Report to GBIF, June 2015

Results of Training Hackathon for Checklist Cross-mapping and Precursor National Checklists Generation from GBIF-mediated data

Authors: Dave Remsen, Wouter Addink, Matúš Kempa, Andreas Kohlbecker, Wouter Koch, Dag Endresen, Rui Figueira, Oskar Kindvall, Nabil Youdjou, Marie-Elise Lecoq, Sophie Pamerlon, Ayco Holleman, Ruud Altenburg, Maarten Schermer, Dmitry Mozzherin, Toril Loennechen Moen

Introduction

Species 2000 organized a hackathon event hosted at Naturalis in Leiden, NL, from 2-5 March, 2015. The event was funded by the Global Biodiversity Information Facility (GBIF) as part of the Capacity Enhancement Support Program (CESP). Goals for the event were:

1. To promote capacity-building amongst GBIF Nodes and curators of national species in Europe.
2. Explore methods to support the development of tools and workflows for creating proto-national checklists or to enhance existing checklists using external data sources and services from the GBIF network, the Catalogue of Life, and other sources.

A total of 20 participants attended the event representing 10 different countries and multiple institutions. Many participants made presentations to introduce their interests and expertise. The full list of participants and their presentations are available here.

Goal of hackathon

The focus on the hackathon was to explore issues related to the creation, maintenance and interoperability of national species checklists, both among each other and with the Catalogue of Life (referred to throughout as *CoL*). With it's global and comprehensive scope, the Catalogue of Life, when complete, would, in principle, include any species referenced in any national species checklist. With nearly 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 GBIF, to support the discovery of missing species candidates to such lists. We sought to address the question, "Can a complete and authoritative Catalogue of Life be used to provide and maintain taxonomic authority records for satellite national lists?

From the original proposal:

> *To create a mechanism through which each checklist can be derived from a common component. This would make national checklists interoperable. This can be through the development of a common nomenclatural framework (tied to the original description of species). The framework should allow people to compare taxonomies and to determine the nomenclatural status of a name. Once determined, it should be reusable so others do not have to do this comparison again.*

This broad topic allowed us to divide the group into three teams focused 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. 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.
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 analyzer.

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. See Appendix 1 for links to these repositories, prototype web applications, and the team notes that form the basis for this report.

The following sections provide a summary of the work and results of each of the three teams.

Team 1 - Cross-mapping

Team Members: Dmitry Mozzherin (leader), Matúš Kempa, Rui Figueira, Wouter Koch, Toril Loennechen Moen

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

User Story 1-1

> *As an owner of a checklist I want to crossmap my names against the CoL and other sources as a comparison framework. I want to know whether I am missing species for my country that are reported to occur there by these 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 GNA name-parsing and matching algorithms
3. Provide recommendations for future checklist cross-mapping tools and interfaces

Results - User Story 1-1

The group compared the Dutch national checklist to a species list derived from occurrence records linked to the Netherlands obtained from the GBIF network. Figure 1 illustrates the cross-mapping process. A target national list is compared to the Catalogue of Life within a system that uses Global Name (GN) services to identify matching taxon names in the national list to as valid taxon names or linked synonyms in the CoL. The GN services use advanced name-parsing and matching tools to compare taxon names and is able to identify matching names that may vary slightly in their form, spelling of authorship etc. The output of this process are three sets of taxa:

1. Matching taxa are those contained in both lists
2. Taxa missing from the national list but contained in the CoL. Given the global scope of the CoL, we will expect most CoL to not occur within a given country. Access to distribution data within the CoL (the work of Team 3) would help identify relevant CoL missing from the national list.
3. Taxa (or possible synonyms) missing from the CoL. This list could have potential value as candidate additions to the CoL.

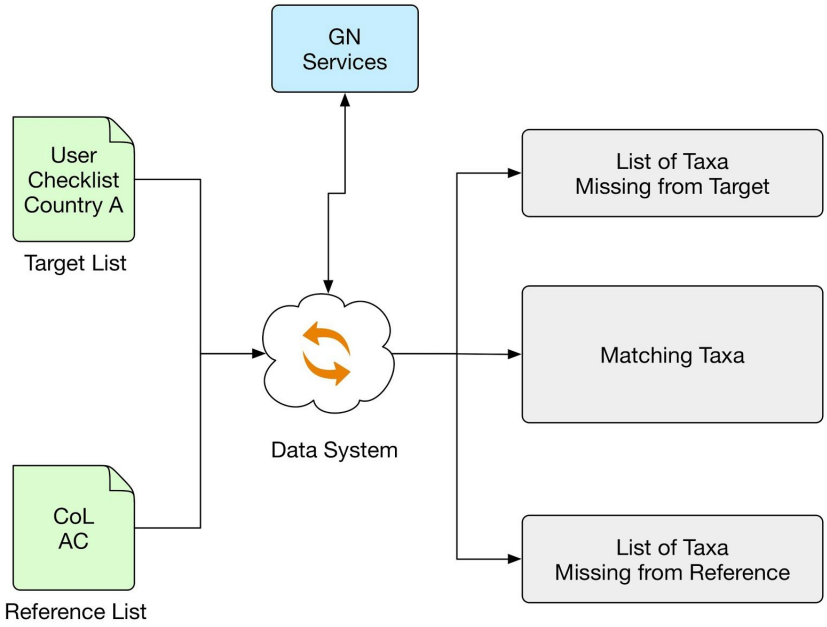


Figure 1 - Cross-mapping group work process

The primary goal of the matching exercise focuses on assessing the capability of the name-matching system to accurately match names so that the curated CoL taxa can be used, via reference, as the core concepts in one or more national lists, hence promoting interoperability.

