Contribuições da ferramenta HerbCoRe no processo de construção de conjuntos de dados de herbário

Connor H. Iwamoto[1], Alexandre Y. Kajihara[1], Diego Bertolini[1], André L. Schwerz[1]

[1]Universidade Tecnológica Federal do Paraná, Campo Mourão (UTFPR)
87.301-899 – Campo Mourão – PR – Brazil

{connorharumi@alunos.utfpr.edu.br}

{alexandrey, diegobertolini, andreluis}@utfpr.edu.br

Abstract. *This paper presents Herbaria Collection Refinement (HerbCoRe), a tool designed to reduce the preprocessing burden in constructing datasets of herbarium specimens for training Machine Learning models, for many complex tasks such as plant identification. HerbCoRe retrieves and filters records from the speciesLink network, utilizes the Leipzig Catalogue of Vascular Plants to validate scientific names, and selects specimens identified by experienced professionals. A case study involving herbarium plants from ten families collected in Brazil demonstrated that HerbCoRe significantly reduces the workload by automating the refinement of botanical datasets.*

Resumo. *Este artigo apresenta a Herbaria Collection Refinement (HerbCoRe), uma ferramenta projetada para reduzir a carga de pré-processamento na construção de conjuntos de dados de espécimes de herbário, para o treinamento de modelos de Aprendizado de Máquina em diversas tarefas complexas, como a identificação de plantas. A HerbCoRe recupera e filtra registros da rede speciesLink, utiliza o Leipzig Catalogue of Vascular Plants para validar nomes científicos e seleciona espécimes identificados por profissionais experientes. Um estudo de caso envolvendo espécimes de dez famílias coletadas no Brasil demonstrou que o HerbCoRe reduz significativamente a carga de trabalho, ao automatizar o refinamento de conjuntos de dados botânicos.*

1. Introdução

A Lista Vermelha de Espécies Ameaçadas, do ano de 2025, assinalou que, entre as 70 mil espécies de plantas do mundo avaliadas por especialistas, em torno de 30 mil corriam perigo de extinção [International Union for the Conservation of Nature 2025]. Esse ritmo acelerado da perda da flora é extremamente preocupante, pois não há como manejar, conservar e utilizar, de forma sustentável, a biodiversidade, sem ter informações taxonômicas das espécies [Ebach et al. 2011]. Assim, nas últimas décadas, pesquisadores têm se esforçado para documentar a diversidade de plantas do planeta, por meio da criação de herbários virtuais, digitalmente interligados e de acesso livre e aberto [Mata-Montero and Carranza-Rojas 2016].

Atualmente, milhões de plantas de herbário digitalizadas estão disponíveis, por meio de redes, como a *species*Link, criada para compartilhar dados científicos de coleções biológicas do Brasil e do exterior [*Species*Link Network 2026]. Um grande desafio, ainda,

é a existência de milhares de espécies depositadas nos herbários, à espera de identificação. Isso porque a determinação manual de plantas é um trabalho extremamente lento, complexo e altamente especializado [Mata-Montero and Carranza-Rojas 2016].

Para lidar com esse problema, métodos de Aprendizado de Máquina têm sido utilizados na tarefa de identificação de espécimes de herbário [Kajihara et al. 2022, Kajihara et al. 2025, Mata-Montero and Carranza-Rojas 2016]. Entretanto, a qualidade dos dados das bases de herbário ainda é um empecilho para o uso desses modelos na identificação automatizada de plantas. Isso porque essas bases de dados têm problemas em comum: falta de padronização de metadados e de imagens de exsicatas; registro de informações incorretas, incompletas ou desatualizadas; erros taxonômicos etc. [Silva et al. 2010].

Considerando que a qualidade de dados de bases botânicas é essencial para que as informações disponíveis sejam utilizadas, de forma eficiente, e não influenciem, negativamente, por exemplo, nos resultados de pesquisas de identificação automatizada de plantas, neste trabalho é apresentada a ferramenta *Herbaria Collection Refinement* (HerbCoRe), que visa contribuir para reduzir a carga de trabalho no pré-processamento, durante a construção de conjuntos de dados de herbário, destinados ao treinamento de modelos de classificação baseados em Aprendizado de Máquina. Além de unificar o pré-processamento em um único sistema, diminuindo a necessidade de *scripts* específicos e de filtragem manual, essa ferramenta contribui para a reprodutibilidade na elaboração de conjuntos de dados.

