Assessing the risk of extinction of Brazil's flora: A computational approach based on micro-services and geospatial analysis

Diogo Souza[1], Eline Martins[1], Eduardo C Dalcin[2]

[1]Centro Nacional de Conservação da Flora - CNCFlora
Rua Pacheco Leão 915 - Rio de Janeiro - RJ - 22460-030

[2]Instituto de Pesquisas Jardim Botânico do Rio de Janeiro - JBRJ
Rua Pacheco Leão 915 - Rio de Janeiro - RJ - 22460-030

{diogo, eline}@cncflora.net, edalcin@jbrj.org

Abstract. *This paper describes a computational tool developed to assess the risk of extinction of flora according to the "B" of the IUCN Red List Categories and Criteria System. The tool consists of a set of systems arranged in a micro-services architecture and performs geospatial analysis in a significant set of data in an automated manner, with relatively low computational cost.*

Resumo. *Este trabalho descreve uma ferramenta computacional desenvolvida para avaliar o risco de extinção da flora brasileira segundo o critério "B" do sistema de critérios e categorias da União Internacional para a Conservação da Natureza e dos Recursos Naturais - IUCN. Esta ferramenta é formada por um conjunto de sistemas organizados em uma arquitetura de micro-serviços, e realiza análises geoespaciais em um conjunto significativo de dados, de forma automatizada, com relativo baixo custo computacional.*

1. Introduction

The extinction risk assessment is the first step in the conservation of a species, and should provide a scientific and objective assessment of the likelihood of a species becoming extinct at one time if the circumstances in which the species is found remains [Mace & Lande 1991]. In Brazil, to carry out the assessment of extinction risk, the system of criteria and categories of the International Union for Conservation of Nature was adopted - IUCN [IUCN 2001]. This system is composed of five quantitative criteria for a robust risk assessment and is scientifically based and can be applied consistently by different people, and also facilitates the comparison between assessments of a species or of different species [IUCN 2001].

The assessment of the risk of extinction of all known species is a global challenge agreed by signatories to the Convention on Biological Diversity (CBD) through the target 2 of the Global Strategy for Plant Conservation (GSPC): "an overall assessment of conservation status of all known plant to guide conservation actions" by 2020 [GSPC 2012]. Currently, Brazil has listed 46,113 species of flora [Flora of Brazil 2020 2016], but even with

a continuous and dedicated effort of the National Center for Plant Conservation (CNCFlora), since 2010, only about 11% of these were evaluated for the risk of extinction..

The remaining large knowledge gap about the conservation status of our flora, together with the challenge of evaluating it completely by 2020, shows that the use of technologies that allow rapid assessment of the risk of extinction supporting trained professionals to make decisions on the final categorization of species is essential.

In this context, this tool performs a risk assessment based on the criteria B of the IUCN [IUCN 2001], using the Extent of Occurrence (EOO), Area of Occupancy (AOO) and the number of subpopulations to categorize the risk of extinction. This criteria mainly uses the spatial information of the occurrence of the species and is the most widely used criteria for extinction risk assessments in Brazil [Martinelli and Moraes 2013].

Occurrence data of flora species comes mainly from herbarium collections, formed by specimens of dried plants mounted on a sheet of cardboard accompanied by a label with data relating to that sample. These data represent the occurrence of a biological specimen in time and space and are the primary source of data for studies on biodiversity and conservation.

In the last decade, significant investments were made to render the data from scientific collections digitally accessible [Beaman et al. 2012, Blagoderov et al. 2012]. More recently, in Brazil, an initiative of the Ministry of Science, Technology and Innovation (MCTI) has promoted the digitalization and provided access to data from collections from different national herbaria [Gadelha et al. 2014, SiBBr 2016]

The initiative of MCTI adopted the Darwin Core standard [Wieczorek et al. 2012] and the Integrated Publishing Toolkit (IPT) [Robertson et al. 2014] for data publication [Gadelha et al. 2014]. Likewise, the Rio de Janeiro Botanical Garden Research Institute (JBRJ) also offers its occurrence data and Species List Flora of Brazil using the same standards and tool.