Four candidate target lists were used to test the cross-mapping services.

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

Table 2 - National species lists used to test cross-mapping tools and their score.

When the prototype system is run, the cross-mapping tool analyzes the two checklists and produces a tabular output showing the results of the cross-mapping. A name-parsing tool splits each scientific name into

its constituent components (genus, species, authorship, year, rank, etc.) using Backus Naur grammar to establish rules regarding the composition of names. This allows the component parts to be identified for comparison using various “fuzzy” text-matching algorithms to estimate when two similar text strings may be matches. The process ends with a matching estimate that is given a match type and a confidence score.

For example, two entries from the compared lists form the following match record with a score of 0.75, indicated an exact match of the canonical forms of the name but non-matching (but similar) author names.

Name 1	**Name 2**	**Score**
Scenedesmus abundans (Kirchner) Chodat	Scenedesmus abundans (O. Kirchner) Chodat	0.75

Matching score options include:

Match Type	**Score**
Exact Match	0.999
Partial Canonical Form Exact Match	0.998
Canonical Form Exact Match, similar authorship	0.995
Canonical Name Only Exact Match	0.75
Genus part matches	0.75

Results

Outputs for all test tables are available on the Team 1 repository. An example, using the Portuguese Bryophyta list is available here.

A summary of the gross cross-mapping results are summarized below.

List	**Matches**	**% matched**	**unmatched species/variety/subspecies**
Portuguese Bryophyta	726/**745**	97.4%	19/0/0
Portuguese Tracheophyta	4458/**4618**	96.5%	121/26/11
Slovakian Checklist	1248/**1279**	97.6%	26/5/0
Norwegian Checklist	10950/**11393**	96.1%	163/2/7

These results indicate a relatively high proportion of matches, which would be predicted if we could assume a relatively comprehensive Catalogue of Life and accurate matching capabilities of the cross-mapping tools. A detailed look at the matching results, however, reveals some degree in ambiguity in the relevance of the matches.

Relevance in matching can be divided into two measures: precision and recall. The cross-mapping tool appears to provide very good recall but the precision (the proportion of false positive matches to true positive) is a measure that should be more thoroughly explored.

The confidence score provides a starting point in assessing precision but a thorough analysis of these scores was not performed during the workshop. A general review would indicate that high confidence scores over 0.995 were likely to be positive matches. Scores less than this could be either true or false positives. For example:

Name 1	**Name 2**	**Score**
Anacolia webbii (Mont.) Schimp.	Anacolia webbii W. P. Schimper, 1876	0.75

This example illustrates a positive match with a lower confidence due to the variation in the authorship, which a domain expert would identify as referring to the same authorship instance.

On the other hand, an equivalent score is given to the following:

Name 1	**Name 2**	**Score**
Campylium stellatum (Hedw.) Lange & C.E.O.Jensen	Campylium	0.75

This example is clearly not an exact canonical match but a more general match between a species and its parent genus.

Another example of an ambiguous match is related to what appears to be non-standardized use of authorship when citing a scientific name. For example, the names

Name 1	**Name 2**	**Score**
Cratoneuron filicinum (Hedw.) Spruce	Cratoneuron filicinum (Fiorini-Mazzanti) Latzel, 1931	0.75

These names could refer to homonyms or to one or more *chresonyms*, where the authorship represents a citation of a particular publication that is not one of the code-regulated citations used to generate a code-compliant nomenclatural act. In these cases, the ambiguity is entirely related to the content and cannot be evaluated by parser rules alone.

Summary of Results

The summary of the results of Team 1 indicates that two species checklists can be matched with a very high degree of recall such that most true positive matches will be identified. More work should be done in refining precision however, particularly in the refinement and assignment of scoring and the fuzzy matching of authorship. Most low-scoring canonical matches appear to be true positives, and thus, separating and

scoring these higher than non-matching authorship (which indicate homonyms or chresonyms) would provide higher and more relevant scoring.

Team 2 - Annotations

Andrea Kohlbecker (leader), Ruud Altenburg, Oskar Kindvall, David Remsen

Sources of biodiversity occurrence data, such as catalogued and indexed by GBIF, may serve as a means to both verify or extend the list of taxa found in national species checklists. They might also serve as the means to start a de-novo national species list. 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?

System Design

Team 2 came up with the following solutions for each step in the workflow described in Figure 2 below.

