Russia2019_04

European North-East Russia mosses occurrence data mobilization on the base of SYKO herbarium mosses collection

FINAL ACTIVITY REPORT

Guidelines on how to complete the activity report are included in italics. You are welcome to remove the guideline text from the document before you submit the report.

Please note that once the activity report has been approved, it will be added to your project page. Therefore, we kindly ask you not to add any contact details of persons in the report unless you have permission from the person to do so.

Contents

1. Executive summary

Provide a brief explanation of the project and its implementation, the objectives achieved, lessons learned and conclusions.

The aim of the project is to digitize and publish label information from at least 8,000 moss specimens in the herbarium (SYKO) moss collection of the Institute of Biology of the Komi Scientific Center of the Ural Branch of the Russian Academy of Sciences. As the non-dataset project deliverable the project will document the process of mobilizing data from herbarium labels in multiple forms: an article in the open access peer-reviewed journal (in Russian); a webinar on the institution's YouTube channel, and a report at one Russian scientific conference. All these tasks have been completed. We were able to develop a data entry system that significantly improves the efficiency of monotonous manual labour. Up to 30 August 2019 there are more than 20,000 labels images uploaded to the database and more than 14,000 of them were digitized with minimum set of fields: catalogue number, species name, collection date, geographic coordinates, names of persons who collected and identified the species in the sample. This number of digitized labels allowed us to publish and to register in GBIF 14,871 mosses occurrences. The project results were presented in regional scientific conference "Symbiosis 2019" (Perm city, 13.05.2019) and in the form of YouTube webinar "Digitization of herbarium label data and their preparation for publication in GBIF. Practical case of bryophytes collections of the SYKO herbarium". The experience obtained by us during the project was published in the form of open access preprint on the ResearchGate. The source code of web-based software (Django project) created during the project was published in open scientific repository Zenodo.

2. Contact information

Provide the name, institutional affiliation, role in the project and contact details of the author(s) of the report.

Ivan Chadin, Institute of Biology of Komi Science Centre of the Ural Branch of the Russian Academy of Sciences (IB Komi SC UB RAS, team leader. E-mail:. Tel: .

3. The project and its objectives

A brief summary of the project to help readers understand its objectives, including, for example:

- *The project's start date and expected duration*
- *A list of project participants and description of the main stakeholders*
- *The targeted capacity needs as outlined in the project proposal*

- *The project objectives and expected deliverables as included in the project proposal*

The project's start date and expected duration

The project's start date: 04.02.2019

Expected duration: 6 months and 26 days (calendar days)

A list of project participants and description of the main stakeholders

Participants:

Chadin Ivan – team leader, program coding, dataset publishing with IPT

Zheleznova Galina – briologist, collection curator, labels data entering, verification of entered data

Shubina Tatyana – briologist, labels data entering, verification of entered data

Rubtsov Mikhail – engeneer, geo-referencing

Litvinenko Galina – technician, labels image digitization, labels data entering

Stakeholder

Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences (IB Komi SC UB RAS). It was organized in 1962. The Institute is the largest the academic centre for ecological and biological research on the European North-East Russia. The Institute consists of six departments and four laboratories, Zoological museum, Botanical garden and herbarium (SYKO). Main areas of research are: Study of biodiversity, structural and functional organization, stability and productivity of taiga and tundra ecosystems; Biological action of ionizing radiation and other physico-chemical factors on cells, living organisms and natural ecosystems; problems of radiation and ecological genetics; Study of physiological-biochemical basis of adaptation and reproduction of plants in cold climates; Study of biologically active compounds in plants of natural flora and introduced plants; Development of methods for monitoring, bioindication; creation of inventories and databases of biological resources of the European North-East with the use of remote sensing and GIS technologies.

The Institute maintains rich biological collections. The SYKO herbarium keep 315 045 samples (including 206 445 samples of vascular plants and 57 000 samples of mosses). IB Komi SC UB RAS is interested in its biological collections digital mobilization.