The access to those data sets provided by the IPT publishing tool, added with the demand of the National Center for Conservation of Flora - CNCFlora to assess the risk of extinction of all flora, served as motivation and made possible the development of this tool.

2. Methodology

2.1. Data gathering

Two data sets are consumed by the tool in their risk assessment process: the *taxonomic set*, which is a list of accepted scientific names and their synonyms, and the *set of occurrences* of these names.

The taxonomic set is formed based on the resource "Lista de Espécies da Flora do Brasil", version 393.53 [Forzza 2014], made public by the IPT tool at http://ipt.jbrj.gov.br/jbrj/. In this version, the set recorded 39,506 accepted names and 18,376 synonyms for these names, totaling 57,882 scientific names.

The occurrences set is formed by made public instances of resources in the following IPTs:

- Inst. de Pesquisas Jardim Botânico do Rio de Janeiro, available at http://ipt.jbrj.gov.br/jbrj/
- Projeto REFLORA, available at http://ipt.jbrj.gov.br/reflora/

- Centro de Referência em Informação Ambiental - CRIA, available at http://ipt1.cria.org.br/ipt/

The occurrences set then totaled 3,632,660 records related to the names present in the taxonomic set mentioned above.

2.2. Analysis Performed

With the collected data sets - taxonomic and occurrences - the tool performs a series of calculations, explained below, for each species, represented then by its accepted name and associated synonyms.

In a first step, the tool selects, for each species, two subsets of occurrence data, based on the date of the sample. The historical subset (M) contains samples collected for over 50 years; and the recent subset (R) contains samples collected in the last 50 years. For each of these subsets, the tool selects for the following calculations only records that have geographic coordinate pairs (latitude and longitude).

In a second step, the tool performs the calculations of the geospatial Extent of Occurrence and Area of Occupancy, as described in IUCN 2012. The Extent of Occurrence (EOO) is the area contained within the shortest continuous imaginary boundary (*minimum convex polygon*) which can be drawn to encompass all known, inferred, or projected sites presently occupied by the taxon and the Area of Occupancy (AOO) is the area within the EOO which is *actually* occupied by the taxon (usually measured by overlaying a grid and counting number of occupied cells). In addition to these two calculations, the tool performs the calculation for subpopulations using the method of "circular buffer" [Rivers at al. 2010].

In its final stage, for each species, the tool confronts the results of calculations performed within these parameters (EOO, AOO and subpopulations) considered by the criterion "B", which is one of the five criteria to define the possible categories and justification of risk assessment in accordance of the methodology established by IUCN (Table 1):

Table 1. Possible results of risk assessments using the IUCN Red List Categories and Criteria - IUCN 2001

Category	Criterion	Justification
{CR,EN,VU}	Geographic range in the form of Extent of occurrence (EOO)- criteria B1	The size of EOO
{CR,EN,VU}	Geographic range in the form of Area of occupancy (AOO)- criteria B2	The size of AOO
{NT, LC}	N/A	The size of EOO and AOO
DD	N/A	Insufficient number of occurrences

CR, EN and VU are threatened categories mean critically endangered, endangered and vulnerable, respectively. NT and LC are categories of risk assessment, but not a threatened categoria, NT: not threatened and LC: least concern. DD: data deficient, means the specie has so few information and it is not possible to evaluate its.

3. Architecture and Technologies

With an architecture based on micro-services [Lewis & Fowler 2014], the tool is divided into different systems, which are independent and interchangeable. The main advantage of this approach is the possibility of independent development of the parties, which have loose coupling.

The systems are based on web services and communicate with each other through APIs using the HTTP standard transport protocol and the JSON format for encapsulating the information.

The Taxadata system, developed in PHP 5.7 programming language, and using as persistence the SQLite database, is responsible for importing and treating the taxonomic data set, as well as exposing the data as a Web service (Figure 1)..