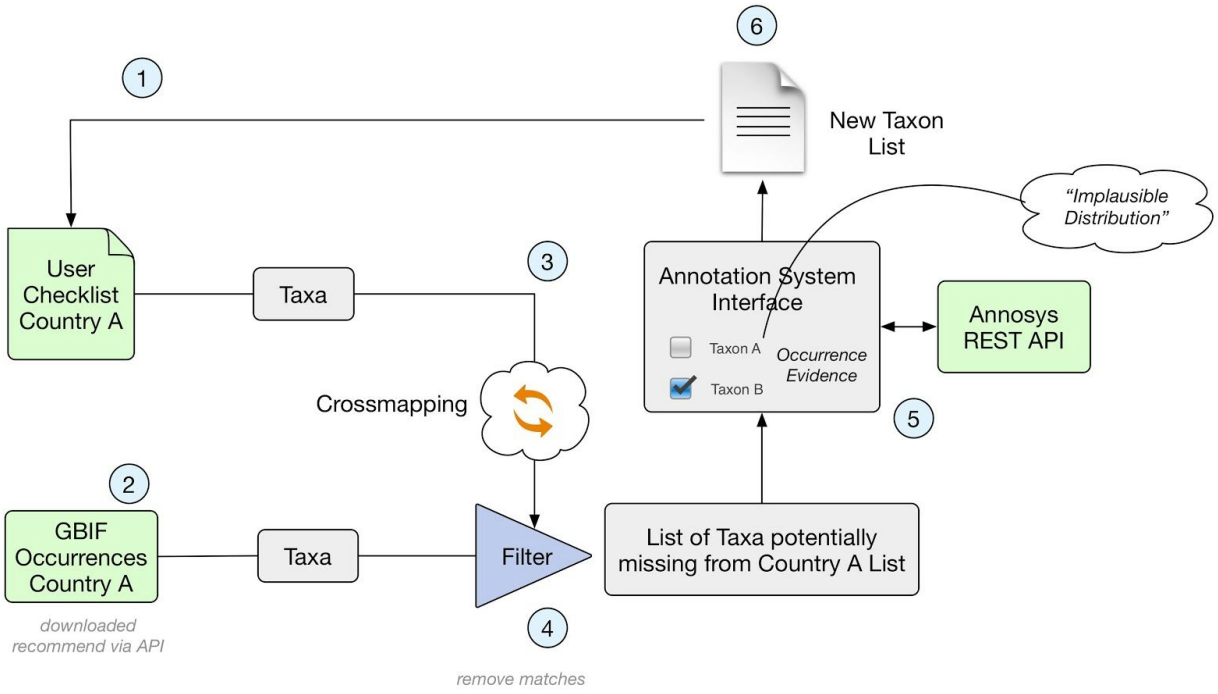


Figure 2 - *Workflow describing the steps needed to enable comparison between a national checklist and the taxa represented by GBIF occurrence data.*

1.
2. GBIF. The term taxonKey appeared to be a more solid choice.

   To retrieve the list of taxa represented by occurrence data, the team used SQL distinct selection on speciesKey, scientificName, genus; specificEpithet, infraspecificEpithet. This list was stored in the database (n = 43 387).

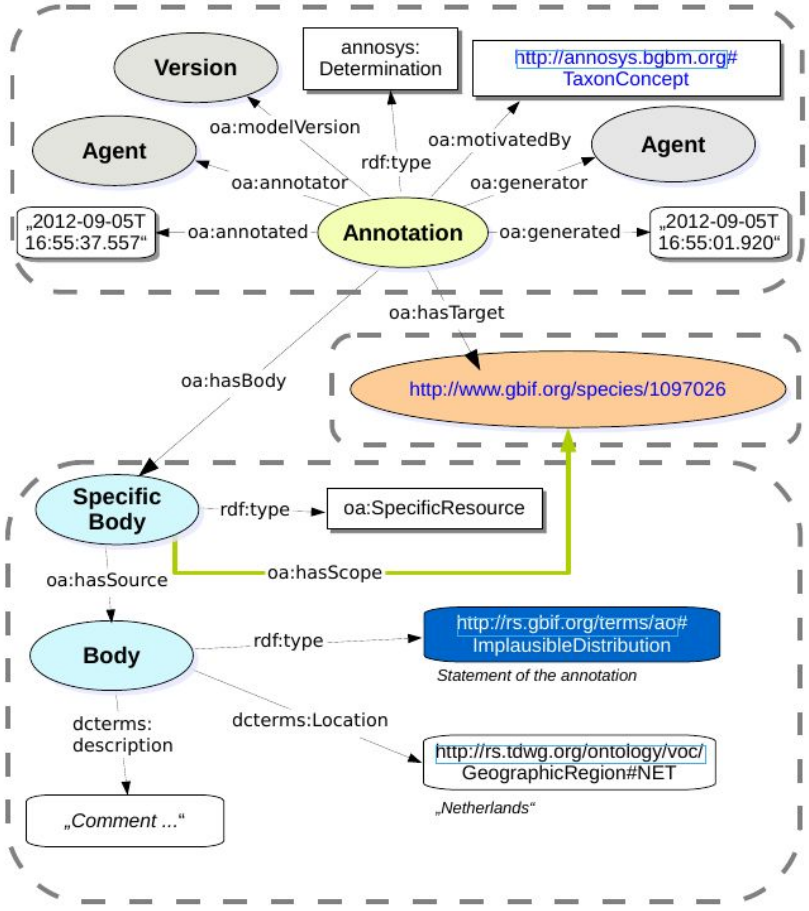
3. Cross-matching the checklist with the Catalogue of Life
4. Filtering out the negative matches (the taxa missing from the original national list)

   A table was created where all taxa represented by GBIF data was inserted. This table included the following columns: taxonKey, AnnosSysUri, scientificName, blacklisted (bool), taxonStatusGBIF, existsInChecklist, checklistStatus (native, introduced etc), occurrenceRecordCount. The fields existsInChecklist, checklistStatus were updated from the Checklist table. Extraction of the potentially missing taxa was then made by selecting which taxa in the table existsInChecklist is false.

