FINAL ACTIVITY REPORT

Extending tools for improving biodiversity data quality to Spanish-speaking communities

Contents

1. Executive summary

The SiB Colombia's capacity enhancement in the final phase of this project include: a) Capacitation of two college interns in VertNet's data quality tools and its implementation to clean up 64 data sets selected by lower quality and regional importance; b) The enhancement of the Data Migrator Toolkit documented in GitHub, according with SiB Colombia's necessities; c) The review and finalization of the toolkit documentation translation to Spanish; d) Experience and knowledge exchange to explore future enhancements in the use of controlled vocabularies. As part of this report, there is an evaluation of the VertNet's quality tool implementation in SiB Colombia's quality workflow, over the cleaning datasets obtained and the quality reports generated.

2. Contact information

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

3. Project summary

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 until this project was implemented, validation and improvement of data quality in SiB Colombia was performed manually or using separate tools for each step of the process. These data cleaning processes were time and resources consuming, particularly for large datasets. We identified a need for improvement and automation of the data quality check mechanism, in a way that would also allow efficient feedback to the data providers.

VertNet has ample experience in development and automation of protocols and tools to improve data quality at different steps of the publication chain. One of such tools is the Darwin Core Data Migrator Toolkit, which performs automatic data quality checks and enhancements and provides reports, facilitating data quality improvement of occurrence records before publication.

This CESP project was a one year collaboration between SiB Colombia and VertNet, both GBIF Nodes, and its main goal was to advance the data quality assessment and improvement processes within the SiB data sharing workflow. In particular, it aimed at transferring knowledge from VertNet to SiB Colombia concerning the Migrator Toolkit, thus enabling more efficient data quality checks and reports at SiB, and at improving the associated documentation, including Spanish versions that could be broadly used by the Spanish-speaking community.

The project was developed in two phases. The first phase (detailed in the mid-term report) consisted of a) capacity transfer concerning the toolkit, b) improvement and translation of the documentation, and creation of new documentation when needed; while the second phase (detailed in this report) consisted of: c) test-usage of the toolkit and adjustments to the SiB Colombia workflow, d) implementation of the toolkit on a series of data sets identified. During both phases we could achieve the expected outcomes, effectively implementing the toolkit into the SiB workflow and documenting the whole process. As a result of the collaboration, changes were implemented on the tool which rendered it more powerful for data cleaning across a broader range of data sets. We

also have been able to provide some insights into the time and effort needed for such implementation, which will be helpful for future implementations.

3.1. Activities completed

We report here the activities performed during the second half of the project. The activities completed in the first half have been explained in detail in the mid-term report.

The second half of the project had as a main goal to implement the Data Migrator Toolkit within the SiB Colombia data quality workflow. It was planned and developed in 4 steps (1-4, below), with other related activities being carried on in parallel and as necessary (5-8):

1. Call for student internship for the first semester 2018.
2. Interns training on the use of the Migrator Toolkit.
3. Migrator Toolkit implementation.
4. Reporting and feedback to the providers, and re-publication of data sets.
5. Modifications to the Data Migrator Toolkit.
6. Documentation review, improvement and translation.
7. Vocabularies.
8. Additional data cleaning activities.

These steps include changes with respect to the activities originally included in the project proposal, which were duly consulted with and approved by GBIF (i.e., reallocation of resources to contract students). The activities and their outcomes are described in detail below. All activities were mainly carried on by SiB Colombia, with VertNet participating in a consultation role to attend technical issues that could arise during the implementation of the toolkit and to introduce the necessary changes in the toolkit itself.

1. ***Call for student internship***

For the implementation of the Data Migrator Toolkit, we opened a call for university undergraduates to take an internship at SiB Colombia. The requirements for the applicants were that they be at least in the 8° semester of Biology or other environmental related careers. These requirements contemplated that the activities to be developed by the interns within the program could be projected in the future within their own institutions, strengthening the links between SiB Colombia and those other institutions. The call contemplated an interview and a technical test and resulted in the selection of two candidates. These interns were incorporated in March 2018, when administrative procedures were completed between the Humboldt Institute and the respective Universities. For more details on the call, including a link to the technical test, see Appendage 1.