The targeted capacity needs as outlined in the project proposal

The amount of data in the GBIF on mosses occurrences in the European North-East Russia (>1 million sq km) is significantly lower than the amount of data accumulated in biological collections. Currently, information on 5,106 mosses occurrences are available for this territory (4,120 of which are published by the authors of this application). At the same time, the SYKO herbarium contains the largest collection of moss in the region, which has more than 57,000 items. On average, for each storage unit there are at least two species of mosses. Thus, at present, information on approximately 100,000 mosses occurrences needs to be digitized and translated into a format that conforms to the Darwin Core.

The project objectives and expected deliverables as included in the project proposal

a. Data

Title of dataset	**Taxonomic/ geographic/ temporal scope**	**Approximate number of records**	**Sampling methodology/protocol used (if relevant)**	**Geographic accuracy for most records (in m or km, or province, country etc.)**	**Current state (e.g. undigitized, digitized)**
Mosses Collection of SYKO Herbarium (Sykrtyvkar, Russia)	Plantae, Bryophyta 1947-2017	8,000	Data will be acquired from herbarium labels. Georeferncing will be performed according Zheleznova G, Shubina T, Degteva S, Rubtsov M, Chadin I (2018): Moss occurrences in Yugyd Va National Park ... http://ib.komisc.ru:8088/ipt/resource?r=mosses_occurrence_yugyd_va&v=1.5	1.6–3.0 km	Undigitized

b. Other deliverables

In the first half of the project period, the experience gained during its implementation will be presented in the form of a webinar on the YouTube channel for biodiversity specialists.

Prior to the end of the project, the results will be presented in the preprint in Russian, which will later be submitted for publication in an open access peer-reviewed journal, as well as in a report at the scientific conference.

4. Project implementation

This section should provide readers with a good understanding of the project, from the original plans to the final implementation, highlighting:

- *The activities that have been completed at the time of writing the report, and those that are ongoing or pending (e.g. longer-term evaluation, follow-up projects/meetings/training events) and your plans for their completion.*
- *How the different partners in the project have contributed to its implementation.*

4.1. Activities completed

Describe the activities that have been completed at the time of writing the report. Explain how the different partners in the project have contributed to its implementation.

The database web interface to it were created. Users able to interact with database through special website. The series of forms allow users to upload label images and to transfer label data into the fields mapped to the terms of the DarwinCore. This work was done mainly by Ivan Chadin with great help from other team members (testing, design decisions).

Digital images of at least 20,000 labels were captured and uploaded into the database. The work was done mainly by Galina Litvinenko (image capture) and Ivan Chadin (image uploading).

Labels data were entered into the database. Catalogue number, species name, collection date, geographic coordinates, names of persons who collected and identified the species in the sample were entered for more than 14,000 labels. Data about substrate were transcribed for about 16,000 labels, The habitat descriptions were transcribed for about 6,000 labels. This work was done by Galina Zheleznova and Tatiana Shubina (species names, names of collectors and determiners, descriptions of substrates and habitats), Galina Litvinenko (dates of specimen collection, catalogue numbers) and Mikhail Rubtsov (georeferencing).

The quality control of labels data transcribing (except georeferencing results) was performed by professional bryologists — Galina Zheleznova and Tatiana Shubina. The georeferencing checking was performed by Ivan Chadin and Mikhail Rubtsov.

The verified dataset with main fields: catalogue number, species name, collection date, geographic coordinates, names of collectors and determinators (14,871 mosses occurrences) was published in the GBIF (DOI: 10.15468/yjdjs4). This was done with IPT installation by Ivan Chadin. An additional label data (substrate and habitat description) after verification by bryologists will be published in GBIF as a new version of dataset.

The experience obtained during the project was presented in the form of webinar on the YouTube channel (https://www.youtube.com/watch?v=IMs6k8PUrN8), in form of report at the scientific conference (Symbiosis–Russia 2019, Perm, 13–15 of May 2019) and in the preprint in Russian (DOI: 10.13140/RG.2.2.21925.24803), which will later be sent for publication in an open access peer-reviewed journal. The source code of web-based software (Django project) created during the project was published in open scientific repository Zenodo. This part of work mainly was done by Ivan Chadin with active help from other team members.

