos problemáticos que ainda não tiverem sido resolvidos.

3.1 Divisão dos nomes

Com a listagem de todos os binômios no SinBiota em mãos, um primeiro passo em nossa estratégia foi separá-los em dois arquivos: um arquivo chamado "binomios-busca.txt" somente com binômios completos (gênero + espécie) e outro com os binômios incompletos ou com algum tipo de anotação que analizaremos mais tarde, chamado "binomios-tratamento.txt".

A seguir, uma lista com os filtros usados para determinar quando um binômio pertence ao segundo arquivo:

- "sp." e variações, como "spp", "sp.2" ou "spTS1";
- Nomes com um ponto ("."), que quase sempre indicam alguma abreviatura como "sp.", "cf.", "aff.", "pr.", "gr.", "p.", "g.", "gen." ou "ciysp.";
- "xxx", "aaa";
- Nomes contendo qualquer caracter que não seja letras ou espaço, tais como pontos de interrogação, aspas, parênteses, traços, barras, sinais de sustenido, etc;
- Nomes com letras acentuadas, como "Polystira formosíssima", "Swartzia acutifólia" ou "Ormosia arbórea";
- Alguns casos em que o binômio era "Sem nome", seguido de um número;

- Outros casos específicos, como "Gêneros Espécies" e "Material não", que são descritos na Seção 4.0.1.

Um registro que não case com nenhum destes filtros deve ser um binômio completo, pronto para ser buscado na base de dados. É importante notar, entretanto, que estes binômios ainda podem ter erros de digitação, que devem ser levados em conta para a busca, ou mesmo com a codificação, que podem causar alguns erros durante as comparações. A Figura 1 mostra o processo de tratamento dos binômios.

O arquivo de entrada tem 12682 linhas/binômios. Em 01 de março de 2013, o arquivo binomios-busca.txt tem 9998 binômios, e o arquivo binomios-tratamento.txt, 2684 linhas.

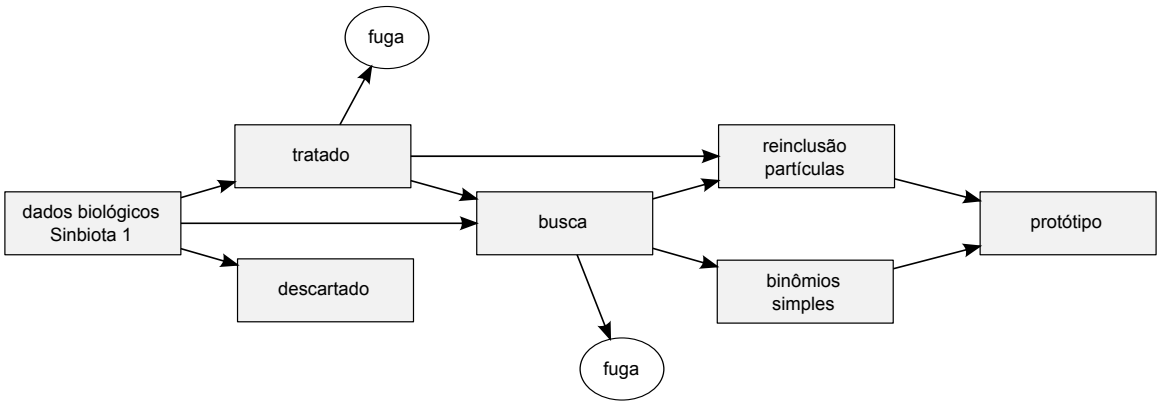

Figure 1: Tratamento dos dados biológicos do SinBiota 1.0

4 Resolução de binômios problemáticos

Esta seção dá maiores detalhes sobre o tratamento dos binômios com algum problema, que foram separados no arquivo "binomios-tratamento.txt" conforme as regras descritas anteriormente.

O arquivo "binomios-tratamento.txt" foi dividido em dois outros arquivos: "casos-tratados.txt" e "casos-ignorados.txt". No primeiro arquivo estão aqueles binômios que puderam ser tratados e passaram a integrar a lista de binômios prontos para busca. No segundo arquivo estão os binômios que não puderam ser aproveitados.