O restante deste artigo está organizado da seguinte forma: na Seção 2 são descritos trabalhos relacionados ao realizado neste estudo; na Seção 3, o método empregado; na Seção 4 é descrita a ferramenta HerbCoRe e suas funcionalidades; e na Seção 5 é apresentado um estudo de caso que demonstra a viabilidade da ferramenta. Por fim, considerações finais e perspectivas para trabalhos futuros são apresentadas na Seção 6.

2. Trabalhos relacionados

No Brasil, um banco cujos dados têm sido muito estudados é o do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. [Silva et al. 2010] analisaram esse banco, e constataram problemas na entrada dos dados da flora brasileira: (i) redação incorreta dos nomes das plantas; (ii) preenchimento parcial dos campos; (iii) imprecisão nas coordenadas geográficas; (iv) importação parcial dos campos que contêm textos longos; (v) e falta de padronização na escrita dos nomes dos coletores, plantas e localidades. Há, também, falta de atualização dos dados.

Para resolver a questão da falta de padronização dos nomes dos coletores das amostras, [Silva 2016] propôs uma abordagem de mineração de dados. Ele analisou 916.799 registros de coletas do Jardim Botânico. Após a limpeza dos dados, aplicou o algoritmo Apriori, tendo considerado como unidade de estudo, o nome de um coletor associado ao local de coleta, para identificar coletas com diferenças na escrita de seu nome. Após o levantamento e a identificação das inconsistências nos nomes suspeitos, efetuou a substituição pelo mais usado, e a padronização pelo formato correto.

Em um trabalho posterior, [Silva et al. 2019] desenvolveram uma ferramenta de importação e validação de dados, para eliminar ou reduzir erros na inclusão de novas informações. Durante o processo de importação, foram aplicados filtros aos dados

por meio de uma planilha, para agilizar a verificação de erros. A ferramenta realizou validações para verificar a ocorrência de erros, por exemplo, taxonômicos, na data e localização geográfica da coleta do espécime e na redação do nome do coletor.

Esses trabalhos, acima citados, visaram melhorar a qualidade de dados de sistemas de herbário. A ferramenta HerbCoRe, desenvolvida neste estudo, destina-se à criação de conjunto de dados de herbário bem rotulados, para treinamento de modelos de Aprendizagem de Máquina.

3. Método

Nesta seção são apresentados a rede *species*Link e o *Leipzig Catalogue of Vascular Plants*, cujos conjuntos de dados foram utilizados para a elaboração da ferramenta proposta neste trabalho.

3.1. A rede *species*Link

A rede colaborativa *species*Link integra informações primárias sobre flora, fauna e microrganismos, armazenadas em museus e herbários, disponibilizando-as, de forma livre e aberta na Internet, de forma a contribuir para a pesquisa, a educação e a elaboração de políticas de conservação e uso sustentável da biodiversidade. Essa rede permite: realizar buscas por famílias, espécies, determinadores, coletores, locais de coleta etc.; comunicar-se com a sua API[1] (do inglês, *Application Programming Interface*); e obter imagens de exsicatas e gráficos [*Species*Link Network 2026].

Neste trabalho, o principal foco da *species*Link foi a sua API, que permite acessar dados da rede por meio de chamadas HTTP GET, e possibilita usar parâmetros de consulta. As respostas são fornecidas em formato JSON. Com essa API é possível obter metadados de espécimes, listas de coleções e instituições participantes, informações sobre conjuntos de dados específicos e registros de biodiversidade [*Species*Link Network 2026].

