

Training hackathon for Checklist Crossmapping and Precursor National Checklists Generation from GBIF data FINAL ACTIVITY REPORT

Contents

1. Executive summary.....	1
2. Contact information.....	1
3. Project summary.....	1
4. Project objectives.....	5
5. Project deliverables.....	6
6. Evaluation: findings and conclusions.....	6
7. Recommendations and lessons learned.....	9
8. Future plans.....	10
9. Signature of the project main contact point.....	10

1. Executive summary

As a capacity building activity, a training hackathon was carried out in which European Node representatives were trained in the use of tools and services to create and maintain national species checklists. The representatives designed a workflow for this and worked together to improve existing tools, and created new ones, to be able to achieve this workflow. During the hackathon they learned from each other. The results of the hackathon were published in various ways and representatives have been reporting using the results in their Node operations afterwards. Therefore it was concluded that an event in the form of a hackathon can be a very effective way of capacity building amongst nodes. Dividing the participants in teams and giving them clear roles, and getting them well prepared to the event, helped in making this activity successful.

2. Contact information

Wouter Addink (Project leader)

Species 2000

PO Box 9517, 2300RA

Leiden, the Netherlands

wouter.addink@naturalis.nl, sp2000@sp2000.org

3. Project summary

3.1. Activities completed

1. REGIONAL TRAINING SUPPORT

The plan for this activity was a regional training support action through a training hackathon with focus on creating tools and a workflow that facilitates the creation of precursor national checklists, or improvement of existing checklists, from GBIF, CoL and regional checklists such as i.e. PESI datasources. The activity was completed according plan and results were

beyond expectations. Participants did give the event high scores in the evaluation, came up with a complete workflow for national checklists creation and curation and reported back that capacity needs were fulfilled and new skills and created tools were used in node activities and being further developed. 10-15 participants were expected (see proposal) but the event was attended by 20 participants, the maximum that Naturalis could host.

The training hackathon event was organized by Species 2000 from 2-5 March 2015, hosted at Naturalis in Leiden, Netherlands. Goals for the event were:

1. Capacity-building amongst GBIF Nodes and national species checklist owners in Europe
2. To try, improve and build tools to create precursor national checklists and to enhance existing checklists using the GBIF network, Catalogue of Life and other services.

A total of 20 participants, representatives from European nodes and trainers attended. There were participants from 10 different countries and multiple institutions with a good coverage over all parts of Europe.

In preparation of the event a survey was done to collect information from the participants, and to ensure that they came well prepared to the event, with a clear goal in mind for them to achieve. With the survey information was collected about goals and expertise of the participants, whether they agreed with open source licensing for any tools created in the hackathon. Capacity needs from the European Nodes were already collected in an earlier stage, also through a survey (see the proposal). Participants could also submit possible use cases for the hackathon. These were further discussed between the participants through email before the event.

The event started with one and half day plenary training to create awareness about the landscape of available tools and services and to explain how to use them. Presentations from the participants are available [here](#).

Topics were:

- Global Name services
- EU-Nomen
- EU Bon Taxonomic Backbone services
- GBIF API
- CDM name catalogue services
- Catalogue of Life services
- Norwegian Species Names Database
- Netherlands Biodiversity API
- Taxonomic Tree Tool
- Annosys
- Catalogue of Life Crossmapper

The remaining two and half days were used for programming in teams and discussing the results. Participants were grouped into three teams focusing on sub-topics:

1. **Crossmapping**, comparing precision and recall in matching two checklists, including 'checklist normalisation' and scoring to improve results
2. **Annotations**, identifying and annotating false positive occurrences with a system to allow people to annotate checklist cross-mappings and GBIF occurrence data.
3. **Distributions**, parsing datasources to gather evidence that a species occurs in a country.

Each group got a team leader and a reporter. The reporter created a report with the group results, which were discussed at the end of the event. Other roles people got in the groups were: back-end programmer, front-end programmer or junior programmer. Junior programmers used the hackathon mainly to get trained by the other programmers. There were also three people that got a role as 'domain expert'. These went from group to group to share their domain knowledge with the programmers.

Each group created one or more user stories related to their sub-topic to work on.

During the hackathon issues were explored related to the creation, maintenance and interoperability of national species checklists and prototype solutions were created. Netherlands Soorten Register (Dutch species checklist) was used as example checklist in the hackathon as well as checklists from Norway, Slovakia and Portugal (brought by the participants).