5. Annotate the missing taxa.
   a. The team used AnnoSys (https://annosys.bgbm.fu-berlin.de/) to store annotations to the taxon occurrence records. AnnoSys was originally intended to annotate biodiversity occurrence records in ABCD; an XML format. The team extended an Annotation class of AnnoSys so it could handle information about a taxon (as opposed to a taxon occurrence). The team further developed a new annotation model, which is also based on the W3C Open Annotation Data Model (http://www.openannotation.org/spec/core/). General techical documentation and documentation of the open annotation model as used by AnnoSys can be found at http://wiki.bgbm.org/annosys/index.php?title=TechnicalDocumentation

The purpose of the annotation in this case was to express that the distribution for the taxon might, or might not, be correct. In order to express the latter, a interim RDF term (http://rs.gbif.org/terms/ao#ImplausibleDistribution) was introduced.

b. The validation information is then supposed to be posted into AnnoSys using its REST API. We suggest that the annotation should be related to the URL representing the taxon page of GBIF i.e. http://www.gbif.org/species/taxonKey.The annotation should be expressed in a way that should be interpreted as: for the taxon with the taxonKey, all occurrences reported for the specified Country where the establishmentMeans do not clearly indicate non natural occurrence, should be considered as beeing expected errors.



c. Figure 3 **Example PUT request** to the AnnoSys to create a new taxon annotation:

6. Filter out legitimate missing species candidates for addition to the national list.

The resulting GUI is shown below. It supports annotation of all listed taxa which according to GBIF occurrence data is likely to candidates to add on to the existing national checklist. For each row a taxon-specific link to the distribution map generated by GBIF is provided in order to help the user evaluate the underlying data and judge whether or not the taxon is likely to exist in the country.

To annotate a taxon the user has to use the checkbox. When checked, a taxon-specific annotation is posted

at the AnnoSys repository. When this is done the application posts the response with the URL to the new annotation in the system database table.

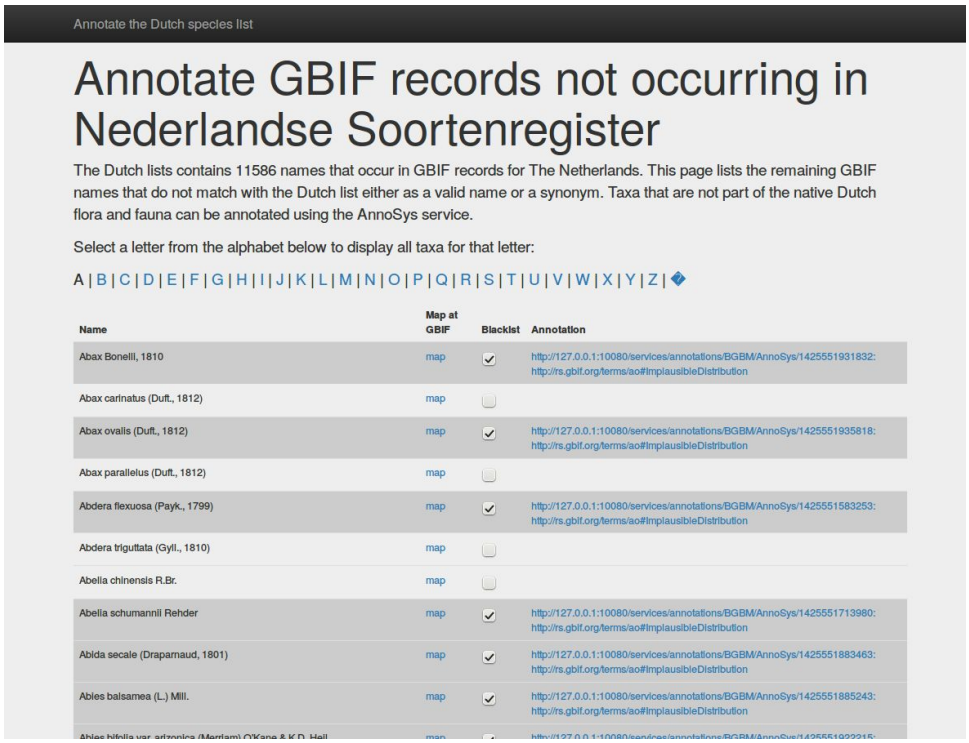


Figure 2. Screen shot of the Annotation GUI with the list taxa potentially missing in the checklist of the Netherlands.