Os binômios com problemas que não foram descartados por um tratamento três etapas:

1. Tratamento automático das partículas em latim, conhecidas como *open nomenclature*;

2. Tratamento semi-automático de binômios com informações adicionais, como subespécie, nome do autor e características da coleta. Estas informações, em geral, introduziram acentos e caracteres especiais, como aspas, underscore e parênteses, dificultando a busca do binômio.

3. Tratamento manual de alguns casos não cobertos pelos casos anteriores, como nomes de espécies abreviados.

As etapas (1), (2) e (3) são descritas em detalhes nas Seções 4.0.2, 4.0.3 e 4.0.4, respectivamente.

Em 07 de março de 2013, o arquivo binomios-tratamento.txt tem 2684 linhas/binômios. O arquivo casos-tratados.txt tem 2639 linhas, e o arquivo casos-ignorados.txt tem 23 linhas. Boa parte dos binômios pode ser aproveitada. Ainda restam 22 binômios a serem avaliados.

4.0.1 Binômios descartados

Alguns binômios foram descartados devido a terem sido utilizados como dados de teste, não fazerem referência a nenhuma espécie, ou serem indeterminados. Os seguintes binômios não puderam ser aproveitados:

1. - -
2. Aaa bbb
3. Fdgdf gfg
4. Fgd ffd
5. Gen esp
6. Gen. sp.
7. Gen1 xxx
8. Gen.N. sp.3
9. Gen.N. sp.6
10. GêN.1 sp1
11. GêNeros Espécies
12. Indet indet
13. Indet Indeterminda

14. Indeterminada sp.
15. Material não
16. Sem nome1
17. Sem nome2
18. Sem nome3
19. Sem nome4
20. Sem nome5
21. Teste teste
22. X x
23. Xxx xxx

Estes binômios estão armazenados no arquivo "casos-ignorados.txt".

4.0.2 Tratamento das Partículas em Latim

Tratamos nomes com partículas em latim, também conhecidas como "open nomenclature", da seguinte forma: retiramos a partícula, buscamos o binômio resultante, e reintroduzimos a partícula. Partículas que passaram por este tratamento foram:

- sp.
- cf.
- aff.
- gr.
- ca.
- pr.

Estas partículas podem aparecer seguidas ou não de espaço, mas, na nossa reintrodução, sempre colocamos com espaço. Elas também podem aparecer tanto no nome do gênero quanto no nome da espécie. Por exemplo:

Aff.Miroculis xxx e Archaespora aff.leptoticha

Table 1: Tratamento das partículas em latim

Partícula	Variações	Exemplos
cf.	cf._, cf_, cf-, c.f., cf*espécie*	Adenomera cf.marmorata, Grubeulepis cf._tebblei, Bulbostylis cf_capilaris, Typhlomyrmex cf-major, Zetzellia c.f.mapuchina, Baccharis cfsemiserrata
aff.	(aff.*espécie*), aff._, aff_, aff*Gênero*	Aff.Miroculis xxx, Archaespora aff.leptoticha, Hylodes sp.(aff.lateristrigatus), Nassarius aff._albus, Laetacara aff_dorsigera, Affmiroculis xxx
gr.	(gr.*Gênero*), *Gênero*gr.*Gênero*, *Gênero*gr.*Espécie*, gr.*espécie*	Cyphomyrmex (gr.Strigatus)olitor, Cyphomyrmexgr.Strigatus strigatus, Cyphomyrmexgr.Rimosus sp.B1, Theridion gr.orgea
pr.	pr-	Rogeria pr.pellectasp.1, Adelomyrmex pr-boltoni, Typhlomyrmex pr-pusillus-sp.1
ca.		Yphthimoides ca.angulares
sp.	diversas (vide exemplos)	Aceria sp, Abacarus sp., Achaearanea sp.21, Acanthognathus ciysp.1, Acentroscelus sp.n., Adiantum spp, Anoplotermes spLCMO1, Apterostigma sp.rrs2, Arenicolidae spp., Attheyella sp.a, Aysha gr.brevimannasp.2, Cletodes sp.nov.6, Diopatra spp._(jovem), Octostruma sp.bhdL, Oxyepoecus boraceiensis.sp.nla1, Oxyepoecus nlasp.1, Phaenochitonia sp.(preta), Pseudomonas Pseudomonas_sp., Pyramica (Smithistruma)-sp.1, Pyrrhopyge sp.(rhacia?), Scutellospora sp1(natalina), Tacebia sp.(cf.africanus), Zischkaia sp.?