Group 1 - Cross-mapping

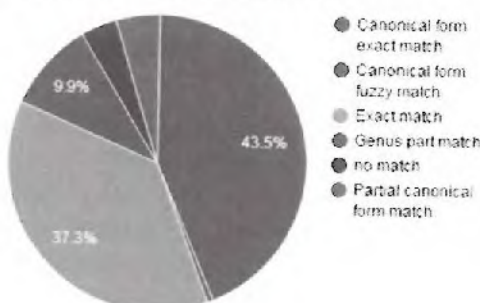
This group compared the Dutch national checklist to a species list derived from Dutch occurrence records linked to the Netherlands obtained from the GBIF network and cross-mapped against the Catalogue of Life.

The following tasks were carried out:

1. Examine how names in national checklists correspond to names in CoL - quantity and quality of positives and negative matches
2. Use these examples to assess and refine the GNA name-parsing and matching algorithm
3. Provide recommendations for future checklist cross-mapping tools and interfaces
4. Provide a qualitative review of existing tools for matching names or taxonomic hierarchies

Improvements implemented resulted in very high matching percentages >95% with the tested checklists.

Norwegian Plants Checklist - CoL



Group 2 - Annotations

This group adapted the Annosys framework, designed to annotate XML documents (ABCD concepts) to be able to create annotations for URLs. The annotation model was based on the [W3C Open Annotation Data Model](#). With the service, a checklist owner can evaluate potentially missing species by comparison of spatial distribution in GBIF and annotate every taxon where GBIF data are likely to be incorrect in terms of indicating national occurrence of that species.

Investigations done:

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 national 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?

The group created a demonstrator with the developed software to 'blacklist' records in GBIF using the Nederlandse Soorten Register checklist as example and provided suggestions for improvement of GBIF API service.

Group 3 - Distributions

This group looked at how to gather evidence for the occurrence of a given potential list of species in a given country with expert opinions about species distribution from sources like the Catalogue of Life and PESI and raw occurrence data from GBIF. Other GBIF information like basisOfRecord, age of occurrence (or last seen), establishment means, was also taken into account.

The group created a demonstrator that shows the results for the Nederlandse Soortenregister and allows to upload your own checklist to gather evidence. The group also made recommendations for improvement of the current GBIF webservices.

Code from all groups was documented and stored in github with an open source license. When the developed components from the groups are combined, a workflow can be established for creating and improving national checklists:

2 GBIF ADVOCACY ACTIONS

The objective here was a scientific publication about the results of the hackathon with as aim to achieve long-term results. The GBIF community expressed that in addition to a role as a global facility for data sharing and access, GBIF should also provide for applied data services that meet countries needs i.e. such as national checklists. Providing services generating national checklists and publishing about this will further engage (more) countries to support and participate in GBIF.

After the hackathon, the participants created a full technical report by combining the group reports, which is available at <http://www.gbif.org/project/2014-checklist-hackathon>. This was also shared with GBIF and the European Nodes. This report was used as the basis for a publication that has been created by Dave Remsen (Trainer) and Wouter Addink (Project leader). It is available at <http://sp2000.org/species-checklist-crossmapping-and-precursor-national-checklists-generation-gbif-mediated-data> and also at <http://www.gbif.org/project/2014-checklist-hackathon>. A publication in PLOS-One was attempted but rejected because a hackathon is not suitable to test proposed use cases and possible solutions in a scientific way.

Participants defined extra actions (not in the proposal) which were taken after the hackathon:

- Creating a full technical report
- Reporting the training hackathon event in GBITS
- Presentation about the results and discussion about further steps in the European Nodes regional meeting

- GBIF community site: a the group has been created and participants were invited to join the group

3.2. Ongoing and post-project activities

The cross-map solution created in the hackathon has been further developed and put into production as GNA services. It is actively maintained.

The annotations solution created as been included in the Annosys code base.

Code developed in the hackathon will be kept available by SP2000 on Github.

