Project ID: CESP2017-0013

Project Title: Sharing VertNet experiences and tools on biodiversity data quality with the Spanish-speaking community

MID-TERM ACTIVITY REPORT

Contents

Executive summary

During the first period of the project we have developed the implementation of the VertNet's Data Migrator Toolkit within the SiB Colombia data sharing workflow. Starting with a workshop, we have trained the partners in the use of the tool and we have reached the level of its actual implementation to data sets shared through SiB Colombia. We have also performed enhancements to the toolkit and have produced and translated documentation accordingly. We have accomplished these tasks ahead of the time planned in the original agenda. The approaches taken for the development of the project have proven efficient and effective, and therefore, we will continue using the strategies utilized to date for the remainder of the project. We anticipate that there will still be issues to solve in the next months, but that the results by the end of the project will be as expected.

Contact information

Representative	Affiliation	Role(s) in the project	Contact
Leonardo Buitrago	SiB Colombia	Main Contact and activities coordination	albuitrago@humboldt.org.co
Paula Zermoglio	VertNet	Participant	pzermoglio@gmail.com
Dairo Escobar	SiB Colombia	Participant	descobar@humboldt.org.co
John Wieczorek	VertNet	VertNet contact and activities coordination	tuco@berkeley.edu
David Bloom	VertNet	Participant	dbloom@vertnet.org

Introduction

In this midterm report we compile the results achieved to date within the project. We inform on the deliverables that are already available and the activities scheduled to accomplish the tasks that remain for the second phase of the project. Additionally, we present a summary of the challenges encountered during this first working period and how we plan to overcome them.

In the project proposal there were two basic criteria to evaluate progress: 1) number of datasets subjected to data quality improvement using the "Darwin Core Data Migrator Tookit"; and 2) amount of documentation translated into Spanish for its broader use by the Spanish-speaking community. As the first part of the project mainly contemplated the training on the use of the tool, the first criterion is best suited for the final report and was not considered in this evaluation. Instead, we used the second criteria, plus an assessment of the results of the training workshop and how this knowledge is being consolidated.

The knowledge transfer about the data quality tool has been tracked during this first phase, and captured in: a) a workshop full report; b) an issue tracker in a GitHub repository; and c) participants' shared documents. All these have been continuously used by all partners as regular self-evaluation instances. Furthermore, monthly meetings have been carried out to assess difficulties and resolve any pending issues. All this allows all participants to be aware of the current state of the project, the difficulties encountered and the agenda planned to solve them. In this sense, the evaluation is an on-going process, rather than a one-time only report. This helps us keep the project running swiftly, overcoming problems as soon as they arise. As the methods listed above have proven efficient for the project partners, they will remain the main mechanisms for evaluation implementation in the second phase of the project.

The project and its objectives

The Colombian Biodiversity Information System (SiB Colombia) has been working for the last decades to identify and solve different challenges concerning biodiversity data quality at different steps of the data sharing process. However, up to date, SiB's data quality validations and improvements are performed manually or using separate tools. This process consumes time and resources, and renders it extremely difficult to validate large data sets. Therefore, a need has been identified to improve and automate the mechanisms that allow not only data quality checks but also providing data quality feedback to the providers.

VertNet has ample experience in the automation and development of protocols and tools to improve data quality at several stages of the publication chain. One of these tools, the Darwin Core Data Migrator Toolkit, is able to generate automatic data quality checks and improvement reports on datasets, facilitating the improvement of the data quality of occurrence records before publication.

This project is a one year (July 2017-July 2018) collaboration between SiB Colombia and VertNet, both GBIF Nodes, and its main goal is to advance the data quality assessment and improvement processes within the SiB data sharing workflow. The expected deliverables are: a) to implement the Darwin Core Data Migrator Toolkit in the SiB's data workflow, and b) to translate its documentation and to create new documentation where needed, in order to enable the use of the tool by the Spanish-speaking community.