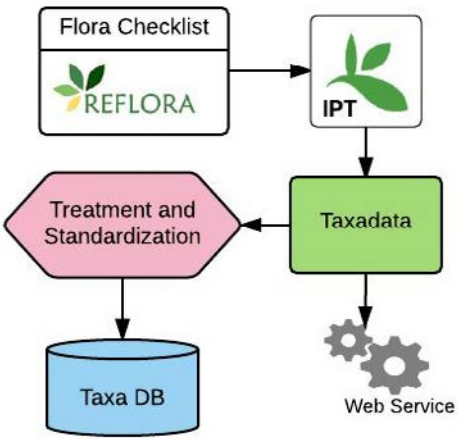

Figure 1. Taxadata system scheme

To aggregate the occurrence data, the Darwin Core Bot system was developed. This can import resources from IPTs, and also expose the set of occurrences as an API web service. This system was developed in Clojure programming language and uses the SQLite database to store the data (Figure 2).

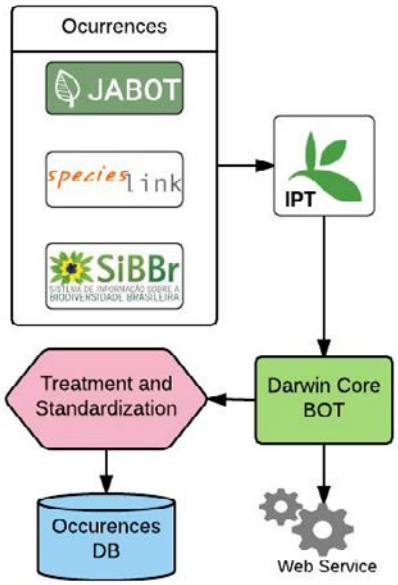

Figure 2. Darwin Core Bot system scheme

The spatial calculations and risk assessment are made through a web service without persistence (Stateless) of the Darwin Core Services tool, developed in Clojure. This system is responsible for receiving a set of events and perform the calculation and evaluation (Figure 3).

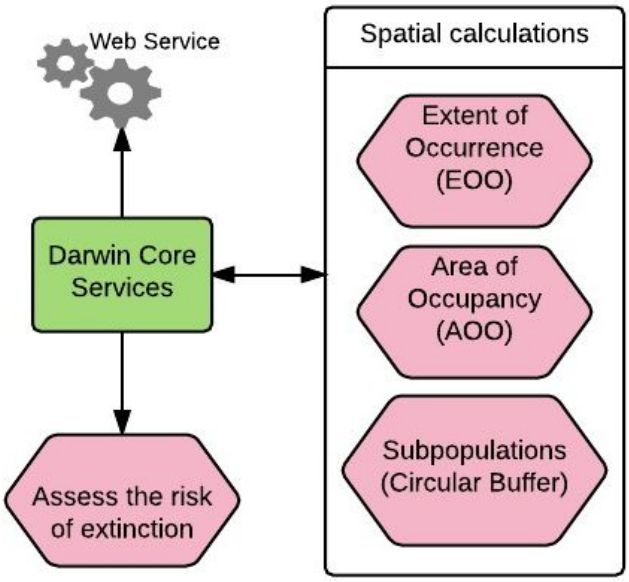

Figure 3. Darwin Core Services system scheme

Another system that is part of the tool is the Biodiv-idx. This system is responsible for consolidating the data of calculations and evaluations made possible by previous systems,

storing this data in a database and in the "ElasticSearch" search engine. This tool was also written using the Clojure programming language (Figure 4).

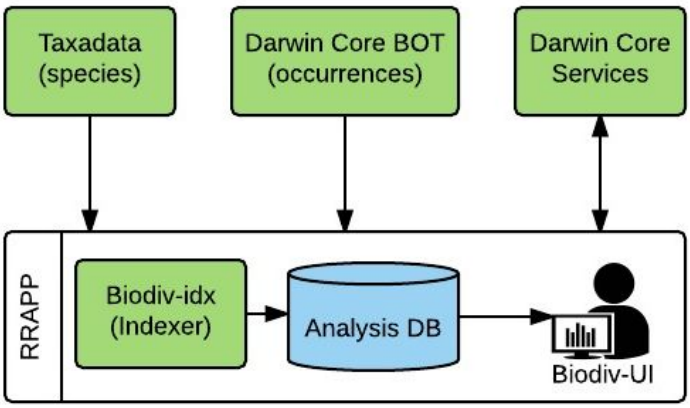

Figure 4. Figure showing the systems and the tool "Rapid Risk Assessment Application" - RRAPP

Finally, the Biodiv-UI system, developed with the PHP 7 language, provides a search interface and displays statistics on the calculations and risk assessments carried out (Figure 5).