4.2. Ongoing and post-project activities

Highlight ongoing or pending activities (e.g. longer-term evaluation, follow-up projects/meetings/training events) and your plans for their completion.

The project team is working on substrate and habitat descriptions transcription after completing this work, a new version of the dataset will be published in GBIF before the end of 2019. The paper based on published preprint should be sent in Russian pear-reviewed scientific journal before the end of 2019.

5. Project deliverables

This section should summarize the project activities completed by the end of the project with a description of the associated outputs and deliverables. Please highlight any changes from the original plans provided in the full proposal by filling in the column 'State by final report'. You are welcome to attach deliverables to the report as annexes or to link to them.

Make sure to include details of data mobilized through the project and/or re-usable information resources or tools. Should your deliverables include data publication to GBIF, please make sure to include the project ID in the dataset metadata.

Also, please comment on the expected milestone for the final reporting as defined in the contract.

a. Data

Title of dataset	Taxonomic / geographic / temporal scope	Approximate number of records	Sampling methodology/protocol used (if relevant)	Geographic accuracy for most records (in m or km, or province, country etc.)	Current state (e.g. undigitized, digitized)	State by final report
Mosses Collection of SYKO Herbarium (Sykrtyvkar, Russia)	Plantae, Bryophyta 1947-2017	8,000	Data will be acquired from herbarium labels. Georeferncing will be performed according Zheleznova G, Shubina T, Degteva S, Rubtsov M, Chadin I (2018): Moss occurrences in Yugyd Va National Park ... http://ib.komisc.ru:8088/ipt/resource?r=mosses_occurrence_yugyd_va&v=1.5	1.6–3.0 km	Undigitized	14,871 mosses occurences published in GBIF

b. Other deliverables

Description	State by final report
In the first half of the project period, the experience gained during its implementation will be presented in the form of a webinar on the YouTube channel for biodiversity specialists.	The record of webinar is available on YouTube: https://www.youtube.com/watch?v=TMenzugsCaY (available only by direct link) Trimmed video: https://www.youtube.com/watch?v=IMs6k8PUrN8 (available for all).
Prior to the end of the project, the results will be presented in the preprint in Russian, which will later be submited for publication in an open access peer-reviewed journal, as well as in a report at the scientific conference.	Methods of labour productivity increase for the digitization of label data of biological collections. Case of bryophytes collections of SYKO herbarium [In Russian] // https://www.researchgate.net/publication/335600975_Priemy_povyseniа_proizvoditelnosti_truda_pri_ocifrovke_etiketocnyh_dannyh_biologiceskih_kollekcij_Opyt_mobilizacii_dannyh_kollekcii_mohoobraznyh_gerbaria_SYKO. DOI: 10.13140/RG.2.2.21925.24803

Expected milestones by final report:

Milestone	Status by final report
1. All mobilized data has been published to GBIF.org 2. (All published data must data meet the minimum requirements outlined in the Data Quality Requirements available at http://bid.gbif.org/en/community/data-quality/) 3. Webinar held (YouTube channel) for biodiversity specialists. 4. Best practices and lessons learned have been documented	1. Done 2. Done 3. Done 4. Done

6. Project communications

Report on the way the results of your project have been communicated and shared with the project stakeholders and broader GBIF community . Please also review the page describing your project available from https://www.gbif.org/project/5ZsAifyl6z0OguyoNTFIlu/mobilizing-moss-occurrences-from-the-komi-science-centre-herbarium*. Highlight any additional documents, events, news items or links that you would like to add to your page.*

In the first half of the project duration, the obtained experience was presented in the form of webinar on the YouTube channel (https://www.youtube.com/watch?v=IMs6k8PUrN8). The webinar was combined with a seminar held in the conference hall of the Institute of Biology. C.a 20 botanists and zoologists from the Institute attended the seminar in conference hall.

The masterclass “The concept of “Open Science” on the example of biodiversity primary data publication through the global portal GBIF.org: theory and practice” for students and teachers of Perm State University was held by Ivan Chadin on 13.05.2019. Short communication about our principles of label data digitization was made during Invasive Alien Species and Data Collaboration Seminar and Workshop (Petrozavodsk, 9-11 September 2019). The oral presentation “Methods of labor productivity increase for the digitization of label data of biological collections. Case of bryophytes collections of SYKO herbarium” will be given on school and seminar “The digitization and publication of the field records and biological collections of reserves and national parks of Russia” (The Prioksko-Terrasny Nature Biosphere Reserve, 1-4 October 2019).