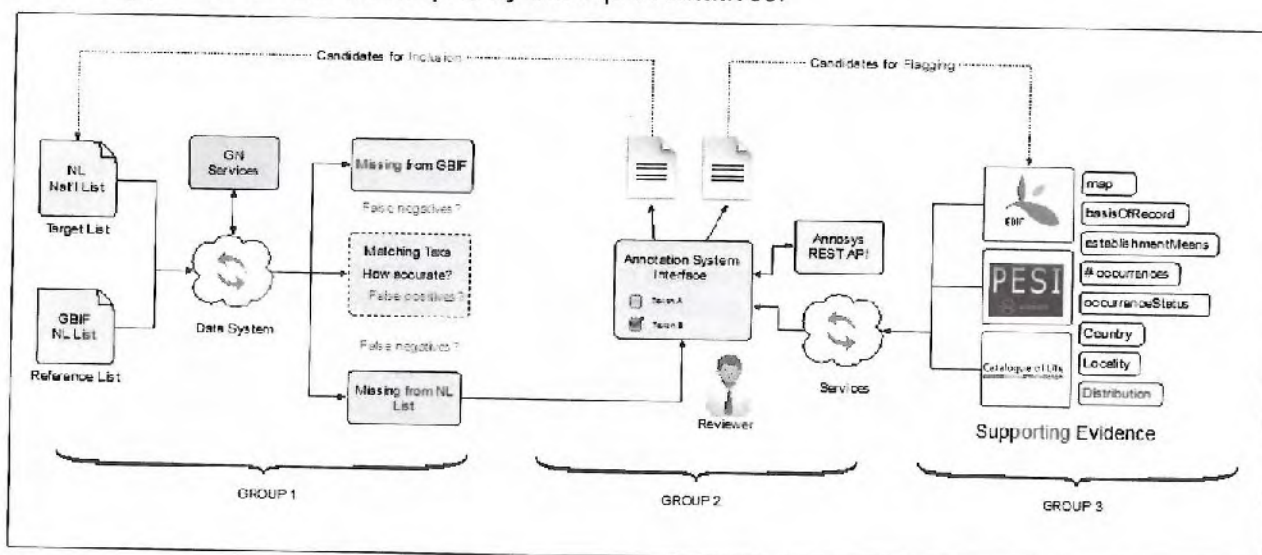
European Country Nodes committed themselves to inventory existing checklists for their country and take actions to make national checklists available in GBIF.

Participants have been reported that they have been using the tools and gained knowledge after the hackathon and improved checklists with them, and it can be assumed that they keep doing this.

4. Project objectives

All project objectives described in the proposal were reached:

- *Train European node representatives in usage of cross-mapping tools with their checklists.* Representatives were trained and used their own checklists.
- *Joint development with European node representatives on methods and protocols for national checklist generation in a hackathon, combined with hands-on training.* The following procedure was developed by the representatives:



See for a full explanation of this procedure the technical report. The procedure can be carried out with the developed demonstrators, but several shortcomings were identified in existing datasources and services that need to be solved to make this workflow fully operational.

- *Exchange experiences with validation tools, identifiers and ontologies.* Reached through presentation by participants and by working together in the Hackathon.

- *Make the developed tools and workflow for national checklist generation available to the GBIF community.* Tools were made available as open source and workflow was presented to the community.

5. Project deliverables

- Precursor software (demonstrators, mash-ups) for semi-automated national checklist generation and validation responding to needs in the GBIF community was created.
- Code repository on Github with code developed in the hackathon was created (<https://github.com/Sp2000/>), and code was stored with documentation and open source licences.
- An article about the hackathon results including recommendations for semi-automated checklist services has been created, and has been published on the [GBIF](#) and [SP2000](#) websites.
- A long term GBIF community site to further develop the tools and workflows based on results from the hackathon has been created.

6. Evaluation: findings and conclusions

After the hackathon, responses from participants were collected through a survey. Also, the email list was used to collect feedback of the capacity building results.

The GBIF Portugal representative reported:

We use one of the outputs of the Hackathon in our Nodes' activity. This is the matching tool, which compares and assesses the comparison of two taxonomic list. We have been using this tool in two cases:

- **Preparation and quality assessment of datasets to be published in GBIF**
The Portuguese citizen science platform Biodiversity4All (<http://www.biodiversity4all.org/>) is willing to become a GBIF publisher, and requested the support of the Portuguese Node to prepare a DarwinCore formatted dataset. The process involves the extraction of the information from the database, formatting and mapping to DwC terms. Users of the platform insert records using a controlled species names list managed by the managers of the project. Furthermore, the project shares its platform with the NL based project Observado, inheriting the species names list from it. Also, many species names in the list lack authorship. We are using the species matching tool gn_crossmap (see more at <https://globalnamesarchitecture.github.io/gna/resolver/checklist/2015/05/11/gn-crossmap-gem.html>) to check species names against:
 - Catalogue of Life (data source ID: 1)
 - GBIF Taxonomic Backbone (data source ID: 11)
 - Checklist da Flora de Portugal (Continental, Açores e Madeira) (data source ID: 176)

We use the results, including the matching scores to identify names not listed in the authority files, and to confirm full name matching, including species authorship and taxonomic rank.

This is a work in progress, so we may expand the crossmap exercise to other authority files, and assessing how to formally include this step in the quality control assessment of candidate data sets to be published in GBIF.