Recommendations

As a result of the explorations the team identified an additional user story worthy of further elaboration:

User story 2-2

> As a curator of GBIF data and harvest processes, I want to be able to look up annotations indicating that imports of occurrence data with a specified combination Country and Taxon may be incorrect in order to handle these records in an appropriate way (e.g., records of lions occurring within the Netherlands and not indicating zoo specimens). They may lead to more accurate assessment of species composition within countries.

Suggested improvements of existing services

In order to make it more feasible to implement a system supporting the User Story investigated here, we can recognize the need for at least two major improvements of existing web services.

1. **GBIF occurrence service**

   a. A method that can utilize the same set of input parameters for querying occurrence data as the existing that delivers the result as a list of represented (observed) taxa with all relevant DwC terms plus the various GBIF taxonKeys. Relevant could be adding a column for number of occurrence records per taxon listed.

2. **AnnoSys API**
   a. Adding functionality that supports a more generic solution for annotations of information.

Improvements of the preliminary solution

The list of potential missing taxa was very long (about 30 000 records). This may partially be explained by the type of matching: full scientific names including authorship were used. Any spelling variations in the names of authors would result in non-matching names. Furthermore, the GBIF data was not tested for semi-invalid data, such as missing lat/lon data (which can be used to verify the location), and an incorrect country label (data provided in the GBIF DwCA file). After extensive pruning, some values can be used to score taxa: basisOfRecord, establishmentMeans, type specimen, number of occurrences etc. The resulting list could be ordered by this score.

A link to the taxon page at GBIF would be useful in the interactive interface in order to support evaluation of the taxon likelihood of actually being missed in the checklist. That could be if the overall distribution pattern for that taxon suggest that the natural distribution do not include the target country.

Demo version:http://134.213.149.111/group_2/

Team 3 - Distribution Evidence

Dag Endresen (leader), Marie-Elise Lecoq, Sophie Pamerlon, Maarten Schermer, Nabil Youdjou

User Story 3-1
I am the manager of a national species checklist and I am looking to verify the occurrence of new species candidates for my list.

User Story 3-2
I am the manager of a national species checklist and I am looking to verify the occurrence of existing species candidates in my list.

Goals

1. Analyze the existing Catalogue of Life and PESI as potential sources for authoritative assertions for regional/national occurrence for species.
2. Analyze the GBIF-mobilized data as a source for documented species candidates for a national checklist.

Introduction

The cumulative efforts of Teams 1 and 2 established the means to compare taxa listed in a national species checklist and the aggregate list of species assigned to the same country within the GBIF index of occurrence records. Recognizing that this latter index may contain evidence for species that should be in a national list, as well as the inevitable errors, an interface was developed that allow a reviewer to provide review and annotate the GBIF records. The objective of Team 3 was to extend this process by gathering evidence for the occurrence of a candidate species within a given country. Other data sources provide this evidence through various service interfaces. Sources include:

- Catalogue of Life: CoL provides information on country level species occurrence as asserted by the custodians of the taxonomic data sectors. A full review of these data was conducted and the results are available.
- Pan-European Species-directories Infrastructure (PESI): Provides verified occurrence information for European countries.
- GBIF Checklist Bank - the GBIF checklist index provides documented species checklists that may contain national occurrence information.
- GBIF occurrence data - the GBIF occurrence data that is used to provide the summary list of species contains additional data that may provide supporting evidence for determining the status of a candidate species. This includes the following Darwin Core data properties.
  * Basis of record: Observation, MaterialSample/Specimen, Living,
  * Establishment means: "native", "introduced", "naturalised", "invasive", "managed"
  * Reported occurring in respective country the latest 10 years, 50 years, 100 years, ...

Species occurrences derived from the GBIF index provided a starting point. Investigated was whether other sources of authority can be used to verify the occurrence. For example, Catalog of Life and PESI provide authoritative assertions of national occurrence of species. Team 3 gathered these data and split them into a set of dimensions such as basisOfRecord, age of occurrence (or last seen), establishment means. There were parsed into a table and presented to the curator of the potential national checklist for further analysis.