2. ***Interns training***

Interns were trained in the following topics:

1. Introduction to SiB Colombia: governing schema, organization, publication model and participation channels.
2. Introduction to the Darwin Core standard and to data quality.
3. Introduction to the Data Migrator Toolkit.
4. Other data quality tools: OpenRefine, geographic validator, GBIF Data Validator.

For the Data Migrator Toolkit training, we replicated the workshop that was carried on at the beginning of this project (see mid-term report), but this time led by SiB Colombia members who had been trained in the previous edition. The basic documentation used was the Spanish version of the Data Migrator Toolkit Use Guide. Interns were made familiar with the use of Microsoft Access (needed for the toolkit) and with the Migrator Toolkit Workflow, using dataset that had been previously migrated. Also, students were trained in the use of GitHub as a tool to register and resolve issues concerning the migrator functioning, using the repository created for this purpose in the first half of the project. After training, a selected data set was migrated, in order to test the students understanding of the mechanism and to make sure they could resolve the common errors that could occur during the migration process, resolve the appropriate vocabularies and generate the reports and final publication files.

3. Data Migrator Toolkit implementation

For the implementation of the Data Migrator Toolkit, we selected data sets with known data quality problems and that had a thematic context, such as belonging to a particular region. We selected two publishers from the Colombian Pacific Region that have 64 resources, of which 57 are published through the IPT of the Instituto de Investigación Ambiental del Pacífico - IIAP, and 7 belong to the Universidad Tecnológica de Chocó - UTCH and are published through the SiB Colombia IPT (Table 1).

Table 1. Resources processed using the Data Migrator Toolkit.

Provider	**# resources**	**# records**
Universidad Tecnológica del Chocó	7	11566
Instituto de Investigaciones Ambientales del Pacífico	57	2573
Total	**64**	**14139**

The data cleaning process consisted of two phases (Figure 1), the first one focused on the results from the migrator toolkit, and the second one consisting of complementing the results and adjusting the cleaning process beyond the known limitations of the toolkit using other tools, such as OpenRefine.