- **Preparation of Taxonomic Checklists to be published in GBIF**

Two taxonomic and distribution checklists are candidate to be published in GBIF:

Borges, P.A.V., Costa, A., Cunha, R., Gabriel, R., Gonçalves, V., Martins, A.F., Melo, I., Parente, M., Raposeiro, P., Rodrigues, P., Santos, R.S., Silva, L., Vieira, P. & Vieira, V. (Eds.) (2010). A list of the terrestrial and marine biota from the Azores. Princípiã, Cascais, 432 pp.

Borges, P.A.V., Abreu, C., Aguiar, A.M.F., Carvalho, P., Jardim, R., Melo, I., Oliveira, P., Sérgio, C., Serrano, A.R.M. & Vieira, P. (Eds.) (2008). A list of the terrestrial fungi, flora and fauna of Madeira and Selvagens archipelagos. Direcção Regional do Ambiente da Madeira and Universidade dos Açores, Funchal and Angra do Heroísmo. 438 pp.

The Portuguese Node is helping on preparing the data sets according to DwC format. The source information, for some biological groups, is lacking the infraspecies rank, and its also important to verify which species are lacking from global authority files like CoL or GBIF Taxonomic Backbone. We are using two tools to do this job:

- GBIF ECAT Name Parser (<http://tools.gbif.org/nameparser/>), to prepare the species names list, because the original source does not have the names parsed
- gn_crossmapp to:
 - obtain infraspecies rank (using GBIF Taxonomic Backbone (data source ID: 11))
 - confirm is species are present in global authority files like CoL or GBIF Taxonomic Backbone
 - depict any spelling error in the source data files

The GBIF Norway representative reported:

From the point of view of the Norwegian GBIF node, I would in particular mention the progress on behalf of the Norwegian Species Information Center with harmonizing and cross-mapping the official Norwegian names checklist against other names resources such as the Catalogue of Life. My impression is that most, if not all, of the hackathon participants expressed that we learned a lot and that the hackathon was a very efficient event for learning new things about the names architecture and coordinating our efforts.

ANTABIF reported:

From the Hackathon I learned how to improve the taxonomy mapping (e.g. from CoL, Aphia) process that I am using in AntaBIF project.

The GBIF Sweden representative reported:

The hackathon gave me valuable insights in the details about the web services. I now understand to much greater extent what the GBIF services can be used for and what their limitations are currently. This is most valuable for me as being in charge of the Swedish LifeWatch development. Beside this it was also very interesting to get information about the more global issues of cross mapping since I have been working with that a lot on the national level in Sweden. I think that this knowledge and all contacts gained during the meeting will help during the forthcoming years when I think very much will happen within this field. And the progress of this activities will depend on how well we collaborate on this. One more thing, I really thought that the hackathon showed how creative it is to come together from different organizations and try to develop things. I would definitively like to become involved in similar things in the future.

Dmitry Mozherrin (trainer, GNA) reported:

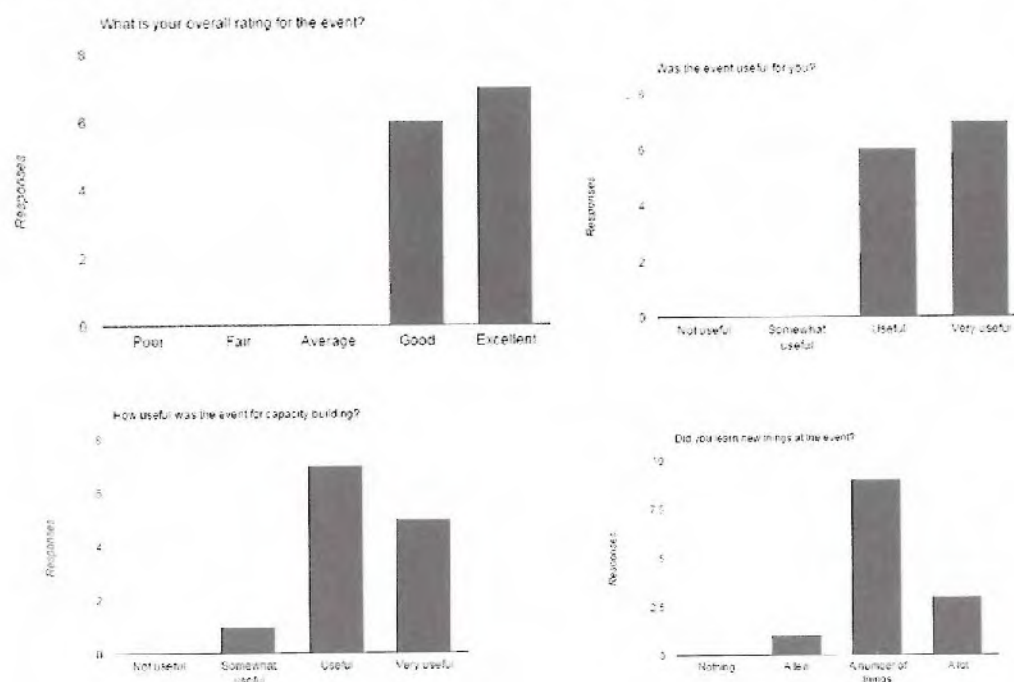
The Hackathon was very useful for me. I understood use cases of national checklists and Catalogue of life much better. During programming phase of the hackathon I learned in practical terms what is required to help national checklists to crossmap their data with CoL. The experience of the hackathon helped me to build better cooperation with CoL, start personal contact with Yuri Roskov, and ultimately helped me to decide on my next career move. So i would say hackathon had been disproportionally positive experience for me.

