Report: CESP 2017 / SiB Colombia - VertNet Workshop

Compartiendo experiencias y herramientas en calidad de datos sobre biodiversidad [Sharing biodiversity data quality experiences and tools] (11-15 September 2017)

Participants (all week *)

- **John Wieczorek** - VertNet (Information Architect - Museum of Vertebrate Zoology, UC Berkeley y Museum of Comparative Zoology, Harvard University.) *
- **Paula Zermoglio** - VertNet (Associated Researcher -Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA-CONICET), Universidad de Buenos Aires, Argentina y Université François Rabelais de Tours, Francia)*
- **Leonardo Buitrago** - SiB Colombia (Content Administration) *
- **Victoria Arciniegas** - SiB Colombia (Cooperation) *
- **Oscar Duque** - SiB Colombia (Information Technologies) *
- **Ricardo Bastidas** - Instituto Humboldt (Information Technologies) *
- **Iván González** - Instituto Humboldt (Biodiversity Indicators) *
- **Ricardo Ortíz** - Researcher, GIS and georeferencing specialist *
- **Dairo Escobar** - SiB Colombia (Coordinator)
- **Javier Gamboa** - SiB Colombia (Report and Synthesis)
- **Carolina Castro** - Instituto Humboldt (Infrastructure and data quality - PEM)*
- **Carlos DoNascimiento** - Instituto Humboldt (Fresh water Fish Collection Curator)
- **Kevin Borja** - Instituto Humboldt (Information Technologies / Biological Collections)

Goals

Main

To transfer VertNet's experience on the automatic improvement of biodiversity data quality on the publication workflow of SiB Colombia (SiB) and the Humboldt Institute (IAvH).

Specific

- Recognize and implement the "Darwin Core Data Migrator Toolkit" within the data quality process of SiB Colombia and the IAvH.
- Exchange experiences among VertNet, SiB Colombia and IAvH about the internal process and quality data tools that would allow optimization and improvement of biodiversity information that is published.
- Generate needed documentation about the "Darwin Core Data Migrator Toolkit" and its translation to Spanish, in order to facilitate the use of the toolkit by the participants and the data publishing institutions within the Spanish-speaking community, extending it to the proper GBIF nodes (in Spanish).
- Exchange experiences about the digitization process in biological collections and how to improve workflows in this context.

General Agenda

TIME	MON 11	TUE 12	WED 13	THU 14	FRI 15
8:00 – 9:00	Welcome and visit of the venue, Venado de Oro	Data *Migrator Toolkit* Start up: Example	*Data Migrator Toolkit* Working session	*Data Migrator Toolkit* Documentation processes	Biological collections workflow / I2D
9:00-10:00	Review of the agenda. SiB Colombia Workflow and context				Exchange of experiences about collections digitization processes
10:00-10:30	Break	Break	Break	Break	Break
10:30-11:30	VertNet workflow and context	Data *Migrator Toolkit* Start up: Example (continuation)	*Data Migrator Toolkit* Working session (continuation)	Exchange of experiences about data quality	Perspectives in workflows improvement - biological collections / I2D
11:30-12:30	Introduction and installation requirements of the *Data Migrator Toolkit*				Conclusions and closure biological collections workflow / I2D
12:30-14:00	Lunch	Lunch	Lunch	Lunch	Lunch
14:00-15:00	Installation *Data Migrator Toolkit*	→ Transfer to Universidad Javeriana for the talk	Vocabularies maintenance *Data Migrator Toolkit*	Perspectives in workflows improvement - data quality: - Kurator - Vocabularies maintenance	
15:00-16:00	General review of the components of the *Data Migrator Toolkit*	Talk: *"El bueno, el malo y el no tan lindo. ¿Cómo lidiar con datos de biodiversidad?"*	Feedback *Data Migrator Toolkit*	Evaluation and closure *Data Migrator Toolkit* and data quality	
16:00-17:00		Paula y John			

Report

General aspects of the workshop

During the workshop several shared documents were used, kept in a Google Drive shared folder: CARPETA CESP-SIB. This folder contains general documents, such as the agenda for the meeting, and different subfolders in which the documentation files (ready and in process), presentations, etc., were organized.

The resources used that are related to the migrator tool can be found in the following links:

- Migrator, GitHub repository: https://github.com/VertNet/toolkit
- Vocabularies, GitHub repository: https://github.com/tucotuco/DwCVocabs

Also, during the workshop it was agreed that a GitHub repository would be created as a communication channel among the SiB Colombia, the Humboldt Institute and VertNet, to consolidate the implementation process of the tool, documenting news and possible failures or improvements (issues) during the process: CESP-GBIF-SiB-VertNet.[1]

Based on the activities developed during the workshop and on what was agreed to carry on in the subsequent months, a tentative chronogram was established for accomplishing some of the activities.

Activities developed and outcomes achieved during the workshop

Monday 11

1) The **workflows and context** of SiB Colombia and VertNet were presented. Similarities and differences in the approaches used were discussed. Presentations: VertNet, SIB Colombia.