3.2. *Leipzig Catalogue of Vascular Plants* (LCVP)

O *Leipzig Catalogue of Vascular Plants* (LCVP) é uma lista com mais de 1 milhão de nomes científicos de todas as plantas vasculares já descritas, e cerca de 350 mil nomes de espécies aceitos oficialmente. Essas informações são disponibilizadas a partir de um pacote na linguagem R. Nesse catálogo, cada nome de espécie é classificado em: nome aceito; nome sinônimo, que equivale a um nome aceito; e nome não resolvido, cuja veracidade é incerta [Freiberg et al. 2020]. Neste trabalho, foi utilizado o LCVP para validar os nomes científicos das plantas disponíveis na rede *species*Link, de forma a prover uma camada de qualidade ao conjunto, visto que não é produtivo, por exemplo, ter duas exsicatas de uma mesma espécie com nomes científicos diferentes.

4. Ferramenta HerbCoRe

A ferramenta HerbCoRe foi desenvolvida para reduzir a carga de trabalho na etapa de pré-processamento, do processo de construção de conjuntos de dados de herbário, destinados ao treinamento de modelos de classificação baseados em Aprendizado de Máquina, a serem utilizados, por exemplo, em tarefas de identificação automatizada de plantas. Além de

[1]Documentação disponível em `https://specieslink.net/ws/1.0/`

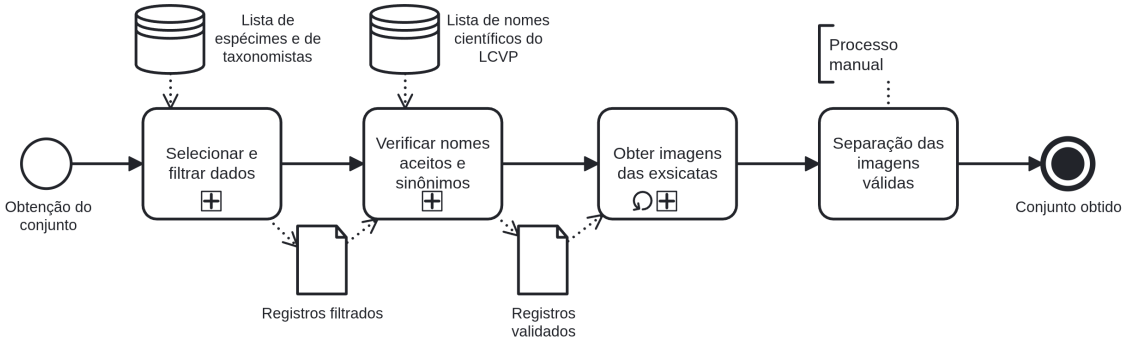
unificar o pré-processamento em um único sistema, diminuindo a necessidade de *scripts* específicos e de filtragem manual, essa ferramenta contribui para a reprodutibilidade na elaboração de conjuntos. A seguir, na Seção 4.1, são apresentadas as principais funcionalidades da HerbCoRe e, na Seção 4.2, são descritos os detalhes de sua implementação.

4.1. Funcionalidades

A Figura 1 ilustra as quatro principais tarefas da ferramenta:

- Coleta e filtragem de dados de amostras de plantas de uma ou mais famílias ou espécies botânicas, a partir da API da rede *species*Link.
- Consulta ao LCVP, para confirmar se os nomes científicos das plantas das amostras são: nomes aceitos, adotados oficialmente pela Botânica; nomes válidos, mas sinônimos de nomes aceitos; ou nomes que não constam nesse catálogo.
- Obtenção das imagens das plantas cujos nomes científicos são válidos, isto é, nomes aceitos e sinônimos;
- Classificação das imagens das plantas, separando-as em imagens de exsicatas (plantas desidratadas) válidas ou imagens de ruídos (por exemplo, plantas fotografadas em ambiente natural ou etiquetas de identificação das plantas).

Figura 1. Fluxo de trabalho para a construção de um conjunto de dados de plantas de herbário



Fonte: Autoria própria (2026).

Cada uma dessas tarefas é detalhada a seguir. A coleta e filtragem de dados são, inicialmente, divididas em subetapas. O processo de obtenção do conjunto de dados é realizado a partir da rede *species*Link, em que são especificadas a(s) família(s) ou espécie(s) botânica(s) desejada(s), e os resultados são armazenados em uma base de dados. Em seguida, é realizada a limpeza e a filtragem, para eliminação de inconsistências e separação dos dados com base em critérios desejados para a formação do conjunto.