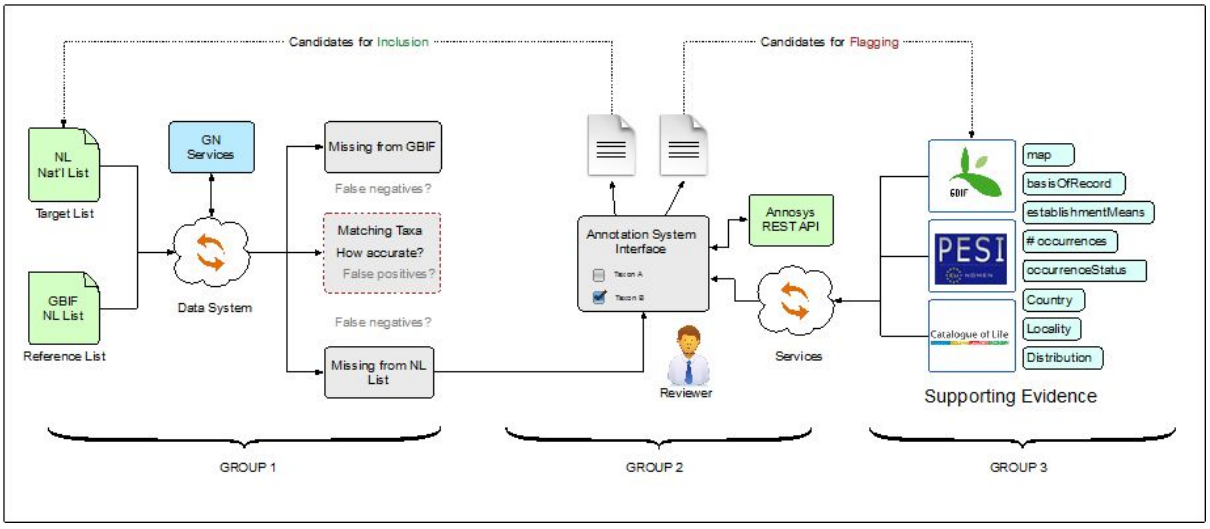
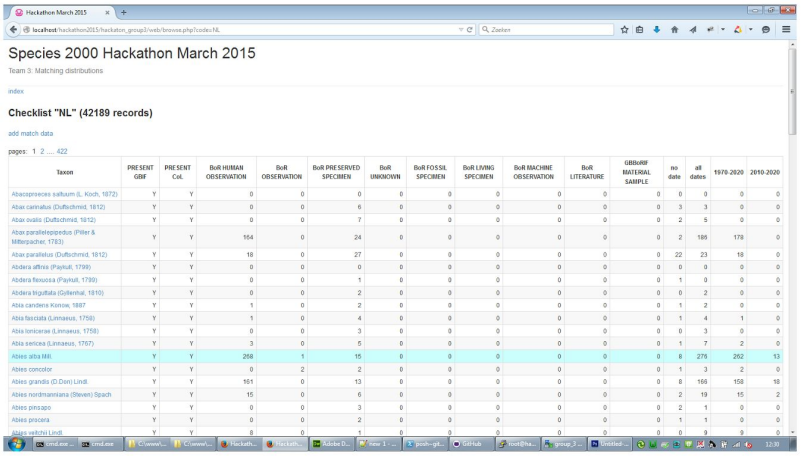


Figure 3 illustrates the relationship between the 3 teams with the efforts of Team 3 providing evidence to validate national occurrence data mobilized through the GBIF network.

Process

The process for providing supporting national distribution evidence of GBIF species occurrence data proceeded as follows:

1. Obtain list of taxon names for target country (Output of Group 1)
2. Retrieve taxon keys from respective working checklists -
    a. For the source national lists either via the source data or through API from checklist source.
    b. Retrieve GBIF taxon key via GBIF API
        i. http://api.gbif.org/v1/species/match?name=[taxon_name]
    c. Retrieve global checklist taxon keys for the Catalogue of Life
    d. Retrieve global checklist taxon keys for PESI
3. Retrieve country-level distribution information from the source checklists through their respective API's via taxon key identifiers
4. Retrieve occurrence counts from raw occurrence data in GBIF
    a. Example "Abax ovalis (Duftschmid, 1812)", taxonKey=5754770:
    b. http://api.gbif.org/v1/occurrence/search?taxonKey=5754770 => count=588
    c. http://api.gbif.org/v1/occurrence/search?taxonKey=5754770&country=NL => count=7
5. Build a front-end to present the summarized evidence.
    a. http://134.213.149.111/group_3/hackathon_group3/web/index.php



b.

c. Figure 1: Example national checklist of the Netherlands annotated with evidence in support of species occurrence in the Netherlands.

Results

Available services were able to be integrated into the Annotation system interface in order to verify the overall concept. The various sources of supporting evidence do provide data that can help a reviewer in determining whether a single or set of GBIF occurrence records represent a legitimate instance of a species for a country that should be added to the national list. A number of observations regarding the existing interface form the basis for feature requests to GBIF.

Request 1 (GBIF)

It takes two API calls to obtain national distribution information for a taxon.

1. Input a species name to obtain an identifier that can be used to obtain distributions (referred to as a usageKey)
    a. Example “Abax ovalis (Duftschmid, 1812)”:
    b. http://api.gbif.org/v1/species/match?name=Abax%20ovalis%20%28Duftschmid,%201812%29
    c. => usageKey=5754770 (aka GBIF taxonKey)
2. Utilize the usage key to obtain a list of distributions.
    a. http://api.gbif.org/v1/species/5754770/distributions => list of “localities” (aka countries)

Current output only shows the source taxon key. It would be useful to include a source dataset key with the distribution data.