While the participants of the project are SiB Colombia and VertNet, other stakeholders will be directly involved in / benefited from its outcomes. For instance, the data quality tool will be applied to data sets from many Colombian institutions that are active providers of the SiB Colombia, including for example the Humboldt Institute, who will get the data quality reports back from the process. The outcomes of this project will also contribute to achieve the goals and milestone of the priority "Improve Data Quality" of the GBIF Strategic Plan 2017-2021, helping to improve data quality of occurrence records within the Spanish-speaking community in general. At a national level, the results of this effort will contribute to support the activities of the Annual Operation Plan 2017 of the SiB Colombia Technical Secretariat.

Project activities completed by mid-term

During the first phase of the project, the following activities were completed:

1. **Meetings and communication**.
    a. **Monthly meetings**. Meetings here held remotely in July (pre-meeting, before the official starting date), August, and November. On site meetings were held during the workshop in September (see below). The agenda and notes from each of those meetings were captured in a Google document (https://goo.gl/HbBywP) . During these meetings we: a) organized the ongoing and future activities (e.g., translation of documentation), b) resolved issues concerning the training on and implementation of the data migrator toolkit.
    b. **Other meetings**. Other remote meetings took place involving some of the participants to solve particular issues (e.g., in the practice / implementation of the migrator toolkit). Notes from this meetings were taken in the document above as well as directly as issues in the GitHub repository.
    c. **GitHub repository**. A GitHub repository was set to capture all issues concerning the process of knowledge transfer and to track all necessary changes. (https://github.com/SIB-Colombia/CESP-GBIF-SiB-VertNet, see Issue tracker). Also, in this repository we keep shared documents (e.g., copies of the documentation produced, workshop reports).

2. **Workshop**. A 5-day workshop was held at the Humboldt Institute, 11-15 September 2017. The goal of the workshop was to transfer VertNet's experience on the automatic improvement of biodiversity data quality on the publication workflow of SiB Colombia and the Humboldt Institute. Thirteen people participated in the workshop (2 from VertNet, 6 from SiB Colombia, 5 from the Humboldt Institute), spanning from information technologies experts to biological collections members. During the workshop the following activities were carried out:
    a. Presentation and intensive practice on the Darwin Core Data Migrator Toolkit to train the SiB and Humboldt Institute members on its use.
    b. Assessment of available documentation and identification of documentation gaps. Translation of key documentation and confection of new documentation (see documentation section below).
    c. Exchange of experiences about data digitization, data publication and data quality processes. (https://twitter.com/sibcolombia/status/908072554392322049)
    d. Planning of the activities to follow concerning the implementation of the tool.

e. Talk at the Universidad Javeriana, open to the general public, about biodiversity data quality issues.
(Video available at: https://www.youtube.com/watch?v=om1TdHOj5B8).
(Invitation available at: http://bit.do/invitationtalkPUJ)

A full report of the workshop was built in Spanish based on the notes taken during that week and translated into English, and both versions are available for download in the joint GitHub repository (EN version: https://github.com/SIB-Colombia/CESP-GBIF-SiB-VertNet/blob/master/Report%20Workshop%20SIB-VertNet_EN.pdf).

3. Darwin Core Data Migrator Toolkit Implementation.

The first phase of the Data Migrator Implementation to the SiB Colombia workflow was initiated during the workshop, and was based on training the parties on the use of the tool. A second phase was developed remotely, and a particular person from SiB Colombia was assigned to the task of learning all the details of the tool by using it on actual datasets and testing the outcomes against those run in parallel by VertNet members. During these two phases and as a result of the use of the tool by SiB Colombia, changes were introduced in the toolkit to enable new functionalities and to simplify some of its processes (see section below). Currently, the tool has been tested by SiB Colombia staff on a dataset from the fish collection of the Humboldt Institute, the quality reports have been produced and we are in the process of getting the feedback from the collection. From here, the next phase will include applying the migration process to other datasets shared through SiB.