2) The **Migrator Toolkit** was presented, as well as its main components in a *demo* session.

3) The **Migrator was installed** in the computers of the attendees. To execute the migrators certain regional configuration and software (Access) configuration are necessary, as well as having installed Unix Utilities. The working computers were configured and the necessary steps were documented. The steps and corresponding explanations were captured in Spanish and are available in the document Pasos Versión ES, section "I. Configurando la computadora para hacer el trabajo". This document is the basis for the

[1] In this repository, issues are already being captured with the activities to be performed and the problems encountered in the migrator execution process.

translation into Spanish of the step by step use guide of the migrator. [2] The English version is available here Pasos Versión EN.
To test that the computers could actually execute the Migrator, VertNet provided a migrator that was already done, based on a real dataset.

Tuesday 12

1) The **practice of use of the Migrator** began using a dataset provided by the SIB Colombia. This dataset was simple in structure, so that the basic functions of the migrator could be easily demonstrated.

2) The **talk**: "El bueno, el malo y el no tan lindo. ¿Cómo lidiar con datos de biodiversidad?" took place at the Universidad Javeriana, and counted with around 30 attendees.

Wednesday 13

1) We continued the **practice of use of the Migrator**. The practice took most of the working day. From this practice, different modifications to the Migrator were proposed. Some of those modifications were incorporated during the workshop, while others were highlighted to be performed in the weeks following the workshop. One of the already incorporated changes is the reduction of the number of queries that need to be customized, given a rearrange in the macros used by the migrator.

2) **Vocabularies**: the subject was addressed during the previous and subsequent days in a saltatory, recurrent fashion. Different alternatives for vocabularies management were discussed, among which it was considered to merge the vocabularies used by the SIB and the Humboldt Institute with those utilized by VertNet, in order to create a common, single repository. It was agreed to make changes to the process of merging vocabularies from the migrator ("Merger") and that such changes were to be documented accordingly[3] . It was decided that the vocabularies topic was going to be revisited further on, in a broader context. In particular, it was agreed that the issue would be taken to the TDWG meeting in Ottawa this year (Oct 2017), during which several discussion instances are planned around vocabularies.[4] A meeting of the CESP SIB-VertNet group was planned post-TDWG to discuss the next steps referring to building and managing vocabularies. In the meantime, it was agreed that an evaluation of the current state of the vocabularies maintained by VertNet was to be performed, assessing potential needs for translation into Spanish of the standard terms.[5]

---

[2] During the workshop, but outside the stipulated agenda, the translation of this part of the document into English was performed, incorporating it to the step by step migrator use guide in English. Guía completa EN. The guides in English and Spanish are currently under review, and will be incorporated in the GitHub Migrator repository as soon as they are ready.
[3] These modifications and the corresponding documentation are actually in process.
[4] During this same meeting the Data Quality Interest Group meeting will take place, when it is planned to propose a new Task Group dedicated to the construction of controlled vocabularies, lead by P. Zermoglio.
[5] To date (Oct 2017) an evaluation of the vocabularies has been perfomed regardin the number of values per term which could potentially be translated into Spanish (available at: VN Vocabs Translation Scope). Next steps will be discussed in the next online meeting.

3) The toolkit **documentation processes** were discussed, identifying a) documents already in existence and complete in English and that need to be translated, b) documents in English that need to be reviewed previous to the translation, and c) new documentation nueva that needs to be generated from this project in both languages.
    a) ***Existing documents***
        i) ***Reports***: In the migrators there are reports that capture the results of the process and the changes introduced to the data. Among the Migrator's files there is a document that explains the structure of such reports. It was agreed during the workshop to add a Spanish version of the report description, as a structural part of the migrator file, for its further use by non-English speaking communities. The translation of that document was initiated before the development of this workshop. During the workshop, a first review of the translation was performed in a shared document.[6]
        ii) ***Migrators - general descriptive document for users***. The original English version of this document, available in GitHub, was reviewed (Version EN). During the workshop the translation was completed Version ES. Also, during the workshop but outside the agenda, explanatory graphs were created of the processes performed by the migrator, in Spanish version (graphic version ES).[7]
        iii) ***Migrators - detailed descriptive documents of the steps*** to follow to run a migrator (contains the computer configuration and the execution of the migrator). When the workshop was ended, two versions were available, a full one in English and one partially translated into Spanish. Pasos versión EN; Pasos versión ES. It was agreed that the English version would be reviewed, as changes would be incorporated to the migrator, and the later those changes were going to be incorporated in the Spanish version to complete the translation of that document.
    b) ***New documentation:*** During the workshop it was agreed that in the following months the following documentation would be produced:
        i) Documentation specifically referring to ***what each type of query/table does*** within the migrator (e.g., legacies).
        ii) ***Test dataset*** (fictitious data, to cover the types of errors that the migrator detects and corrects) and a migrator already executed on that test dataset, against which one could test the practical way of building/running a migrator. This documentation will be added to the general documentation of the toolkit.
        iii) ***Document for the general public***, explanatory about Migrators and data quality enhancement in general, to show the added value of using the toolkit.
        iv) ***Vocabularies management***. Documentation referring to the "Merger" tool. Given that this will be subjected to changes in the Merger, the production of the corresponding documentation will be dependant on the stabilization of the Merger as a process. Also, it is expected that

[6] After the end of the workshop, modifications were made to the migrator that included changes in the generated reports. Such changes were incorporated to the English and Spanish versions of the reports explanation, and pdf documents were generated that are available in GitHub: version EN, version ES.