Nessa primeira etapa, a HerbCoRe realiza o levantamento dos taxonomistas que foram responsáveis pela identificação de cada amostra, isto é, pela determinação da família, gênero e espécie dessa planta que já é conhecida pela Botânica. Esse tipo de informação é muito importante, pois a tarefa de identificação é altamente complexa e difícil, e por isso aumenta a confiabilidade da determinação, se tiver sido executada por um especialista com grande experiência em uma família ou espécie botânica.

A HerbCoRe ranqueia os taxonomistas, em ordem decrescente, com base na frequência de amostras de uma determinada família ou espécie botânica identificada por

cada um deles. Esse processo é realizado por meio do método da deduplicação *fuzzy*, e da biblioteca *rapidfuzz*, do Python. A ferramenta: agrupa os nomes dos taxonomistas, a partir da similaridade entre eles, utilizando algoritmos de distância do *rapidfuzz*; cria subgrupos de nomes dos taxonomistas; e escolhe o nome mais frequente, como seu representante. Por exemplo, no *species*Link, o nome do taxonomista "Ravenna, P. F.", que identificou amostras da família botânica Amaryllidaceae, foi redigido de várias formas: Ravenna, P.; Ravenna; Ravenna, P.; P. Ravenna etc. Todos esses nomes são reunidos em um subgrupo criado pelo *rapidfuzz*, e unificados sob o nome de "Ravenna, P. F.".

Na lista em ordem decrescente, o taxonomista "Ravenna, P. F." obteve a segunda colocação, tendo identificado 335 amostras da família Amaryllidaceae. Essa seleção, que prioriza as quantidades de identificações realizadas, parte da premissa de que um volume superior de trabalho reflete maior experiência e competência naquele grupo botânico. Dessa forma, busca-se assegurar a confiabilidade taxonômica das amostras a serem utilizadas na composição do conjunto de dados, com a seleção daquelas que foram identificadas por especialistas com experiência consolidada. A HerbCoRe oferece, também, um mecanismo de seleção manual. O usuário pode indicar nomes específicos de taxonomistas. Esse recurso é útil para estudos que exigem elevado rigor taxonômico, pois possibilita compor conjuntos de dados com amostras identificadas por um grupo seleto de especialistas em determinadas famílias ou espécies botânicas.

Em uma segunda etapa, a HerbCoRe realiza consulta ao LCVP, para confirmar se os nomes científicos das espécies, que constam nos metadados, são válidos para a Botânica. As amostras das plantas são categorizadas em: com nomes aceitos oficialmente; com nomes válidos, mas que são sinônimos, e por isso precisam ser substituídos pelos nomes oficiais; e com nomes científicos não resolvidos, cuja inconsistência será anotada, para ser resolvida manualmente. Há, também, a possibilidade de o nome científico não constar no LCVP; nesse caso, essa amostra é excluída do conjunto. Com a atualização dos sinônimos por nomes aceitos, os registros passam a ser considerados válidos.

Após a validação dos nomes das plantas, são obtidas as imagens das plantas desidratadas de herbário. Como não é possível ter acesso a essas imagens, por meio da API da rede *species*Link, é preciso utilizar um *crawler* elaborado por [Kajihara 2023], que permite obter as imagens a partir de seus códigos de barras. Essas imagens são adicionadas em uma pasta por esse *crawler*.

No *species*Link é possível encontrar imagens que não são de plantas desidratadas apresentadas em exsicatas, mas de plantas vivas, fotografadas em ambiente natural ou logo após a sua coleta. Essas imagens de plantas vivas, assim como imagens de etiquetas de identificação de plantas, representam ruídos, e por isso suas amostras são excluídas do conjunto. O *species*Link não apresenta um campo específico para imagens de plantas desidratadas (exsicatas) ou vivas; assim, é preciso realizar essa separação após a filtragem dos registros. Esse processo deve ser feito manualmente. Uma vez obtidas as imagens de exsicatas, pode ser iniciada a tarefa de classificação das plantas. Como parte de um trabalho em andamento, estão sendo anotadas amostras, para treinar um modelo preditivo que seja capaz de identificar, automaticamente, imagens de exsicatas.

