Species Checklist Cross-mapping and Precursor National Checklists Generation from GBIF-mediated Data

David Remsen[1], Wouter Addink[2,3]*

[1] Marine Biological Laboratory, Woods Hole, MA, USA

[2] Species 2000, Leiden, Netherlands

[3] Naturalis Biodiversity Center, Leiden, The Netherlands

* Corresponding author

E-mail: wouter.addink@naturalis.nl

Abstract

In a hackathon, three teams worked together to investigate the workflow needed for establishing or improving national species checklists. We investigated existing open-source tools and services, improved them where needed and created new open source solutions to establish a workflow for national species checklist creation and curation. When the components are further developed and suggested improvements are implemented, the tooling will be in place to establish an internationally standardized, common workflow.

Introduction

Regional checklists of species provide a consolidated catalog of biodiversity related to a geographic area and are important tools for collecting and managing biodiversity data. Without complete national checklists, for example, countries cannot effectively implement national scientific biodiversity networks [1,2]. A common approach for establishing national checklists does not yet exist. Establishing a common approach would reduce duplication of work and could also introduce interoperability between national checklists in international scientific biodiversity networks . With its global and comprehensive scope, the Catalogue of Life (CoL)[2,3], when complete, would, in principle, include any species referenced in any national species checklist. With over 1.6M species already in the Catalogue, we sought to investigate how it might compare to existing national lists in coverage and how it might be used, in combination with other sources, such as the Global Biodiversity Information Facility (GBIF) [4], to support the discovery of missing species candidates to such lists [5]. The GBIF portal provides access to over 500 million biodiversity data records, from over 13,000 different source databases and many of these contain geographic information that may be linked to a country. Our goal was to explore whether such data might be used to locate species missing from existing regional/national checklists or as the basis for generating new lists. In the form of a hackathon we explored issues related to the creation, maintenance and interoperability of national species checklists, both among each other and with the Catalogue of Life.

Materials and Methods

A hackathon was organised by Species 2000 as part of the GBIF Capacity Enhancement Support Program [6] and hosted at Naturalis in Leiden, The Netherlands, from 2-5 March, 2015. A total of 20 participants attended the event, representing 10 different countries and multiple institutions, all involved in creating national checklists, managing national checklists, or developing tools for national checklists.

This broad topic allowed us to divide the group into three teams focusing on sub-topics:

1. **Team 1 - Cross-mapping** refers to the comparison of a target list to a reference list with the goal of identifying overlapping and distinct taxa within each list [7]. This group focused on the normalization of taxon names for comparison and the development of a confidence score in matches.
2. **Team 2** - **Annotations** refers to a system that allows curators to annotate checklist cross-mappings with GBIF occurrence data to provide feedback on the presence of legitimate or suspect species occurrences within a country[8].
3. **Team 3 - Distributions** of species, sought solutions to gather evidence that a species occurs in a country to support accurate annotations.

Each team was composed of 4-5 persons with each individual assigned roles to support the objectives of the group. This included, at least, a team leader, a senior programmer and a reporter. For capacity-building reasons, people could also have a junior programmer role, to learn from the other programmers in the team. Others identified as domain-experts and moved between groups to share expertise. Further roles included tester and data analyst.

The teams worked in separate rooms with an initial task to define the user stories that would form the basis of their work. Work proceeded for the remainder of the meeting with regular meet-ups and discussions at the end of the each workday. Code and documentation was stored in a GitHub repository and open source licences were added.

The following user story formed the basis of work for the Cross-mapping team (team 1):

User Story 1-1

As an owner of a checklist I want to know whether I am missing species for my country that are reported to occur there by other authoritative sources.

Goals

1. Assess how names in national checklists correspond to names in CoL - quantity and quality of positives and negative matches resulting from current parsing and naming tools.
2. Use these examples to assess and refine the Global Names Architecture name-parsing and matching algorithms (http://gni.globalnames.org/parsers/new)
3. Provide recommendations for future checklist cross-mapping tools and interfaces

The team implemented a system in Ruby programming language that uses Global Name (GN) webservices (http://resolver.globalnames.org/data_sources.json, http://resolver.globalnames.org/data_sources) to identify matching taxon names in a target list (a national checklist) against a reference list. CoL was used as reference list containing valid taxon names and synonyms. Four candidate target lists were used to test the name-parsing and matching services (Table 1). Test results were used to improve the GN webservice.

Table 1. National species lists used to test cross-mapping tools.

List	Number of Taxa
Portuguese Bryophyte	726 species
Portuguese Tracheophyta	4,458 species
Slovakian Checklist	1,248 species
Norwegian Checklist	10,950 species

Team 2 focused on a system design that could be used to present assertions of a taxon occurrence within a country - to a presumed expert curator, who might then use their knowledge to assess the assertion and determine whether the taxon should or should not be added to the list. The authoritative Catalogue of Life record - linked through the cross-mapping efforts of Team 1, would then form the record-of-authority for the national list. In addition, negative matches (i.e, species asserted to occur within the country but determined to not belong there - might be linked to a comment or annotation that could serve to inform future users of the GBIF network to the nature of the suspect occurrence. This led to the articulation of the following user story.

User Story 2-1

> As an owner of a national checklist I want to load my checklist into a system and compare it to the list of taxa assigned to my country within the GBIF index. Matches missing from my national list may 1) represent legitimate missing taxa that should be candidates to add to my list. They may also 2) represent taxa erroneously applied to my country that should be annotated with their suspect status for future users of the record.

Goals

1. Can the federated GBIF portal be used to support the identification and qualification of novel species occurrence records in the development of national or regional species inventories?
2. Can annotation interfaces be used, in combination with authoritative regional or national species lists, to identify and annotate potentially erroneous species occurrences and thus inform future users of GBIF-mobilized data as to this erroneous assessment?