Request 2 (GBIF)

It would be useful to add a datasetKey as a search parameter e.g., http://api.gbif.org/v1/species/5754770/distributions?datasetKey=90d9e8a6-0ce1-472d-b682-3451095dbc5a

Request 3 (GBIF)

Also very useful might be to add search parameter country (as interpreted by GBIF) and to get a TRUE or FALSE response only:
http://api.gbif.org/v1/species/5754770/distributions?country=NL => T/F

Request 4 (GBIF)

Given a national checklist with taxa occurring in a country, we are interested to check the GBIF occurrences if there are additional taxa observed as occurring in the given country. For this operation we would propose an inventory service from the GBIF occurrence API.
http://www.gbif.org/developer/occurrence#inventories
http://api.gbif.org/v1/occurrence/counts/species/?country=NL => taxonKey(s)

Request 5 (COL)

Given a taxon identifier, we would like to check the CoL for national distribution data in a consistent and standardized format. Current CoL distribution data is inconsistent across taxonomic sectors and heterogeneous in format. We performed a detailed analysis of CoL distribution data conducted during the hackathon. See the resulting report, Distribution Data in the Catalogue of Life and the associated summary of distribution data across Global Species datasets.

Request 6 (COL)

Given a country identifier, consider a service that provides a list of matching taxon identifiers.

Request 7 (PESI)

Access to PESI, via LSID, is inconsistent across source datasets. To be precise: ERMS, Fauna Europaea and the Index Fungorum provide an LSID and the Euro+Med Plantbase does not, it provides a GUID instead.

Further actions

The hackathon made a start with exploring technical aspects that could lead to a larger national-checklist strategy. For this, a project proposal would be needed to address technical, as well as procedural/collaborative aspects. This has the attention of the European Nodes. Actions that have been discussed to be carried out after the hackathon:

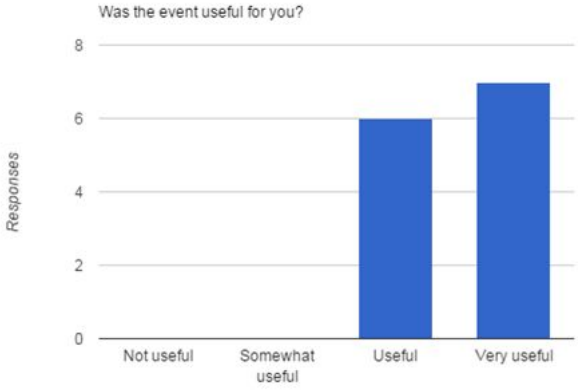
- Write a report for the GBIF newsletter, GBits
- Propose the subject of cross-mapping for a Google Summer of Code (Dima)
- Report the results in the next European GBIF Nodes meeting (Wouter)
- Report the results to the Catalogue of Life global team (Wouter & David)
- Create an interest group on the GBIF communications portal (Wouter)
- Include recommendations in this report and send it to GBIF and the European Nodes

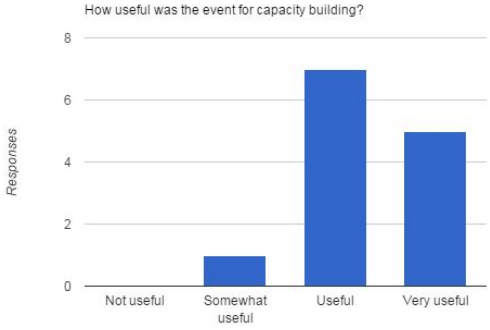
Status (19 june): Almost all actions have already been carried out, the last one is in progress.
The preliminary code of the cross-mapping group is formally released now as a command line application, and a web-interface is being worked on. See:
http://globalnamesarchitecture.github.io/crossmap/gna/2015/05/31/gn-crossmap-0-1-6.html

Evaluation results

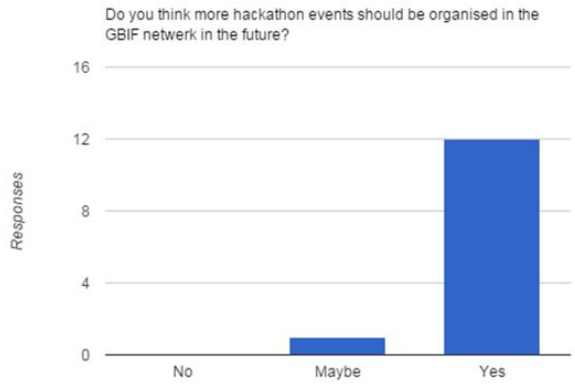
After the hackathon, an anonymous evaluation form was sent to the participants to get feedback (node participants only, not the trainers) about the event. All participants replied.

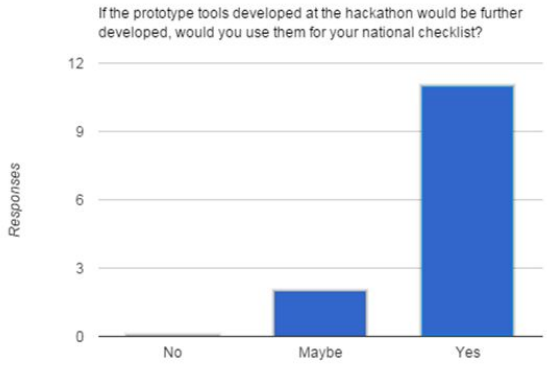

What is your overall rating for the event?
Responses
8
6
4
2
0
Poor
Fair
Average
Good
Excellent