Além da ausência de espaço, estas partículas também apresentaram outras variações, como a substituição do espaço por um hífen ou um underscore. Listamos algumas destas variações na Tabela 1. Em todos estes casos, supusemos que foi um erro tipográfico, e que a intenção era colocar a partícula da forma descrita acima, abreviada com um ponto e seguida por espaço.

O tratamento de remover a partícula, realizar a busca, e reintroduzir a partícula foi aplicado nas diferentes partículas usadas, exceto em algumas ocorrências das partículas “gr.” e “sp.”.

Partícula gr. No caso da partícula "gr.", além da remoção da partícula, também foi necessário aplicar um tratamento para decidir se o conteúdo que acompanhava a partícula deveria ou não ser removido. Por exemplo, em

Bokermannohyla gr.circumdata

a espécie circumdata foi mantida. Já em

Cyphomyrmex (gr.Strigatus)strigatus

o gênero Strigatus foi removido, resultando no binômio "Cyphomyrmex strigatus".

Outras ocorrências, apesar de apresentaram o mesmo padrão, precisaram ser analisadas com maior cuidado. Por exemplo, a partir de

Cyphomyrmexgr.Rimosus sp.B1

foi obtido o binômio "Cyphomyrmex rimosus", já que rimosus é uma espécie válida para o gênero Cyphomyrmex. Pelo mesmo motivo, a partir de

Cyphomyrmexgr.Strigatus sp.B3

foi obtido o binômio "Cyphomyrmex strigatus". Entretanto, no caso

Cyphomyrmexgr.Strigatus strigatus

foi removido o termo que acompanhava a partícula "gr." pois, apesar de ser semelhante ao caso anterior, se mantivéssemos o termo Strigatus, o binômio resultante seria "Cyphomyrmex strigatus strigatus" que, assim como na ocorrência "Cyphomyrmex (gr.Strigatus)strigatus", não seria encontrado.

Partícula sp. Indicações de gênero sem espécie definida apareceram com diversos sufixos, incluindo

sp.1, sp.3, sp1, sp.n., sp, sp.act1, sp.act2, sp.B1, spAD1, sp.nov.1, sp.a, bhdsp.F, sp.aat1, sp.aat2, s.p.1

ou até mesmo, em certos casos, com as iniciais do autor da coleta:

ciysp.1, sp.rrs1, spLCMO1

Todos estes sufixos foram substituidos por sp. somente. Também supusemos que spp é o mesmo que sp [5].

Dentre as partículas usadas, "sp." e suas variações foram as que mais ocorreram nos binômios a serem tratados. Aproximadamente 54% dos binômios para tratamento continham alguma variação desta partícula.

Outras partículas Houve casos em que outras partículas foram utilizadas como, por exemplo, a partícula "xxx":

Aff.Miroculis xxx ou Affmiroculis xxx

Nestes casos, removemos a partícula aff, buscamos Miroculis, e reintroduzimos a partícula aff. Desprezamos o xxx, trocando-o por sp. O mesmo se aplica às ocorrências com o termo "indeterminada", como "Difflugia indeterminada". Às vezes o sufixo "indeterminada" pode estar escrito incorretamente, como em "Picramnia Indeterminda".

Há termos também onde aparece uma interrogação ou um asterisco num dos nomes, por exemplo:

Agistemus floridanus?
?Dicrothrix sp.
Eugenia cuprea(?)
Sphiggurus insidiosus*
Stryphnodendron* adstringens
Tribelos (?)
Xispia sp.?

Estes foram tratados exatamente como as partículas cf., aff., etc. Por exemplo, o nome "Tribelos (?)" foi buscado como "Tribelos" e exibido ao usuário como "Tribelos ?".

Em alguns registros encontrados no sistema, há somente um ponto no lugar do segundo nome ("Buchenavia ."). Nestes casos, também consideramos como sp. Em outros, um dos nomes é seguido por um ponto, apesar de estar completo (por exemplo, "Alchornea. triplinervia" e "Vernonia virgulata."). Nestes casos, desprezamos o ponto.