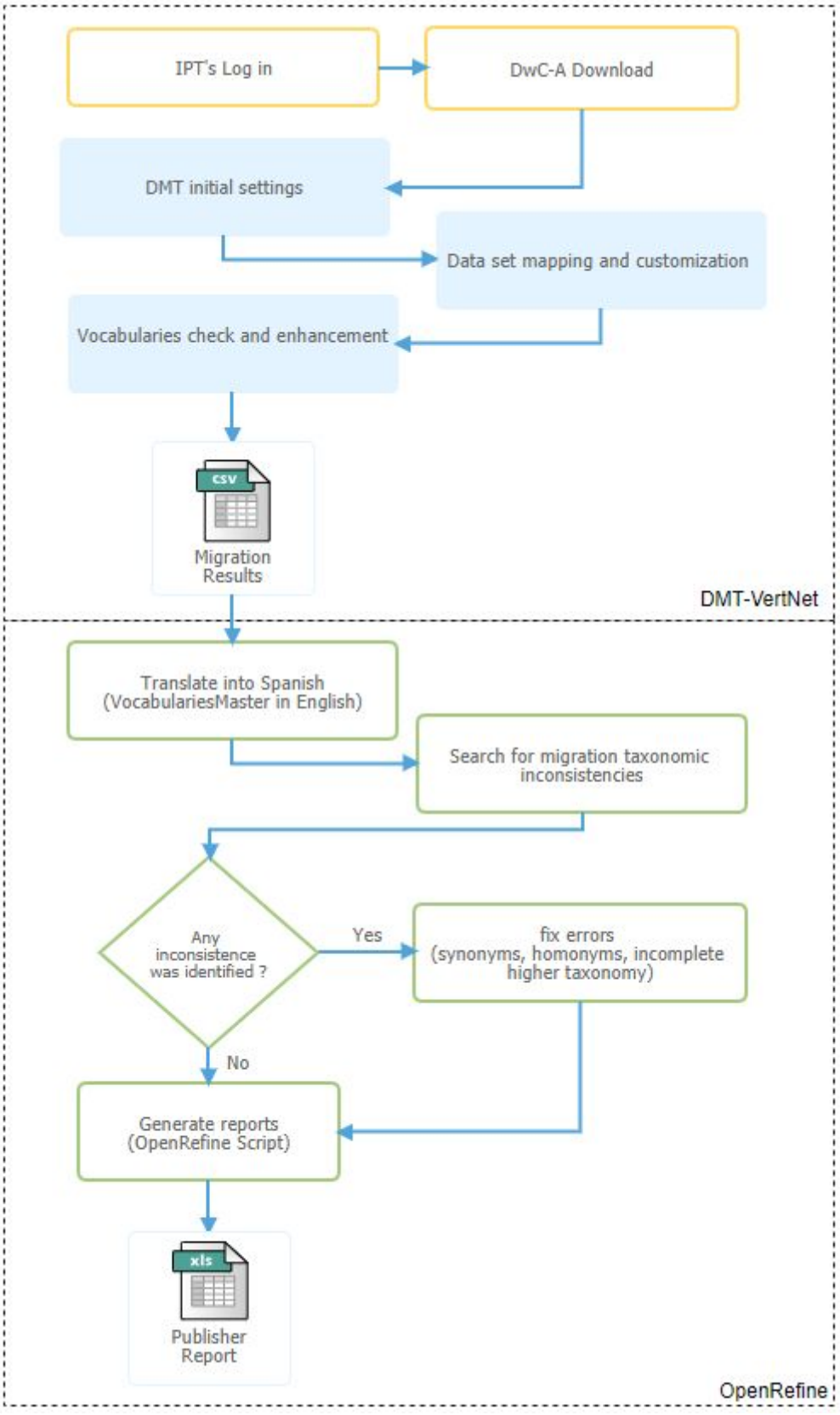


Figure 1. Data cleaning workflow implemented using the Data Migrator Toolkit (DMT) and OpenRefine.

Once the migration process was completed on the data sets, revisions were performed to ensure the necessary changes had been introduced and that no inconsistencies were found given the particularities of each data set. In doing so, we identified 17 data sets for which taxonomic issues could not be readily resolved using the migrator toolkit. These errors were mainly due to changes in homonyms among *Aves*, *Insecta* and *Plantae*, and were documented in the GitHub repository and were discussed with the VertNet team. We agreed on an alternative solution to these issues using taxonomic reconciliation in OpenRefine.

Several changes were introduced to the records that were processed using the Data Migrator. Considering the terms selected for reports, we constructed a graph to represent the percentage of records for which data quality was improved through the migration process (Figure 2). The most frequent change was the incorporation of attribution license from the information provided in the resources metadata. Terms such as institutionCode, institutionID, collectionCode and collectionID were standardized for all resources published by the same organization, which will render easier searches in the data portals.

Regarding taxonomic information, significant changes were made in terms spanning from class to genus. Errors in scientific names were corrected for 2687 records (19% of the migrated records). These changes will contribute to reducing the error associated with the estimation of the number of species recorded and published through SiB Colombia.

Regarding geographic information, the process allowed to a) correct for format errors in the coordinates, mainly due to language use of punctuation, and b) resolve inconsistencies in the stateProvince and county fields due to misspellings or use of currently non-official names for these geographic entities.