4. Changes introduced to the toolkit based on its implementation at SiB Colombia.

During the implementation phases described above, the following changes were made to the migrator toolkit:

- Restructuring of the Access databases within the tool, in order to simplify the use of the macros and allow faster processing of the data.
- Restructuring of the migrator to accommodate much larger datasets.
- Addition of extensive commentary and indicators of required and to-be-customized queries in the main macro of the migrator.
- Review of the data quality reports that result from the use of the tool, adding new reports.
- Change to the vocabulary management process so that it can be collaborative and allow merging vocabularies from different sources.
- Addition of vocabularies checked by the toolkit (i.e.,coordinatePrecision, identificationVerificationStatus, license, nomenclaturalStatus, organismScope, day, and taxonomicStatus).
- Enabling Event and Taxon-based migrations.
- Update of the migrator to comply with the complete current version of Darwin Core.

All changes were tracked in detail in the toolkit repository: https://github.com/VertNet/toolkit/commits/master?after=609679224c445d715d49201d311975abf0b66d9e+34.

5. Documentation

Many of the documents available for the migrator toolkit prior to this project were meant for internal use in VertNet, and were therefore very technical in nature and in

need of improvement for a more general use. In this period we updated most of the migrator's documentation to include detailed and graphical explanations in English and Spanish to facilitate the broader use of the tool. New documentation was also created for this purpose.

a) **Migrator Workflow explanation.** The original file in English was improved by adding an explanatory graphic (https://github.com/VertNet/toolkit/wiki/Migrator-Workflow-(EN)) and the translation of both the text and a graph were performed (https://github.com/VertNet/toolkit/wiki/Migrator-Workflow-(ES)).
b) **Step by step explanation of the use of the migrator toolkit.** This document, which previously consisted in a brief list of steps in a text file, was fully remade to reflect the changes in the migrator and to include: explanations on the preliminary computer settings necessary to run a migrator, a detailed explanation of each step of the migration process including screen captures and newly made schemes to exemplify the migrator's functioning. The document was then translated into Spanish and incorporated to the migrator's downloadable files in its GitHub repository as an integral part of the toolkit (https://github.com/VertNet/toolkit, see files: README_Instructions for use_EN, and README_Instrucciones de uso_ES).
c) **Migrator Data Quality Reports explanation.** Previously the migrator included a text file with the explanation in English of the data quality reports. This file was meant for the data providers to understand what each migrator report contains. That file was expanded in content and its presentation was improved. A Spanish version of it was made, and both are now available as an integral part of the toolkit (https://github.com/VertNet/toolkit/tree/master/reports).
d) **Internal functioning of the toolkit:** explanation of what each of the tables, queries and macros within the migrator do. A thorough explanation was built based on Google spreadsheets (https://docs.google.com/spreadsheets/d/1RiAcRosAm-IekXq_0_uKRBqx3NhCCE4OVasrZIc-ngk/edit?usp=sharing). This explanation is still in English but will be translated into Spanish and incorporated to the documentation of the toolkit in its GitHub repository shortly.

Project communications

Results of the project up to date are already published in the GitHub repositories mentioned above in their different forms (e.g., improvements made to the toolkit, new documentation, etc.). The stakeholders directly involved in the project (e.g., the Humboldt Institute) participate in the meetings and in shared documents, and are therefore aware of all the activities and progress made during this period.

In order to communicate the results more broadly, we have planned to build a document describing the data quality process within SiB Colombia, particularly using the Migrator Toolkit, and emphasizing the importance of biodiversity data quality. This document targets an audience with diverse technical skills, and will therefore contemplate using non-technical language for its broader understanding. It will be distributed among the data providers that publish their data through SiB Colombia, as well as distributed to the other GBIF Spanish-speaking Nodes.

Mid-term evaluation findings and recommendations for the remaining project implementation period

General comments

During this first part of the project we have focused on training the participants in the use of the toolkit and on determining the best ways to implement it within the SiB Colombia data publication workflow. All the activities developed during this period were in line with or ahead of the proposed calendar (see below), and we consider that they were accomplished satisfactorily.