4.0.3 Tratamento de Informações Adicionais

Além do tratamento das partículas, foi preciso identificar e tratar informações adicionais contidas nos dados. Nestes casos, temos que corrigir a entrada para que contenha somente gênero e espécie. Por exemplo, alguns binômios repetiam o gênero, abreviado ou não, no segundo nome:

Enterobacter Enterobacter_aerogenes, Cerdocyon C.thous

há casos ainda em que a entrada inclui toda a referência para a espécie, como em

Cymbella Cymbella_naviculiformis_Auerswald_ex_Heiberg

Também existem entradas que incluem subespécies ou variedades, como

Cymatium parthenopeum_parthenopeum, Mimosa dolens_var.acerba

para estas entradas, desconsideramos a subespécie ou variedade e fazemos a busca até o nível de espécie.

Existem binômios com um subgênero, entre parênteses, como em

Amphiura_(Ophionema) intricata

Para estes casos, executamos a busca sem o subgênero.

Outros comentários entre parênteses, como (jovem), (branco), (nova), (laranja), (preta), também foram desprezados.

O único caso em que o conteúdo entre parênteses não foi desprezado foi o Mesene (pyrippe), onde (pyrippe) é o nome da espécie.

Aspas em gêneros ou espécies também são ignoradas para a busca.

Informações adicionais com underscore Diferente do tratamento das partículas, a ausência de um padrão na descrição das informações adicionais dificultou o tratamento, que muitas vezes teve que ser avaliado caso a caso. Como exemplo, mencionamos o caso em que o primeiro ou segundo nome continham vários termos concatenados por underscore, onde não foi possível identificar um padrão para reconhecer se algum dos nomes deveria ou não ser removido. A Tabela 2 lista alguns destes casos. Ao total, 150 binômios da lista de tratamento continham algum underscore.

Outras informações adicionais Além das entradas contendo underscore, alguns binômios pontuais foram identificados e tratados manualmente. Por exemplo, no caso da entrada "Coussarea meridionalisvarporophylla", o segundo nome contém a espécie e a variedade, sem nenhum caracter separador. A única indicação é o termo var no meio do nome meridionalisvarporophylla. A variedade, como nos casos relatados anteriormente, foi removida.

Outros casos específicos foram os termos "Pseudopaludicola aff.falcipesI" e "Pseudopaludicola aff.falcipesII", onde foram usadas letras para indicar números. A busca, em ambos casos, foi realizada com o binômio "Pseudopaludicola falcipes".

No caso da entrada "C.(Tanaemyrmex) balzani", o gênero Camponotus foi abreviado, e o subgênero aparece entre parênteses. O binômio, após o tratamento, passou a ser "Camponotus balzani".

4.0.4 Tratamento com consulta ao SinBiota

Em alguns casos não foi possível tratar o binômio usando somente as informações da entrada. Para esta etapa do tratamento, foi preciso recorrer à outras informações fornecidas pela ferramenta de busca de coletas do protótipo do sistema SinBiota 2.0 [12].

Table 2: Tratamento de nomes com underscore

Entrada	Tratamento	Saída
Amphiura_(Ophionema) intricata	O primeiro nome indica o gênero e o subgênero. Somente o gênero é mantido.	Amphiura intricata
Astyanax scabrinnis_paranae	O segundo nome contém a espécie e a subespécie. Somente a espécie deve ser mantida	Astyanax scabrinnis
Enterobacter Enterobacter_aerogenes	O segundo nome repete o gênero, que é removido.	Enterobacter aerogenes
Euchlanis incisa_incisa	O segundo nome repete a espécie duas vezes, e uma delas é removida.	Euchlanis incisa
Eunotia Eunotia_arcus_Erenberg	O segundo nome contém o gênero, a espécie e o nome do autor. Somente a espécie deve ser mantida	Eunotia arcus
Clypeaster_(Stolonoclypus) subdepressus_subdepressus	Ambos os nomes contém underscore. No primeiro, o subgênero é desprezado; no segundo, a espécie é repetida, e uma das repetições foi removida.	Clypeaster subdepressus
Onuphis eremita_oculata	O segundo nome contém espécie e subespécie. Somente a espécie é mantida.	Onuphis eremita
Onuphis_Eremita oculata	O primeiro nome contém o gênero e a espécie. Ambos são mantidos, e a subespécie (segundo nome) não é utilizada na busca	Onuphis eremita