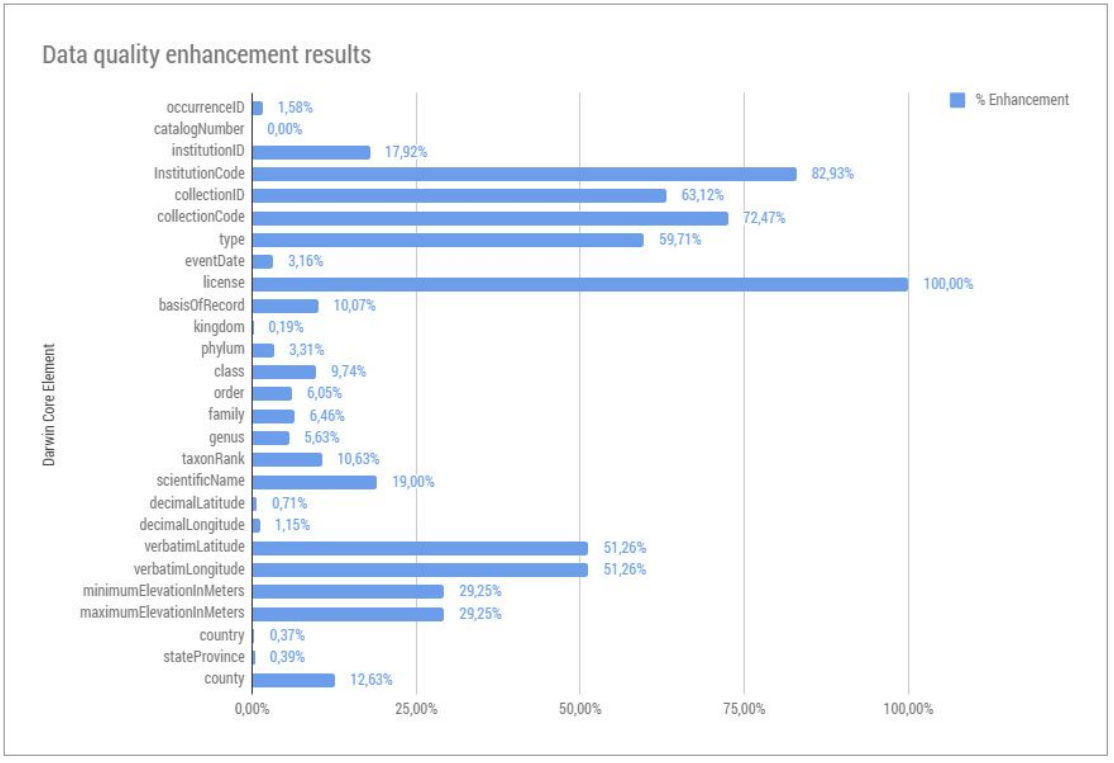


Figure 2. Percentage of records for which data quality was enhanced with the DMT-VertNet, for each Darwin Core element.

As part of the process, we identified 7 data sets that had been published but which had not been registered in GBIF, as IPT administrators from IIAP had omitted this step. These data sets will be registered and re-published with the data quality enhancements incorporated.

In order to evaluate the efficiency of the implementation of the Data Migrator Toolkit, we estimated the time needed for the evaluation of the data quality of the 64 data sets from the selected providers, during a period of 45 days (second week of March to the third week of April 2018, Table 1). During this period we documented the problems found during the process as issues in the CESP-VertNet-SiB GitHub repository (https://github.com/SIB-Colombia/CESP-GBIF-SiB-VertNet). The estimation of the number of datasets migrated per week is presented in Figure 3. The patterns observed were as expected, with longer times needed at the beginning both due to familiarization with the tool and provided that the first data sets to be migrated were those from UTCH, which contained larger amounts of records and hence needed more processing time. When IIAP data sets were migrated we observed a sharp increase in the learning curve slope. This might be related with the datasets containing fewer records (less processing times) and them already having fields named after Darwin Core terms, which made the mapping process much easier.

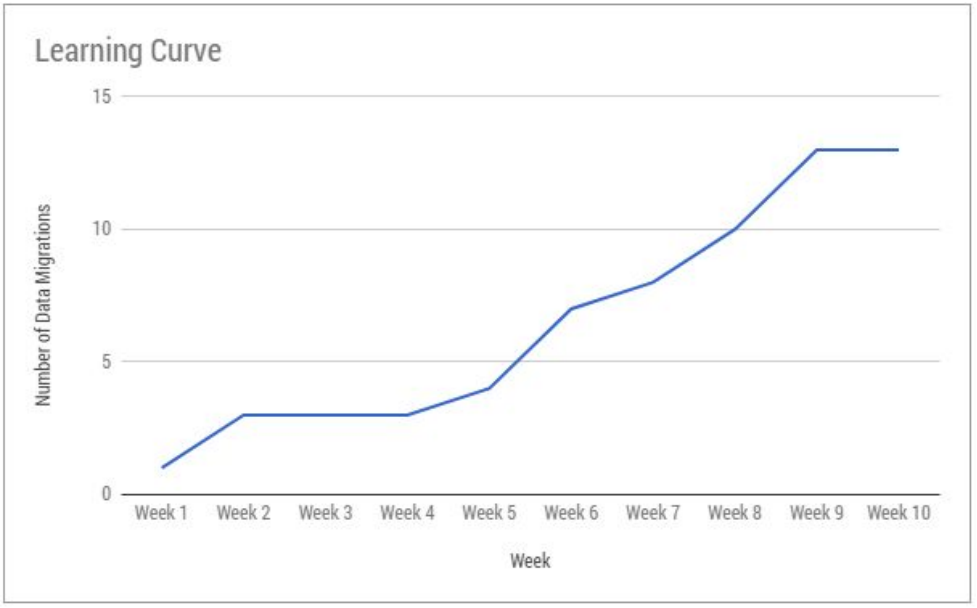


Figure 3. Estimation of the learning curve for the use of the Data Migrator Toolkit.