Communication

The workshop held in September proved very useful, as it allowed us to set the scene for the whole implementation process. Meeting face to face improved understanding of the scope and potential issues to be encountered during the next months, and enabled us to establish a clear action plan to respond to the issues to come. It also provided all parties with a broader perspective and helped us to place the project in the bigger picture. The exchange of experiences in digitization and data quality processes was useful, as it allowed us to identify synergies and to envision future cooperations.

The communication venues established among the members of the project (e.g., the GitHub repository) constituted an invaluable tool to keep track of the progress and issues found during these months. Having monthly meetings was also useful to keep all members of the project involved at all times in all the activities that were taking place. We will keep using these communication methods in the following months, as they proven effective and efficient.

Toolkit implementation

The activity that presented most challenges was the implementation of the toolkit. This was due to the complexity of the tool and the novelty that it represented for the SiB Colombia staff. As with any other tool, we observed a learning curve. The first phase of the implementation required much interaction between SiB Colombia and VertNet to make sure every step was understood and carried out correctly. Such interaction was extremely useful for all parties. As a result of SiB Colombia use of the tool, we were able to identify aspects of the tool that could be improved, and many enhancements were made to it from VertNet's side and immediately tested by SiB Colombia. This cooperation rendered a better tool which now has new capabilities and can be more broadly utilized.

Issues that arose during the implementation were tracked as explained above and the person that was particularly assigned to the migrator implementation took thorough notes of the progress done. In his expert opinion, by the beginning of November, around 70% of all the functionalities of the migrator toolkit were learnt, and many of the remaining challenges were related to the particularities of the data sets that are migrated (i.e., some data sets require special customization of the migration process). Since then, more issues have been resolved and the migrator has already been implemented to a dataset (see below).

In the original calendar we had planned to start the Colombian datasets evaluation with the VertNet Data Migrator Toolkit in January 2018. Given the progress made in the implementation during the informed period, we have been able to already start this task, testing the toolkit with the Fish Collection from the Humboldt Institute. This advancement

enables us to identify issues earlier in the process and provides us with more time for implementation of the toolkit to more datasets within the time frame of the project. We believe we are now ready to bring the implementation to its next level, applying it to more data sets shared through the SiB Colombia.

Documentation

With respect to the documentation process, we have changed the original plan. The proposed agenda contemplated the production and translation of documentation starting on February 2018. However, we realized that: a) documentation was important to have during the implementation process, and b) it was most effective to have it been developed as we go, capturing newly made changes. An advantage of having the documentation developed in parallel with the implementation is that we can view the processes from a new user perspective, and capture all necessary information for others who have never used the tool to be able to understand it. This has been a unique opportunity to do so. Also, we find value in translating on-the-go (rather than at the end, as originally proposed) because it makes the documentation available as soon as possible. Our approach has proven effective and has strengthened the collaborative spirit between SiB Colombia and VertNet.

Translations of the documents were done differently from what was projected in the original proposal. We had planned that the translation would be done by a translator, whose salary would be paid using part of the approved budget. During these months, however, it appeared clear that such strategy would not render the expected results. The toolkit documentation has a heavy technical component, and its translation would require that the person in charge of it was familiar with biodiversity data processes, data quality concepts, biodiversity data standards and the toolkit itself. Employing someone without any one of those skills would require us to double-check all the documentation for correctness, and therefore would render the whole process more time and resources consuming. Therefore, it was decided that the best strategy would be for us, who are already familiar with all those concepts, to translate the documents ourselves. This has proven quite efficient, and much documentation has already been produced and translated and is broadly available. The section of the budget originally destined to translations is to be reallocated to support undergraduate students internships instead. It is expected that two biology students in their last semester will be dedicated to data quality validation using the data migrator toolkit on the data published through SiB Colombia, generating data quality reports for the publishers within the network. This change in the the original plan not only avoids wasting resources, but adds value to the project by directly involving the next generation of biodiversity data managers and users, therefore broadening its overall impact.