Algumas entradas continham o gênero escrito corretamente, mas o nome da espécie estava abreviado. Por exemplo, em "Automeris b." não é possível inferir qual o nome da espécie abreviada, já que no gênero Automeris existem muitas espécies cujo nome se iniciam com a letra B:

- Automeris balachowskyi
- Automeris banus
- Automeris basalis

- Automeris beckeri
- Automeris belti
- Automeris beutelspacheri
- Automeris bilinea
- Automeris boops
- Automeris boucardi
- Automeris boudinoti
- Automeris boudinotiana

Para tratarmos esta entrada, através do sistema SinBiota pesquisamos todas as coletas que haviam cadastrado o termo "Automeris b.". Neste exemplo, encontramos duas coletas, com códigos 4868 e 5312, ambas do autor Vitor O. Becker e cadastradas por Francini Osses. Ao consultar as outras informações da coleta, verificamos que o termo "bilinea" havia sido cadastrado no campo subespécie, em ambas coletas. Com esta informação adicional, inferimos que o binômio tratado seria "Automeris bilinea". Repetimos o tratamento para os casos similares, que ocorreram somente nos registros cadastrados por Francini Osses.

4.0.5 Casos ainda não tratados

Ainda existem três situações em que ainda não sabemos qual seria um tratamento adequado. Todos os binômios que ainda não foram tratados se enquadram em algumas destas situações, conforme apresentado na Tabela 3. Todos os binômios não tratados estão armazenados no arquivo "casos-nao-tratados.txt".

4.1 Base de dados

Foi feito o download da base de dados completa do ITIS (Integrated Taxonomic Information System) [1], do endereço `http://www.itis.gov/downloads/index.html`, escolhendo a opção "Full ITIS Data Set (MySQL bulk load)". Três tabelas são relevantes para nossos propósitos:

longnames liga o nome completo (campo completename) ao identificador (TSN)

hierarchy liga uma unidade taxonomica a seu pai (por exemplo, espécie a genus): campos TSN e parent_TSN

synonym_links liga os sinônimos ao nome mais recente: campos TSN (sinônimo) e TSN_accepted (nome mais recente)

Executando uma busca exata dos 10070 binômios em binomios-busca.txt (608189 linhas) no arquivo itis-binomios.txt, 3339 binômios foram encontrados. Destes, 290 são sinônimos.

A seguir, adicionamos à lista de nomes os binômios contidos nos inventários do Biota Neotropica [6], encontrados em `http://www.biotaneotropica.org.br/v11n1a/pt/item?inventory`. Esta lista adicionou 16099 binômios aos conhecidos, totalizando 624288 linhas no arquivo itis-inv-binomios.txt. Com este novo arquivo de referências, 5182 binômios do arquivo binomios-busca.txt foram reconhecidos.

Adicionamos então à busca os 2656 nomes tratados, num total de 12459 nomes. Destes, 6912 foram identificados (55.5% dos buscados, 54.5% do total). Posteriormente incluímos novos nomes até atingirmos um total de 12522 nomes utilizados na busca e 6961 foram identificados (55.6% dos buscados, 54.9% do total). Finalmente, após intensiva avaliação de nomes, obtivemos a identificação de 7022 de um total de 12589 nomes.

Foi implementada uma busca utilizando o web service do Catalogue of Life, que, em testes iniciais, identificou cerca de metade dos nomes que ainda não haviam sido reconhecidos. Entretanto, buscar em uma amostra maior ou em todos os nomes necessitará de mais tempo, pois há um intervalo de 10s entre cada consulta, para não sobrecarregar os servidores. De uma amostra de 300 nomes, 99 foram encontrados usando o Catalogue of Life. Adicionando o webservice do uBio [2], foram encontrados 162 nomes. Os resultados do uBio, entretanto, ainda precisam ser melhor tratados.

Incluímos uma busca aproximada para os nomes que não foram encontrados nas buscas exatas. Aqui buscamos por nomes na base local que tenham distância de Damerau-Levenshtein menor ou igual a 2 em relação ao binômio buscado. Também foi implementada uma busca utilizando o método `get_close_matches` da biblioteca difflib, que utiliza um algoritmo diferente, mas optamos por utilizar Damerau-Levenshtein. A distância de Damerau-Levenshtein conta operações de inserção, deleção, substituição de um caractere e transposição de dois caracteres adjacentes, todas com o mesmo peso. Com esta busca, alcançamos 216 de 300 binômios da amostra. Apesar de ser feita somente sobre a base local, a busca aproximada implementada demora alguns segundos por nome.