4.2. Implementação

A HerbCoRe é escrita em *Python*, e tem três arquivos *.py*, que visam: coletar espécimes, no *species*Link; verificar nomes aceitos; e selecionar taxonomistas que realizaram o maior número de identificações de plantas de uma determinada família ou espécie. Como bibliotecas importantes, utilizadas na implementação, podem ser citadas: a *rpy2.robjects*, para empregar a biblioteca em R do Leipzig, a *lcvp_plants* [Freiberg et al. 2020], *os*, *argparse*, *json*, *pandas*, *pymysql* e *csv*. A ferramenta encontra-se disponível no repositório `https://github.com/connorharu/HerbCoRe`.

5. Estudo de caso: dez famílias de plantas angiospermas

Nesta seção, é apresentado um estudo de caso em que a ferramenta HerbCoRe foi utilizada para a construção de um conjunto de dados de herbário, formada por dez famílias de plantas pertencentes ao grupo das angiospermas (plantas com flores), coletadas no Brasil e selecionadas, de forma aleatória, na rede *species*Link. Considerando que as angiospermas formam o maior grupo botânico do planeta, e que 45% de suas espécies correm risco de extinção [Antonelli et al. 2023], é urgente criar ferramentas que contribuam para a identificação das plantas desse grupo.

Foram obtidas e armazenadas 243.588 amostras das dez famílias (Figura 2), cada uma composta pelas seguintes quantidades de espécimes: 16.013 (Amaryllidaceae), 22.914 (Calophyllaceae), 39.390 (Combretaceae), 9.530 (Portulacaceae), 42.857 (Plantaginaceae), 13.706 (Peraceae), 34.720 (Oxalidaceae), 8.495 (Escalloniaceae), 30.611 (Dilleniaceae) e 25.352 (Dioscoreaceae). Foram excluídas as amostras coletadas no exterior ou que não apresentavam informação sobre o local de coleta da planta. Registros com erros de ortografia na palavra “Brasil” foram corrigidos. Em função de flutuações na rede *species*Link, decorrentes de remoções de registros e adição de novos, pode ocorrer uma diferença no total de amostras obtidas, em uma nova requisição.

Figura 2. Imagens de exsicatas, com espécimes de algumas famílias do estudo (da esquerda para a direita): Dioscoreaceae, Amaryllidaceae, Oxalidaceae, Dilleniaceae e Calophyllaceae

Fonte: *species*Link Network (2026, não paginado).
Nota: Exsicatas do Herbário Alexandre Leal Costa (ALCB).

Nas coleções de herbário, o código de barras que acompanha uma amostra, por ser exclusivo de um espécime, permite a sua identificação, de forma simples e confiável [Symbiota2 Collections of Arthropods Network (SCAN) 2025]. Em função disso, nessa primeira etapa, foram filtrados e mantidos apenas os registros com códigos de barras. Informações sobre os estados brasileiros onde foram realizadas as coletas

são relevantes para a realização de diversos tipos de estudos, como análises do nível de conservação de uma determinada espécie botânica, a partir de sua distribuição geográfica ao longo do tempo. Considerando a importância dessa informação, foram mantidos no conjunto de dados apenas os registros que continham o estado onde a planta foi coletada.

Como é possível observar na Tabela 1, a exclusão das amostras coletadas no exterior, sem código de barras, sem estado onde foi feita a coleta e com campos em branco ou com informações preenchidas em campos incorretos, resultou em uma redução média de 50,47% do conjunto inicial. A família Escalloniaceae foi a que teve a maior perda, com redução de 80,74% do total de espécimes. O principal fator que levou à perda de amostras foi a exclusão das plantas coletadas no exterior.