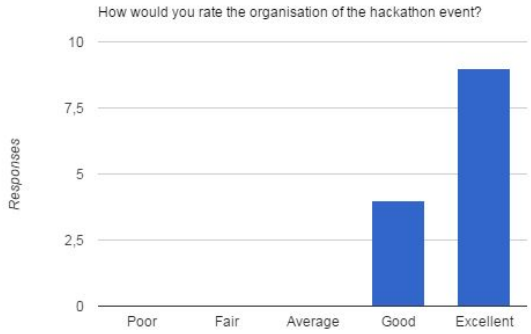

Was the event useful for you?
Responses
8
6
4
2
0
Not useful
Somewhat useful
Useful
Very useful


How useful was the event for capacity building?
Responses
8
6
4
2
0
Not useful
Somewhat useful
Useful
Very useful


Did you learn new things at the event?
Responses
10
7,5
5
2,5
0
Nothing
A few
A number of things
A lot







Suggested topics for a next hackathon were:

- •GBIF name/concept resolution
- •Data quality/data quality improvement & routines
- •GBIF API
- •Persistent identifiers
- •Annotating biodiversity data
- •Checklist hackathon part 2

Suggested topic for further training were:

- •Name evaluation by checklists managers
- •Improving, refining and completing checklists
- •Quality checks and data quality feedback to the owner of a checklist
- •Identifying candidate missing taxa within borders, black-listing
- •Cross-mapping different checklists
- •Annotating taxa
- •Increasing the inter-linkability of data, data-sharing
- •Thematic backbone taxonomy

Other input received:
“I heard one colleague saying it was the best meeting he had attended in some years. He was used to high-level LifeWatch meetings and felt this meeting brought the right group together and was focused and productive.”

“I was very happy for this week. It was interesting to see how well we discussed things and how we collaborated in a creative way.”

All respondents answered “yes” to the questions “Do you plan to stay in contact with colleagues you met at the event?” and “Are you interested to participate in follow-up actions?”.

Lessons learned for a next hackathon, as suggested by the participants:

- If a hackathon would be organized in a more intimate, relaxed atmosphere like a rented house, participants would be able to work together well into the evenings.
- A second Checklist Hackathon which bases on the results and experiences of the first one could lead to more mature tools. Also would it be great to focus during this second round more on cross mapping of taxa than only on names. Cross-mapping of taxa could not really be covered in the Hacking sessions due to lack of time.
- Separate brainstorm phase from tools design and development phase.

Appendix 1 - Links to team notes, repositories and prototype web applications.

A demonstration server was created to install prototypes created for demonstration purposes: http://134.213.149.111. On the last day the results were demonstrated and further steps were discussed.

Team 1 (Cross-mapping)	https://github.com/Sp2000/hackathon_group1
Team 2 (Annotations)	https://github.com/Sp2000/hackathon_group2
Team 3 (Distributions)	https://github.com/Sp2000/hackathon_group3

GitHub repositories for each of the three teams.

Each team drafted their own team report.

Team 1 (Cross-mapping)	https://docs.google.com/document/d/1I_MSlIE_js03CHKMi5-1xbnZvRY0kyJwIjP2P4rGrQk
Team 2 (Annotations)	https://docs.google.com/document/d/108MVZ9oLq6rdJgdZZ0sN38sMhIc0AIrg2an_T8_xJug
Team 3 (Distributions)	https://docs.google.com/a/naturalis.nl/document/d/1Gg5qCenUw9MEJPKS8_3XSTvvkEqIIkgFkq2Bhr4WYR4

Catalogue of life distribution investigations:
https://drive.google.com/open?id=0B-WDk-H2QjRKcFktSERZd0NHRUE
https://drive.google.com/open?id=0B-WDk-H2QjRKcnpHZmZjR2NhOHM

Appendix II - Summary of review of existing cross-mapping tools

A review of existing cross-mapping tools resulted in the identification of eleven different services, many of them consisting of overlapping and common core components. Given the expertise among the group, the GN resolver formed the basis of the work.

Datasystem	**Geographical scope**	**Coverage**	**Fuzzy matching?**	**Technology (web service?)**	**Has Hierarchies**
www.eu-nomen.eu	Europe	ERMS (marine) FaEu (land/freshwater) IF (fungi) E+M (plants)	Tony Rees Algorithm + name parser	-web application -soap (#50 names)	yes
www.marinespecies.org (Worms)	global	marine	Tony Rees Algorithm + name parser	-web application -soap (#50 names)	yes
iPlant TNRS	global	Plant	Tony Rees Algorithm + name parser	- web application - rest API	no
GN resolver	global	all	Tony Rees Algorithm + name parser	- library - web application - rest API (1000 names)	yes
NBIC	Norway	All (except bacteria and viri)	Yes	Web application API CSV export	yes
GBIF	global	all	yes	web service API	yes
CoL	global	all		web service	yes
i4Life	global	all	no	web service	yes
Taxonomic Tree Tool (TTT)	China	all			yes
Dyntaxa	Sweden	all	no?	web application	yes
EU Bon	EU	all	rely on other systems (pesi, worms, col …)		

Table 3 - Results of survey of existing cross-mapping tools and services