Aos webservices, adicionamos buscas a The Plant List [3], chegando a 229 acertos, e ao Global Names Index (GNI), com mais 22 nomes encontrados. Os resultados do GNI, entretanto, não consideramos completamente confiáveis.

Borboletas é uma classe em que há muitas espécies não encontradas. Procuramos um site para nos ajudar com estas buscas, mas há muitos sites sobre borboletas, e não é simples decidir quais são os mais confiáveis. Tentamos o Museu Britânico de História Natural (`nhm.ac.uk`), mas não chega ao nível de espécie. Lepidoptera.pro

parece bastante popular, mas não sabemos qual seu nível de seriedade científica. Assim, ainda não dispomos de um site ideal para borboletas. Por enquanto, vamos usando o Wikispecies. Entretanto, da amostra de 300 nomes, nenhum nome que ainda não havia sido descoberto foi identificado pelo Wikispecies.

Foram adicionados à lista local os nomes da lista de espécies de Lepidoptera da Mata de Santa Genebra `http://www.stagenebra.cnpm.embrapa.br/borbolet.html`. Na busca exata local, chegou-se a 7344 acertos. Na amostra de 300, 4 nomes não reconhecidos antes foram identificados exatamente e mais 2 na busca aproximada.

Para anfíbios, uma boa fonte, citada na checklist sobre anfíbios da revista Biota Neotropica, é o site mantido pelo *American Museum of Natural History* do `http://research.amnh.org/vz/herpetology/amphibia`.

Os nomes restantes parecem ser, em sua maioria, binômios com erros cuja versão correta não está na lista local de nomes e, portanto, não são identificados pela busca aproximada implementada.

5 Resultados

A base de dados local reúne informações do ITIS, dos inventários do Biota Neotropica [4], de uma versão anterior do Catalogue of Life, da lista de espécies de Lepidoptera da Mata de Santa Genebra e de duas famílias de aracnídeos, Gonyleptidae e Stigmaeidae, extraídas dos arquivos

"Gonyleptidae.txt" e

"Stigmaeidae.txt",

respectivamente encontrados em

`https://insects.tamu.edu/research/collection/hallan/acari/Family`.

A cada nome da base local, foi adicionada uma referência à fonte do dado, permitindo verificar posteriormente. Mais dois binômios foram adicionados separadamente, com suas respectivas referências. A Tabela 4 lista as bases utilizadas, na ordem em que são requisitadas.

No momento, nossa base local não inclui subespécies ou variedades, vai apenas até o nível de espécie.

Com estes dados, foram reconhecidos 7966 de 11632 nomes buscados, ou seja, 68,5% de identificação (o número de nomes buscados é menor, pois as duplicatas foram removidas para este teste).

Os webservices requisitados são: Catalogue of Life [8], uBio [2], The Plant List [3] e WoRMS [15]. Com estes quatro webservices, mais a busca aproximada, foram identificados 345 nomes em uma nova amostra de 400 (86.25%). A busca aproximada

também foi alterada. Nos casos em que é buscado só o gênero (como quando é usada a partícula "sp."), em vez de buscar por nomes com uma distância até 2, buscamos apenas com a distância de 1, pois, sem a informação extra da espécie, havia muitos falsos positivos, especialmente em nomes mais curtos.

Dos 54 nomes válidos (um foi para os casos ignorados) que não foram encontrados nesta segunda busca, fizemos uma terceira busca, levando em conta somente o gênero e utilizando todos os métodos de busca anteriores. Destes, 40 tiveram seus gêneros encontrados.

6 Conclusão

Apresentamos uma estratégia para a resolução de problemas e tradução de binômios encontrados no sistema SinBiota para uma classificação taxonômica suportada por uma iniciativa internacional.

A aplicação da estratégia resultou no reconhecimento e identificação de cerca de 68,5% dos binômios armazenados no SinBiota.

Acreditamos que a estratégia utilizada para a resolução de binômios nesse caso particular do SinBiota pode ser generalizada em situações de migração de sistemas legados de classificação taxonômica ou na verificação de correspondências entre binômios de classificações taxonômicas distintas.