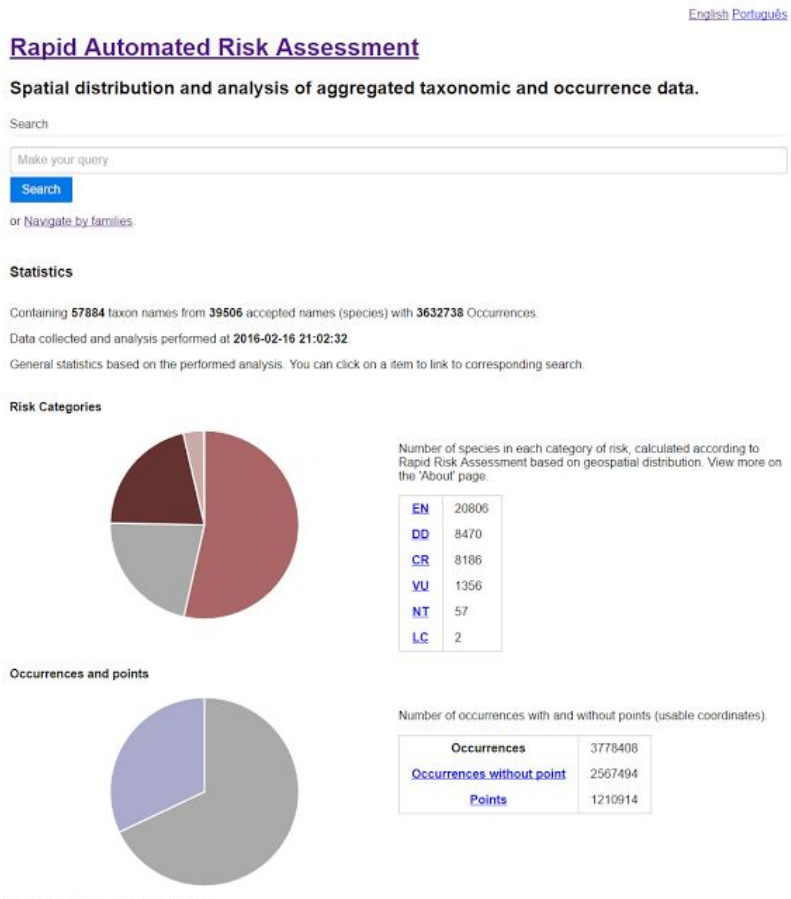

Figure 5. The web interface of the "RRAPP" Tool (partial)

All systems are deployed in a single virtual server with 4GB of RAM, 4 processing cores, and a 40GB SSD, where 30GB of which are allocated. The systems run in isolation and is kept in Docker containers.

3. Results

Regarding the performance of the systems, we recorded the following results:

- The creation of taxonomic set by Taxadata system totaling 57,882 names, was performed in 15 seconds;
- The composition of the occurrences set by the Darwin Core Bot system, totaling 3,632,660 occurrences was performed in 30 minutes;
- The calculation and assessment by Darwin Core Services and Biodiv-idx systems of a total of 39,506 species were conducted in 11 hours and 22 minutes.

Regarding the data analysis, the following results were recorded:

- The results of the analysis show that no occurrence records were found for 2,871 species (7.3%);
- 36,365 species (92%) have occurrence records;
- 3,907 (10.7%) species have up to 2 records;
- 8,578 species (21.7%) do not have occurrence records with valid coordinates;
- 6,363 species (16.1%) have only 1 or 2 records with valid coordinates.

These analyses also indicate that 8,470 species (21.4%) are "data deficient" (DD) and 30,348 species (76.8%) are in the category of "threat" (vulnerable, endangered or critically endangered).

The analyses also are available for the entire set of Brazilian flora species and also for each family and each species individually.

All technologies used are open source, free, and available for use, collaboration, and modification by third parties at https://github.com/diogok/biodiv-compose. The results are available on the portal at http://rrapp.jbrj.org or http://rrapp.cncflora.net.