4. ***Reporting and feedback to the providers, and re-publication of data sets.***

For updating the resources already published through IPT with the data quality enhancements we decided to adopt the VertNet model. The recommendation for best practices of this model is to notify the providers of the changes suggested after the migration process, providing the corresponding reports, and letting the providers review the changes before implementing them in the re-publication of the data. The toolkit generates around 30 distinct reports, which include empty reports for those tests in which no potential improvements to data quality were identified. Considering that this might confuse the providers and in order to speed up the authorization for the changes, a new report model was adopted, focused on prioritized elements that would provide added value to data interpretation. The reports from the migrator were adapted to OpenRefine in a

way that the original data sets could be crossed with the migrated data over a set of 26 Darwin Core terms (Table 2).

Table 2. List of Darwin Core terms used to generate the reports for the data providers.

Darwin Core element	Darwin Core element
basisOfRecord	minimumDepthInMeters
type	maximumDepthInMeters
institutionCode	verbatimCoordinates
institutionID	decimalLatitude
collectionCode	decimalLongitude
collectionID	scientificName
license	kingdom
eventDate	phylum
country	class
stateProvince	order
county	family
minimumElevationInMeters	genus
maximumElevationInMeters	taxonRank

For each of the 26 prioritized terms the new report model contains a column with the occurrenceID of the records, a column corresponding to the migrated element, a column corresponding to the original element, and a column with a boolean result of the validation, where 1 indicates that the original and the migrated value are the same, and 0 indicates that the migrated value contains a change with respect to the original value.

The reports were sent to the providers together with the migrated data sets. We expect to receive their feedback in the incoming months, given the number of data sets/records involved.

5. Modifications to the Data Migrator Toolkit

Based on the issues identified during the migration process on SiB data sets, changes were made to the migrator toolkit itself. Some of the most significant modifications introduced are listed below, but all changes can be tracked in detail in the toolkit GitHub repository (https://github.com/VertNet/toolkit/commits/master).

- Addition of scientificName authorship to the “Update Taxon2 - scientificName” query and erase condition scientificName is Null.
- Change the “Update Location - higherGeography key” query to keep higher geography. This avoids duplicating values in higher geography, which potentially had inconsistencies.
- Correction of the call to the Aggregator script in the “Append Plants to SimpleDwCForIPT” query.
- Addition of a step to automatically construct bibliographicCitation.
- Added processing for dates with the pattern yyyy-mm-.
- Added three missing reports.
- Fixed query "ReportSummary - coordinateUncertainties out of range".

- Added processing of trailing single quote character (') in queries legacy coordinates processing.
- Modified PurgeNonprintingCharacters.sh to NOT remove spaces before double quotes. Spaces are not non-printing characters and removing them this way makes it very hard to find the real problems of non-printing characters.
- Amended macro to invoke query "Report - elevations reversed" correctly.
- Isolated modified queries to use one field only to avoid "resources exceeded" on large data sets.

6. Documentation review, improvement and translation

The documentation related to the Migrator Toolkit was mostly translated during the first half of the project (see mid-term report). However, during the implementation phase some adjustments were introduced to improve understanding by the data providers. Changes to the basic use guide also needed to be made following changes made on the toolkit (described above). All changes were incorporated as well in the Spanish versions of the documents. The following documentation was reviewed/updated:

1. **Migrator workflow explanation**. Adaptation of the original document (in English), adding a graphical component (https://github.com/VertNet/toolkit/wiki/Migrator-Workflow-(ES)), and translation into Spanish (https://github.com/VertNet/toolkit/wiki/Migrator-Workflow-(ES)).
2. **Migrator instructions for basic use**. This document was reformulated and translated into Spanish during the first half of the project, based on the original use document from VertNet. During implementation phase we introduced changes to both the English and Spanish versions (e.g., addition of step 15.1). The document is now an integral part of the Data Migrator Toolkit, and is downloaded together with the tool from the GitHub repository (https://github.com/VertNet/toolkit, see files: README_Instructions for use_EN, and README_Instrucciones de uso_ES).
3. **Migrator Data Quality Reports explanation.** This document aims at explaining the results of the migration process, which are captured in the reports, to the data providers. Adjustments were made to this explanation in order to facilitate interpretation of the reports, particularly in the Spanish version. This document is now incorporated as an integral part of the Data Migrator Toolkit (https://github.com/VertNet/toolkit/tree/master/reports).
4. **Data Migrator Toolkit internal functioning.** This document is an explanation of each of the tables, queries and macros included in the toolkit (https://docs.google.com/spreadsheets/d/1RiAcRosAm-IekXq_0_uKRBqx3NhCCE4OVasrZlc-ngk/edit#gid=763999206). This document was translated into Spanish and was incorporated to the toolkit documentation corpus in the GitHub repository. (https://drive.google.com/open?id=1atUOWANUI7n9KTzZz4XBhMN0iXT7Hmrpd0scmEaIEOk)

7. Vocabularies

The migrator toolkit utilizes controlled vocabularies already available, which greatly reduce the time needed for data quality revision, allowing for automatic replacement of values for known variations (e.g., M to Male, Espécimen Preservado to PreservedSpecimen). However, there exist

limitations regarding the vocabularies, especially considering language differences. Although the migrator identifies and corrects the values correctly in both languages, the new, standardized values introduced are invariably in English. This generates the need to feed the Master Vocabularies files in the migrators with Spanish terms. Since the construction of multilingual vocabularies does not currently have a standardized mechanism, during the migration implementation, the values for some terms were translated outside the migrator. Some of the terms for which outside translations were performed are establishmentMeans, occurrenceStatus, sex, lifeStage, geodeticDatum (e.g., forced to WGS84), preparations, taxonRank, etc. This constituted an additional process performed in OpenRefine, but greatly facilitated by the Data Migrator, as the values had already been standardized to a few values. Ideally, a Master Vocabulary in Spanish would be constructed to incorporate directly in the migration process.

8. Additional data cleaning activities

As an additional activity to implement the data quality workflow, an exploratory quality diagnostic was performed on the datasets published through the CR-SiB IPT (http://ipt.biodiversidad.co/cr-sib/). This is an IPT modified by SiB Colombia to generate certificates of permission to collect specimens in Colombia in order to comply with current internal laws. Each one of this datasets contains usually 41 Darwin Core occurrence elements with event, location, and taxonomic information.

The results of this diagnosis showed the completeness and general data quality of 498.718 occurrence records published between 2014 and 2016. The data cleaning process was performed using the tools and data quality workflow acquired through the transfer of VertNet's capacities during the CESP project. In that order, 821 quality reports were generated and provided to notify at CR-SiB publishers and so improve the data quality for each resource.

3.2. Post-project activities

After the completion of the project, we plan to continue to collaborate and broaden the implementation of the Migrator Toolkit. The related activities we have planned are:

- Follow-up on feedback and reports from the institutions whose data sets have been migrated and re-publication of the clean data sets to SiB and GBIF.
- Implementation of the Migrator Toolkit on more data sets that are shared through SiB Colombia. We will perform this based on a set of priorities that we are currently establishing, which include regional, taxonomic and taxon status criteria.
- Plans are currently being outlined for broader communication of the data quality workflow applied at SiB to the institutions and organizations the provide data through the SiB.
- Continue the collaboration between SiB Colombia and VertNet to improve the tool and its implementation. We will continue to register and resolve issues found in the GitHub repositories (the project one and the toolkit one). This might include both troubleshooting and changes to the tool.
- Join efforts to build Spanish versions of the available controlled vocabularies. This will be framed within a broader context in the community, following efforts currently being taken by the Biodiversity Information Standards (TDWG) organization, and in coordination with other Spanish-speaking countries in the region who have already shown interest in participating in such efforts.

- Explore the possibility of expanding the use of the Migrator Toolkit or its components, many of which are already incorporated as Kurator actors and workflows (http://kurator.acis.ufl.edu/kurator-web/) within the region. Conversations about this point are currently being held with colleagues in several countries in the region.