Melhorias futuras na estratégia de resolução podem incluir o uso de identificação semântica por meio de ontologias.

7 Agradecimentos

Os autores gostariam de agradecer a Priscila Biller e João Paulo Pereira Zanetti pela ajuda com os scripts utilizados nesse trabalho.

References

[1] Integrated Taxonomic Information System (ITIS). Retrieved 2012-12-11 from `http://www.itis.gov/downloads`.

[2] ubio. `http://www.ubio.org/`.

[3] The Plant List, version 1, 2010. `http://www.theplantlist.org/`.

[4] *Biota Neotropica*, volume 11 n1a, 2011.

[5] Peter Bengtson. Open nomenclature. *Palaeontology*, 31(1):223–227, 1988.

[6] Biota Neotropica, 2012. URL=www.biotaneotropica.org.br.

[7] V.S. Chavan, N.S. Rane, A. Watve, and M. Ruggiero. Resolving taxonomic discrepancies: Role of electronic catalogues of known organisms. *Biodiversity informatics*, 2, 2005.

[8] Bisby F., Roskov Y., Culham A., Orrell T., Nicolson D., Paglinawan L., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Hernandez F., De Wever A., and Kunze T. (eds). Species 2000 & ITIS Catalogue of Life. Digital resource at `www.catalogueoflife.org/col/`.

[9] P. Feijao, C. Mira, C. Macario, J. Meidanis, and C.A. Joly. Taxonomic Data Migration in a Legacy Biodiversity Information System. In *Proceedings of The 8th International Conference on Ecological Informatics, 2012*, December.

[10] Life Science Identifier, 2012. URL=http://sourceforge.net/projects/lsid.

[11] R.D.M. Page. Taxonomic names, metadata, and the semantic web. *Biodiversity Informatics*, 3, 2006.

[12] Protótipo do Sistema de Informações Ambientais 2.0 do Programa BIOTA/FAPESP, February 2013. URL=http://sinbiota.biota.org.br.

[13] Resource Description Framework, 2010. URL=www.w3.org/TR/REC-rdf-syntax/.

[14] Sistema de Informações Ambientais do Programa BIOTA/FAPESP, September 2012. URL=http://sinbiota1.biota.org.br.

[15] Appeltans W., Bouchet P., Boxshall G. A., De Broyer C., de Voogd N. J., Gordon D. P., Hoeksema B. W., Horton T., Kennedy M., Mees J., Poore G. C. B., Read G., Stöhr S., Walter T. C., and Costello M. J. (eds). World Register of Marine Species. Digital resource at `http://www.marinespecies.org`.

Table 3: Binômios não tratados

Caso não tratado	Binômios
Detalhamento do binômio começa a partir da família, e muitas vezes não chega ao nível de gênero e espécie.	1. Annonaceae Annona 2. Araneidae gen. 3. Fabaceae-Caesalpinioideae Pterogyne 4. N.Gen. cf.vinnius 5. Thomisinaegen.N. sp.1 6. Thomisinaegen.N. sp.2 7. Thomisinaegen.N. sp.3 8. Thomisinaegen.N. sp.4 9. Thomisinaegen.N. sp.5 10. Thomisinaegen.N. sp.6
No binômio existem duas opções de gênero e nenhuma espécie, ou um gênero e duas opções de espécie. Se mantivermos as duas opções, o binômio não é encontrado na busca.	11. Caulobacter/Asticcacaulis sp. 12. Eugenia (?)neoglomerata/multicostata 13. Eugenia/Myrcia bicarinata 14. Leptodactylus “ocellatus,chaquensis”
O binômio contém alguns termos que não sabemos o significado. Com isso, há dúvidas se é preciso remover ou adaptar partes do binômio.	15. Glossoscolex cj1 16. Glossoscolex cj2 17. Phyllomedusa 3n 18. Phytoseius gp.plumifer 19. Pyramica(Md.Bocadelobo) appretiata 20. Pyrrhopyge sp.(rhacia?) 21. Uninucleate-Rhizoctonia sp 22. Zanthoxylum Z8

Table 4: Bases de dados utilizadas

Locais	Remotas
ITIS	CoL
Checklists	uBio
CoL	The Plant List
Lepidoptera Santa Genebra	WoRMS
insects.tamu.edu	
Extras	