4. Conclusion

The increasing availability of open data on biodiversity offered on internationally accepted standards enables the emergence of new analytical techniques that seek to explore and identify patterns that would not be apparent otherwise [Kelling et al. 2009]. These new techniques rely on implementing computing solutions capable of manipulating and analyzing a massive amount of data with a computational cost compatible with the scale of analysis proposed.

In this work, we demonstrate a tool that follows this new paradigm, to assist the achievement of the second goal of the Global Strategy for Plant Conservation - GSPC for 2020: An assessment of the conservation status of all known plant species, as the far as possible, to guide conservation action.

The analyses also show that the spatial and taxonomic qualification of scientific collections deposited in the herbaria is vital to the quality of the results of the analyses which use these sources of information.

As future work, we include: (i) the development of a "GBIF-bot", able to integrate the set of occurrences offered by the Global Biodiversity Information Facility - GBIF; (ii) the development of a data quality system in order to select more thoroughly the occurrences with geospatial coordinates; (iii) a comparison of the risk assessments made by the tool to the evaluations conducted by CNCFlora professionals, using the criterion B of the IUCN, so as to analyze the tool efficiency.

References

Beaman, R. S. and Cellinese, N. (jan 2012). Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys, v. 17, n. 209, p. 7–17.

Blagoderov, V., Kitching, I. J., Livermore, L., Simonsen, T. J. and Smith, V. S. (jan 2012). No specimen left behind: industrial scale digitization of natural history collections. ZooKeys, v. 146, n. 209, p. 133–46.

Flora do Brasil 2020. Jardim Botânico do Rio de Janeiro. Disponível em: < http://floradobrasil.jbrj.gov.br/ >. Acesso em: 29 Fev. 2016

Forzza R (2014): Brazilian Flora Checklist - Lista de Espécies da Flora do Brasil. v393.53. Instituto de Pesquisas Jardim Botanico do Rio de Janeiro. Dataset/Checklist. doi:10.15468/1mtkaw

Gadelha, L., Guimarães, P., Moura, A. M., et al. (2014). SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira. VIII Brazilian e-Science Workshop (BRESCI 2014). Proc. XXXIV Congress of the Brazilian Computer Society,

GSPC (Global Strategy for Plant Conservation). (2012). Global Strategy for Plant Conservation. Botanic Gardens Conservation International, Richmond, Inglaterra. 38p.

IUCN (International Union for Conservation of Nature and Natural Resources). (2001). The International Union for Conservation of Nature and Natural Resources. Guidelines for Application of IUCN Red List Criteria at Regional levels: Version 3.0.

IUCN. (2012). IUCN Red List Categories and Criteria: Version 3.1. Second edition. Gland, Switzerland and Cambridge, UK: IUCN. iv + 32pp.

Kelling, S., Hochachka, W. M., Fink, D., et al. (2009). Data-Intensive Science: A New Paradigm for Biodiversity Studies. BioScience, v. 59, n. 7, p. 613–620.

J. Lewis and M. Fowler (2014). Microservices. Disponível em http://martinfowler.com/articles/microservices.html. Acessado em 01/03/2016.

Mace G.M. & Lande, R. (1991). Assessing extinction threats: towards a re-evaluation of IUCN threatned species categories. Conservation Biology, n. 5, p. 148-157.

Martinelli G. & Moraes M.A. (2013) Livro Vermelho da Flora do Brasil. Andrea Jakobsson Estúdio: Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rio de Janeiro.

Rivers, M. C., Bachman, S. P., Meagher, T. R., Lughadha, E. N. e Brummitt, N. A. (2010). Subpopulations, locations and fragmentation: Applying IUCN red list criteria to herbarium specimen data. Biodiversity and Conservation, v. 19, n. 7, p. 2071–2085.

Robertson, T., Döring, M., Guralnick, R., et al. (6 ago 2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE, v. 9, n. 8, p. e102623.

SiBBr - Sistema De Informação Sobre a Biodiversidade Brasileira. Disponível em http://www.sibbr.gov.br/. Acessado em 26/02/2016.

Wieczorek, J. R., Bloom, D., Guralnick, R. P., et al. (6 jan 2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE, v. 7, n. 1, p. e29715.