A seguir, foi realizado o levantamento dos taxonomistas responsáveis pelas identificações das plantas. Optou-se pela seleção das amostras identificadas pelos 20 taxonomistas mais atuantes em cada família botânica; no entanto, esse parâmetro pode ser ajustado, de acordo com os objetivos da pesquisa em que o conjunto de dados será utilizado. O uso desse critério levou a uma redução significativa das amostras válidas (Tabela 1). Por exemplo, na família Combretaceae, a filtragem inicial resultou em 21.749 espécimes, e apenas 5.969 foram identificados pelos 20 taxonomistas mais atuantes. Isso significa que 15.780 amostras foram identificadas por um grande número de botânicos que têm se dedicado ao estudo dessa família.

Tabela 1. Número de amostras e redução percentual, na primeira etapa

Família	**Coleta n (%)**	**Filt. inicial n (%)**	**Identif. 20 taxonomistas n (%)**
Amaryllidaceae	16.013 (100%)	7.152 (55,34%)	2.609 (83,71%)
Calophyllaceae	22.914 (100%)	13.969 (39,04%)	5.315 (76,80%)
Combretaceae	39.390 (100%)	21.749 (44,79%)	5.969 (84,85%)
Portulacaceae	9.530 (100%)	4.851 (49,10%)	1.848 (80,61%)
Plantaginaceae	42.857 (100%)	17.540 (59,07%)	6.583 (84,64%)
Peraceae	13.706 (100%)	8.975 (34,52%)	3.007 (78,06%)
Oxalidaceae	34.720 (100%)	17.602 (49,30%)	6.983 (79,89%)
Escalloniaceae	8.495 (100%)	1.636 (80,74%)	657 (92,27%)
Dilleniaceae	30.611 (100%)	18.333 (40,11%)	6.224 (79,67%)
Dioscoreaceae	25.352 (100%)	11.980 (52,75%)	4.774 (81,17%)

Fonte: Autoria própria (2026).

Para a validação do nome científico de cada planta, disponibilizado como metadado, foi utilizado o LCVP. A consulta a essa lista permitiu verificar se o nome científico atribuído à amostra correspondia a um nome aceito pela Botânica, a um sinônimo de um nome aceito ou a um nome inexistente nesse catálogo. Os sinônimos foram atualizados, e os respectivos nomes aceitos foram registrados em novos campos. Essas amostras foram mantidas no conjunto de dados e considerados como nomes válidos.

A quantidade de sinônimos encontrados foi grande (Tabela 2). Por exemplo, no grupo das Combretaceae, em 553 amostras os nomes eram sinônimos. Essa família foi justamente aquela em que, na etapa anterior, constatou-se que 15.780 amostras foram identificadas por taxonomistas que não estavam na lista dos 20 mais atuantes. Essa situação exemplifica o alto grau de dificuldade da tarefa de identificação botânica e, portanto, a *expertise* necessária para a sua realização. Na Tabela 2 também são apresentadas

as quantidades de amostras cujos nomes não constavam no LCPV. A família em que houve o maior número de exclusão de amostras foi a Dilleniaceae (n=189).

Tabela 2. Número de amostras e redução percentual, na segunda etapa

Família	**Nome inválido (ausente no LCVP)** n	**Sinônimo** n	**Nome válido** n **(%)**
Amaryllidaceae	117	458	2.492 (84,44%)
Calophyllaceae	93	16	5.222 (77,21%)
Combretaceae	45	553	5.924 (84,96%)
Portulacaceae	9	35	1.839 (80,70%)
Plantaginaceae	180	481	6.403 (85,06%)
Peraceae	45	38	2.962 (78,39%)
Oxalidaceae	88	329	6.894 (80,14%)
Escalloniaceae	25	12	632 (92,56%)
Dilleniaceae	189	61	6.035 (80,28%)
Dioscoreaceae	58	394	4.716 (81,40%)

Fonte: Autoria própria (2026).