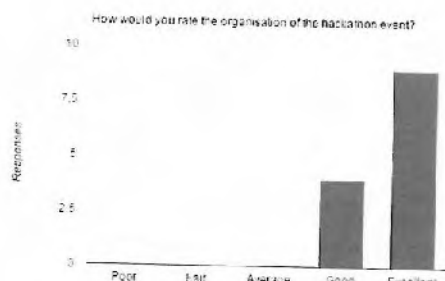
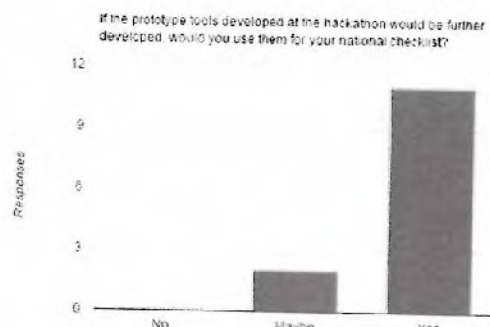
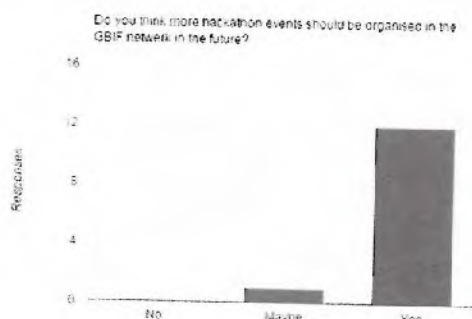
As a practical outcome there is now a `gn_crossmap` command line tool based on the code developed during hackathon, and web GUI is in works as well. (Note: see also <http://globalnamesarchitecture.github.io/gna/resolver/checklist/2015/05/11/gn-crossmap-gem.html>). The Web GUI did already become operational after this report from Dmitry: <http://resolver.globalnames.biodinfo.org/>).

Artsdatabanken (Norway) reported:

The hackathon has given us additional insight into taxonomy-related institutions, initiatives, challenges and their status. In addition to the actual implementation, it will also require an effort on our side to make our systems and workflow compatible with the tools that come out of this. The obvious great value of these tools is an additional incentive to make the necessary improvements locally.

Results of the anonymous evaluation form (all participants filled in the form):





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?".

7. Recommendations and lessons learned

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.

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 topics 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

To effectively use the data in the GBIF network and in CoL for the user stories defined in the hackathon, both the data and the services have to be improved. The participants made several recommendations for this, see the [technical report](#) for details.

Node participants were able to finance more of their travelling than expected. Therefore there was no need to ask GBIF for a second installment (there was budget left). On the other hand, getting trainers without paying their travel expenses was difficult and we had to spend more budget on that.

The email list has been used after the hackathon for further communications, but the group in the communications portal does not seem to be used although several people subscribed. This was foreseen and the participants therefore also came up with other suggested actions for further communications after the hackathon, which were carried out (see reported activities).

From attempts to create a scientific publication about the hackathon can be learned that it might be better to instead create a report about the activity for GBIF and a separate paper describing developed/improved software to be published in the "software tools" section of a scientific journal.


8. Future plans

Species 2000 aims to use the publication as starting point for project proposals to establish the workflow described by the participants in the community. Options for project funding have been discussed further by the European Nodes in the regional nodes meeting, but no opportunities were found yet.

9. Signature of the project main contact point

Signed on behalf of the project partners

Date


W. Adkins

30-3-16