The project results and experience was presented in the preprint in Russian (DOI: 10.13140/RG.2.2.21925.24803), which will later be sent for publication in an open access peer-reviewed journal. The source code of web-based software (Django project) created was published in open scientific repository Zenodo (DOI: 10.5281/zenodo.3385382)

7. Evaluation: findings and conclusions

An assessement of the overall outcomes and impacts of your project, including strengths and weaknesses in its implementation and results. Try to identify clear conclusions from your experience during the implementation of the project. If any changes have been made to the project plans please clearly indicate this in the report and the reasons for this. Also report on any feedback on the project’s relevance from the partners and stakeholders

All project activities have been carried out according to the plan stated in project proposal. Digital images of herbarium labels began to be obtained with the help of a home-made planetary scanner. The database and the web interface were constructed with the help of Django framework. The workflow of data entering was modified during project implementation and after several trials and errors the high productivity workflow was implemented. Entering data from labels is basically manual labour and therefore, we have applied to it the well-known Frederick Taylor principles of increasing the productivity of manual labour. The “Scientific management” principles were implemented with modification of web interface. All labels information were splited in portions and for every portion the series of special web-forms were created: form to input the elements of the collection date; form for entering species names; form for entering the names of collectors and determinators. Since labels are stored in the catalogue in a certain order, the preservation of this order during label images uploading allowed to apply another method of acceleration of manual processing: the “Copy from previous” button. This button speeds up data entry by several times, keeping the ability to quickly modify incorrectly entered data. The most difficult to decipher labels data is the georeferencing of occurrence on the base of text description.

This work was simplified by combining labels into groups by the name of the collector, and the year of collection of samples. It is known that botanists make most of their routes by walking and in the same date they usually work in radius 5-10 km. So it is usual to have tens of labels with the same or close geographic coordinates that were collected in one day by the same collector.

Up to 30 August 2019 there more than 20,000 labels images uploaded to the database and more than 14,000 of them were digitized with minimum set of fields: catalog number, species name, collection date, geographic coordinates, names of persons who collected and identified the species in the sample. Data about substrate were transcribed for about 16,000 labels, The habitat descriptions were transcribed for about 6,000 labels. This number of digitized labels allowed us to publish 14,871 mosses occurrences in GBIF (Zheleznova G, Shubina T, Rubtsov M, Litvinenko G, Chadin I (2019). SYKO Herbarium Mosses Collection. Version 1.5. Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences. Occurrence dataset https://doi.org/10.15468/yjdjs4 accessed via GBIF.org on 2019-09-12). All digitized data were checked by bryologists. The geographic coordinates were checked for gross error with the help of occurrence location visualization in QGIS software.

The results of the project were presented to the professional community in form of masterclass “The concept of “Open Science” on the example of biodiversity primary data publication through the global portal GBIF.org: theory and practice” for students and teachers of Perm State University on the regional scientific conference “Symbiosis 2019” (Perm city, 13.05.2019), in the form of webinar “Digitization of herbarium label data and their preparation for publication in GBIF. Practical case of bryophytes collections of the SYKO herbarium” and in open access preprint published in social network “ResearchGate”: “Methods of labour productivity increase for the digitization of label data of biological collections. Case of bryophytes collections of SYKO herbarium [In Russian]” (DOI: 10.13140/RG.2.2.21925.24803). The source code of web-based software (Django project) created during the project was published in open scientific repository Zenodo (DOI: 10.5281/zenodo.3385382).

The feedback of these events showed that researchers in the field of biodiversity have a great interest in the possibilities of publishing their data through GBIF. At the same time, the level of awareness about the procedure of publication and use of data with help of GBIF remains relatively low. The feedback of main stakeholder — Institute of Biology of Komi Science Centre of the Ural Branch of the Russian Academy of Sciences was very positive. Our colleges were very interested in possibilities to publish their primary data through IPT instance of our Institute. The director of Institute Svetlana Degteva (she is a professional botanist) was impressed by the productivity of project team. The SYKO herbarium has several collections (fungi, lichens, algae) with number of specimens comparable with number of labels digitized by our team in 4 months.