Apesar de a redução percentual ter sido pequena, entre o final da primeira e segunda etapas, a realização da validação dos nomes científicos é importante para a formação de um conjunto de dados confiável. Quando não é feita a atualização de um sinônimo por um nome aceito, duas plantas com nomes científicos diferentes são consideradas, equivocadamente, com sendo de duas espécies diferentes. Por exemplo, na família Peraceae, *Pera anisotricha* Müll.Arg. é o nome aceito de uma das espécies. Entretanto, esse nome tem outros três sinônimos, ou seja, nomes científicos possíveis, mas que não são considerados oficiais (*Pera barbinervis* Pax & K.Hoffm., *Pera bahiana* Ule e *Spixia barbinervis* Klotzsch). Nesse caso, temos uma única espécie, e não quatro. Em estudos de identificação automatizada de plantas, é fundamental que o treinamento dos modelos de Aprendizado de Máquina seja realizado com classes confiáveis e livres de duplicações.

A seguir, foi realizada a busca das imagens das plantas desidratadas, apresentadas em exsicatas. A exclusão das imagens, consideradas como ruídos, resultou nos resultados apresentados na Tabela 3. A taxa de ruído variou bastante entre as famílias. Por exemplo, 8 amostras da Escalloniaceae continham ruídos, e 848 da família Combretaceae.

Ao final do pré-processamento (Tabela 1), realizado pela HerbCoRe, a redução percentual, entre a amostra inicial e a final, foi significativa, em todas as famílias, visto que foi $> 82\%$. A maior redução foi observada na família Escalloniaceae: entre as 8.495 amostras retornadas pelo *species*Link, restaram apenas 461 ao final do pré-processamento, isto é, houve uma perda de 94,57% da quantidade inicial. Do conjunto inicial de 243.588 amostras retornadas, inicialmente, pelo *species*Link, referente às dez famílias botânicas, restaram 32.251 que preencheram todos os critérios de inclusão empregados nas três etapas do pré-processamento.

6. Considerações finais

A crescente demanda por conjuntos de dados botânicos de alta qualidade, capazes de fornecer informações confiáveis que não comprometam resultados de estudos, como os de identificação automatizada de plantas, motivou o desenvolvimento da HerbCoRe. Esta

Tabela 3. Número de amostras e redução percentual, na terceira etapa

Família	**Imagens de exsicatas** n	**Ruídos** n	**Amostras válidas** n **(%)**
Amaryllidaceae	2.058	78	1.980 (87,64%)
Calophyllaceae	4.543	519	4.024 (82,44%)
Combretaceae	5.684	848	4.836 (87,72%)
Portulacaceae	1.580	190	1.390 (85,41%)
Plantaginaceae	4.777	313	4.464 (89,58%)
Peraceae	2.812	529	2.283 (83,34%)
Oxalidaceae	4.993	371	4.622 (86,69%)
Escalloniaceae	469	8	461 (94,57%)
Dilleniaceae	5.424	794	4.630 (84,87%)
Dioscoreaceae	3.683	122	3.561 (85,95%)

Fonte: Autoria própria (2026).

ferramenta foi projetada para apoiar a construção de conjuntos de dados de herbário, contribuindo para a padronização, validação e filtragem de informações botânicas relevantes.

Neste estágio do trabalho, foi conduzido um estudo de caso com espécimes de dez famílias botânicas, disponíveis na rede *species*Link. Os resultados evidenciaram que a filtragem dos dados, a validação das identificações das plantas, realizadas por taxonomistas experientes, a validação dos nomes científicos e, ainda, a validação das imagens, são fundamentais para a obtenção de conjuntos de dados consistentes, especialmente quando destinados ao treinamento de modelos de Aprendizado de Máquina para, por exemplo, identificação de plantas de herbário.

Como trabalhos futuros, pretende-se: ampliar a HerbCoRe, por meio da integração com outros bancos de dados botânicos, como o do GBIF[2], que é a maior rede de dados - de acesso aberto - de biodiversidade do mundo; automatizar a configuração, validação e verificação de campos dos registros; e incorporar modelos de Aprendizado de Máquina para a filtragem automática de imagens de plantas desidratadas de herbário. Espera-se, assim, tornar o processo de curadoria mais escalável e reprodutível, fortalecendo o uso de dados de herbário tanto em estudos botânicos quanto em aplicações computacionais.