4. Project objectives

This project was a one year collaboration between SiB Colombia and VertNet, both GBIF Nodes, and its main goal was to advance the data quality assessment and improvement processes within the SiB Colombia data sharing workflow. In particular, we aimed at transferring capacity from VertNet to SiB Colombia concerning the use of the Darwin Core Data Migrator Toolkit, which allows data quality assessment and enhancement. The deliverables expected from this projects were: a) the implementation of the Data Migrator Toolkit in the SiB workflow, and b) the translation and improvement of the documentation in order to facilitate the use of the toolkit within the Spanish-speaking community. The activities developed throughout the project were specifically directed towards attaining our goal and allowed us to achieve the expected outcomes (see Activities section).

Aside from VertNet and SiB Colombia, other interested parties were directly involved or were benefited by implementation and the results of this project. The migrator toolkit will be implemented in the data sets provided by some Colombian organizations who are already SiB data providers. These include the Humboldt Institute, who will provide feedback reports from the data cleaning process, particularly for the fish collection, as well as the Instituto de Investigaciones Ambientales del Pacífico (IIAP) and the Universidad Tecnológica del Chocó (UTCH). In this sense, we have established as additional objectives to disseminate the use of the toolkit results and to foster feedback from the different institutions which publish data through the SiB by improving the communication flow between parties and organizing an information dissemination plan. These objectives, although arising from this project, are mid- to long-term objectives, which are starting to be tackled by the project members.

Although not a direct objective of this project, its results are already contributing to achieving the milestones of the “Improve Data Quality” priority of the GBIF Strategic Plan 2017-2021, by helping to improve the data quality of occurrence records within the Spanish-speaking community. Also, at a national level, the results are contributing to support the SiB Colombia Technical Secretariat Annual Operation Plan 2017.

5. Project deliverables

The deliverables expected from this project were the implementation of the Data Migrator Toolkit within the SiB Colombia workflow and the improvement and translation into Spanish of the corresponding documentation. We were able to provide both deliverables following the activities explained in the Activities section. These include:

- Darwin Core Migrator Toolkit incorporated to the SiB Colombia workflow.
- Datasets re-published after data cleaning with the toolkit.

- ***UTCH data sets DOI's.*** *This data sets was already published with the data quality enhancement. And the quality reports was sent to the publishers.*

#	Data set Name	# Records	DOI
1	Colección del Herbario de la UTCH	4373	http://doi.org/10.15472/viv3ue
2	Colección Científica de Referencia Zoológica ...	2394	http://doi.org/10.15472/sz41lm
3	Colección Teriológica de la UTCH	1323	http://doi.org/10.15472/nhehct
4	Colección Entomológica de la UTCH	1156	http://doi.org/10.15472/rmbpsh
5	Colección Hidrobiológica del Chocó	1001	http://doi.org/10.15472/5tflih
6	Colección Limnológica del Chocó	1000	http://doi.org/10.15472/fzpiwc
7	Colección Ornitológica de la UTCH	319	http://doi.org/10.15472/xgbet6

- ***IIAP data sets DOI's.*** *This data is in process to be published with the quality enhancement, after a meeting with the IIAP's IPT Administrator scheduled for the third week of September. There are other 6 data sets not registered and without DOI assignment, but were cleaned and are ready to be publish.*