8. Recommendations and lessons learned

This section should be addressed to others preparing similar projects in the future. Try to identify your experiences that could help others to design and implement projects more effectively, including the best practices to adopt and the pitfalls to avoid.

We believe that we were able to achieve all the stated goals of the project due to the combination of the technology and the project team list. It is highly recommended including into the project team persons who are curators of the collection to be digitized or persons who make significant contribution for the collection enhancing. The software architecture developed by us in the project may be used as basis for requirements specification for developing database and web-interface to it taking into account the unique feature collection

to be digitized. There two obvious modifications that will further improve the quality and performance the software: maintain individual logs of each operator's actions and ability to issue individual task for one week for every operator.

To reach a wider professional audience and more rapid feedback it is necessary to regularly blog the project in social networks like ResearchGate (and maybe others).

9. Future plans

A description of how the partners involved will build on the results of this project in their future work. This could include future collaborative activities, such as plans to complete any unfinished project activities and how the future impact of the project could be monitored or measured.

The digitization of the mosses collection and its publication in GBIF will be continued. If we can attract additional financial support, all mosses collection (57,000) digital mobilization may be completed in 2-3 years (for the field set: species names, collector and determinator names, date of collection, decimal latitude and longitude).

The software developed during the project will be maintained and new version will be published under GPL-like license. Translation of user interface into English will allow to expand potential user audience.

The preprint “Methods of labour productivity increase for the digitization of label data of biological collections. Case of bryophytes collections of SYKO herbarium” will be sent in the open access peer-reviewed journal. The results of the project may be present in National scientific conference with international participation “INFORMATION TECHNOLOGIES IN BIODIVERSITY RESEARCH” Yekaterinburg, Russia, 20-25 April 2020.

The future impact of the project could be monitored or measured by:

– monitoring of dataset new versions “SYKO Herbarium Mosses Collection” dataset publishing;

– links to open repository with new version of herbarium labels digitization software;

– links to scientific papers published with citation of “SYKO Herbarium Mosses Collection” dataset.

All new links may be published on project on the GBIF website (https://www.gbif.org/project/5ZsAifyI6z0OguyoNTFIIu/mobilizing-moss-occurrences-from-the-komi-science-centre-herbarium).

10. Annex – Sources of verification

1. Zheleznova G, Shubina T, Rubtsov M, Litvinenko G, Chadin I (2019). SYKO Herbarium Mosses Collection. Version 1.5. Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences. Occurrence dataset https://doi.org/10.15468/yjdjs4 accessed via GBIF.org on 2019-09-12

2. The Program of XI All-Russian Congress of young scientists-biologists with international participation: "Symbiosis–Russia 2019» http://imbiocom.ru/wp-content/uploads/2019/05/PROGRAMMA-Simbioz-Rossiya-2019-1.pdf

3. The webinar “Digitization of herbarium label data and their preparation for publication in GBIF. Practical case of bryophytes collections of the SYKO herbarium” recording: https://www.youtube.com/watch?v=TMenzugsCaY. Trimmed video: https://www.youtube.com/watch?v=IMs6k8PUrN8

4. Methods of labour productivity increase for the digitization of label data of biological collections. Case of bryophytes collections of SYKO herbarium [In Russian]:

https://www.researchgate.net/publication/335600975_Priemy_povysenia_proizvoditelnosti_truda_pri_ocifrovke_etiketocnyh_dannyh_biologiceskih_kollekcij_Opyt_mobilizacii_dannyh_kollekcii_mohoobraznyh_gerbaria_SYKO. DOI: 10.13140/RG.2.2.21925.24803

5. Information system for digitization of moss collection label data [In Russian]: https://zenodo.org/record/3385382#.XW9Wq6VS9hE, https://doi.org/10.5281/zenodo.3385382

Signed on behalf of the project partners Date

-
-
-
-
-
-
-