7. Agradecimentos

Agradecemos pelo apoio financeiro da UTFPR, por meio do Programa Institucional de Iniciação Científica e Tecnológica.

Referências

Antonelli, A., Fry, C., Smith, R. J., Eden, J., Govaerts, R. H. A., Kersey, P., Nic Lughadha, E., Onstein, R. E., Simmonds, M., and Zizka, A. (2023). *State of the World's Plants and Fungi, 2023: Tackling the Nature Emergency: Evidence, Gaps and Priorities*. Royal Botanic Gardens, Kew. DOI:https://doi.org/10.34885/wnwn-6s63.

Ebach, M., Valdecasas, A. G., and Wheeler, Q. (2011). Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics*, 27:550–557. DOI: https://doi.org/10.1111/j.1096-0031.2011.00348.x.

[2]https://www.gbif.org/

Freiberg, M., Winter, M., Gentile, A., Zizka, A., Muellner-Riehl, A. N., Weigelt, A., and Wirth, C. (2020). LCVP, the Leipzig Catalogue of Vascular Plants, a new taxonomic reference list for all known vascular plants. *Scientific data*, 7(416). DOI: https://doi.org/10.1038/s41597-020-00702-z.

International Union for the Conservation of Nature (2025). The IUCN Red List of Threatened Species. Version 2025-2. Disponível em: https://www.iucnredlist.org. Acesso em: 6 fev. 2026.

Kajihara, A. Y. (2023). Segmentação e classificação de espécimes de herbário: um estudo de caso com a família Piperaceae Giseke. Master's thesis, Universidade Tecnológica Federal do Paraná - Campo Mourão. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/32342. Acesso em: 25 jan. 2026.

Kajihara, A. Y., Bertolini, D., and Schwerz, A. L. (2022). Identification of herbarium specimens: a case study with Piperaceae Giseke family. In *Proceedings of the 29th International Conference on Systems, Signals and Image Processing (IWSSIP'22)*, pages 1–4, Sofia. DOI: https://doi.org/10.1109/IWSSIP55020.2022.9854444.

Kajihara, A. Y., de Queiroz, G. A., Caxambú, M. G., Oliveira, L. E. S., Bertolini, D., and Schwerz, A. L. (2025). A database for automatic identification of herbarium specimens in Piperaceae family. *Multimedia Tools and Applications*, 84:44427–44455. DOI: https://doi.org/10.1007/s11042-025-20883-2.

Mata-Montero, E. and Carranza-Rojas, J. (2016). Automated plant species identification: Challenges and opportunities. In *ICT for promoting human development and protecting the environment. WITFOR 2016. IFIP advances in information and communication technology*, pages 26–36. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-44447-5_3.

Silva, L. A. E. (2016). A data mining approach for standardization of collectors names in herbarium database. *IEEE Latin America Transactions*, 14(2):805–810. DOI: https://doi.org/10.1109/TLA.2016.7437226.

Silva, L. A. E., Barros, R. O., Dalcin, E., Zimbrão, G., and de Souza, J. M. (2010). Abordagem colaborativa para a melhoria da qualidade de dados em bases de dados botânicas. In *XXX Congresso da Sociedade Brasileira de Computação*, pages 535–544, Belo Horizonte. Disponível em: https://sol.sbc.org.br/index.php/wcama/article/view/32184/31986. Acesso em: 12 março 2026.

Silva, L. A. E., Oliveira, F. A., Lima, R. O., Bellon, E., Ribeiro, R. d. S., Clemente, L. d. S., Medeiros, E. v. S. d. S., and Magdalena, U. R. (2019). Tool for validation and import in herbarium database. *Rodriguésia*. Article no. e03222017. DOI: http://dx.doi.org/10.1590/2175-7860201970032.

Symbiota2 Collections of Arthropods Network (SCAN) (2025). Barcodes overview – Lichens. Disponível em: https://scan-all-bugs.org/?page_id=1499. Acesso em: 12 março 2026.

*Species*Link Network (2026). *species*link. Disponível em: https://specieslink.net/. Acesso em: 20 fev. 2026.