[7] After the end of the workshop, the graphic representation of the migrator was translated, and is now available in the toolkit's Wiki in the GitHub repository.

broader discussions regarding the construction and management of vocabularies may constitute delaying factors in the production of this documentation.

v) ***Graphic examples for each type of data.*** Based on the general schemes of the migrator functions that were produced during the workshop (versión en ES), it was proposed to generate similar graphs to demonstrate the process for specific types of data (e.g., dates, taxonomy, geography, etc.).

Thursday 14

1) An exchange of **experiences on data quality** was carried out, during which the workflows to enhance data quality were showed, as well as different tools created by members of the team, both from SIB Colombia and from the Humboldt Institute. Among them, a data quality validator was presented, capable of detecting errors in datasets based on Google Spreadsheets, and a tool for species distribution modelling with taxonomic and geographic checks written in R, aside from protocols for locality georeferencing. The details of the tools and processes presented are shown in Table 1.

Table 1. Exchange of experiences on data quality: processes and tools presented during the workshop.

Tool or resource	Organization in charge	Presenter
SiB documents	SiB Colombia	Leonardo
SiB publication sheets	SiB Colombia	Leonardo
Publication Wiki	SiB Colombia	Leonardo
Data quality guides	SiB Colombia	Leonardo
T-Rex	SiB Colombia	Leonardo
Validations and controlled vocabularies	Instituto Humboldt	Carolina
Data quality validator	SiB Colombia	Oscar
Geo-validador	SiB Colombia	Leonardo
Georeferencing protocol	SiB Colombia/ Instituto Humboldt	Ricardo
Geo IAvH Tool	Instituto Humboldt	Iván
Biomodelos	Instituto Humboldt	Iván

2) The **Kurator toolkit** was presented in the context of the perspectives in enhancement of workflows and data quality. Presentation Kurator.

3) The **evaluation of the workshop and closure** of the section regarding the Migrator and the quality of the data was carried out. Advantages and disadvantages related with the use of the toolkit were identified, and are presented in Table 2.

Table 2. Evaluation of the Migrator Toolkit.

ADVANTAGES	DISADVANTAGES	SOLUTION
Complete. Data quality validation across the whole spectrum of DwC.	Complex.	1. Simplify tasks within the migrator. 2. Training.
Allows structuring original data with all DwC elements.	Structuring requires high level of detail and attention.	1. Have knowledge about the original data and about the terms in the DwC standard.
Particular sections for data quality enhancement can be implemented.	NOT a universal tool (e.g., does not include georeferencing, currently more apt for English speakers).	1. Improve the documentation. 2. Translate into other languages. 3. Incorporate vocabularies in other languages. 4. Use in combination with other tools.
Tool developed for administrators, given its complexity.	NOT a tool designed for publishers, given its complexities.	1. Use at the aggregator level. 2. Prior use of simpler data cleaning tools at the provider level. 3. Training.

Friday 15

1) The **biological collections workflow** / I2D from the Humboldt Institute were presented.

2) An **exchange of experiences regarding digitization** of biological collections and how to optimize its performance was carried out. Different alternatives were discussed and resources were shared (e.g., IRS report about biodiversity data sharing projects of the JRS Foundation; data digitization via crowdsourcing: Notes from Nature; presentation with estimated digitization rates; etc.)

Concluding Remarks

MIGRATOR
The Data Migrator Toolkit was presented and its installation and running were performed. During the implementation, necessary modifications to the toolkit were identified, many of which were performed during the working week. Others are to be implemented later on, based on an agenda that was set up for testing and documenting the changes in the GitHub repository that was established during the workshop.

CONTROLLED VOCABULARIES
The implementation and use of controlled vocabularies in Spanish within the toolkit, integrated with the already established ones in English, remains fundamental for the validations that the migrator performs. This is a topic that will be addressed and solved during the whole proyect implementation period.

DOCUMENTATION
A large part of the toolkit documentation was compiled and translated into Spanish before and during the workshop, and this not only allows the consolidation of the information regarding the Migrator in Spanish, but also helps to improve the original documentation in English.

EXPERIENCES EXCHANGE
The exchange of experiences in data quality that took place during the working week allowed identifying possible synergies and the implementation of modules that would make the migrator toolkit more robust, along with the integration and enhancement of already developed data quality tools.