#	Data set Name	# Records	DOI
1	Caracterización de anfibios del Cerro de ...	141	http://doi.org/10.15472/nz9npw
2	Reptiles de un sector del cerro Tacarcuna ...	118	http://doi.org/10.15472/iaifof
3	Caracterización ecológica de reptiles present ...	111	http://doi.org/10.15472/tlbxc2
4	Caracterización de la ornitofauna del Alto del...	106	http://doi.org/10.15472/6cmtih
5	Comunidad de Peces un Área Degradada por ...	79	http://doi.org/10.15472/5mlsnq
6	Monitoreo de la avifauna presente en el ...	79	http://doi.org/10.15472/6jtc2b
7	Caracterización Florística del Páramo Tatamá	76	http://doi.org/10.15472/utj40k
8	Monitoreo de la flora presente en el páramo ...	69	http://doi.org/10.15472/ornrkk
9	Fauna Asociada a los Cativales del Municipio ...	68	http://doi.org/10.15472/wydg4q
10	Análisis ecológico de la ornitofauna de la ...	64	http://doi.org/10.15472/hpg9hm
11	Caracterización Ecológica de la Fauna del ...	64	http://doi.org/10.15472/v6cmhw
12	Caracterización de la Ornitofauna del Alto ...	55	http://doi.org/10.15472/r0zpwo
13	Anfibios presentes en las lagunas del mun ...	52	http://doi.org/10.15472/bqtnks
14	Caracterización vegetal de una zona de alta ...	52	http://doi.org/10.15472/vvokvd
15	Caracterización de la ornitofauna del Cerro ...	50	http://doi.org/10.15472/zyaxya
16	Caracterización Ecológica de la avifauna del ...	45	http://doi.org/10.15472/w0ktlb
17	Anfibios de un sector del cerro Tacarcuna, ...	41	http://doi.org/10.15472/znktkt
18	Caracterización de Reptiles del Cerro de ...	38	http://doi.org/10.15472/gkxzuy
19	Comunidad de Macroinvertebrados de un ...	37	http://doi.org/10.15472/xhcrtv
20	Caracterización Florística del Páramo de ..	32	http://doi.org/10.15472/kr6fut
21	Caracterización Ecológica de los Macroinv ...	31	http://doi.org/10.15472/vfzswq
22	Ornitofauna de un Área Degradada por ...	30	http://doi.org/10.15472/rx3iso
23	Caracterización Ecológica de los Mamíferos ...	29	http://doi.org/10.15472/6ukz9t
24	Caracterización Ecológica de la Mastofauna ...	29	http://doi.org/10.15472/fr2gym
25	Caracterización Ecológica de la Fauna del ...	29	http://doi.org/10.15472/p0hncg
26	Caracterización de Anfibios del Cerro Tacarcu ...	29	http://doi.org/10.15472/o84vrs
27	Mastofauna presente en el Alto del Buey ...	28	http://doi.org/10.15472/2aqllc
28	Caracterización de la fauna del páramo del ...	28	http://doi.org/10.15472/vjtirq
29	Caracterización de Reptiles del corredor ...	27	http://doi.org/10.15472/jomypn

30	Análisis ecológico de la mastofauna en la par ...	27	http://doi.org/10.15472/zm1htq
31	Caracterización de la avifauna presente en el ...	26	http://doi.org/10.15472/uvke6a
32	Macroinvertebrados de la ciénaga La Honda, ...	26	http://doi.org/10.15472/4on6vi
33	Caracterización de anfibios del corredor bio ...	26	http://doi.org/10.15472/36p97m
34	Caracterización ecológica de la comunidad ...	26	http://doi.org/10.15472/h3bpiq
35	Caracterización Ecológica de la Ornitofauna ...	26	http://doi.org/10.15472/7cf6c1
36	Comunidad de Herpetos Presentes en la Cién ...	24	http://doi.org/10.15472/5wem2r
37	Caracterización de la ictiofauna de la ciénaga ...	23	http://doi.org/10.15472/tovn1s
38	Ictiofauna Presente en la Ciénaga de Monta ...	22	http://doi.org/10.15472/g3vtqf
39	Análisis ecológico de la fauna en las fuente ...	22	http://doi.org/10.15472/tljx5i
40	Ictiofauna Asociada a los Cativales del Mun...	22	http://doi.org/10.15472/aukumf
41	Herpetofauna del Alto del Buey, Bahía Sola ...	22	http://doi.org/10.15472/flxa71
42	Mastofauna Presente en la Ciénaga de Monta..	22	http://doi.org/10.15472/qicotv
43	Mastofauna presentes en la ciénaga de Gui ...	21	http://doi.org/10.15472/7eplaw
44	Comunidad de Reptiles presentes en el Cerro ...	20	http://doi.org/10.15472/w9fpvg
45	Anfibios presentes en el Cerro Jánano	18	http://doi.org/10.15472/magebe
46	Herpetofauna del Alto Galápagos, San José ...	18	http://doi.org/10.15472/ijpswu
47	Herpetos de un Área Degradada por Activid ...	18	http://doi.org/10.15472/3asdrh
48	Caracterización Ecológica de la Mastofauna ...	17	http://doi.org/10.15472/pe96rs
49	Caracterización de la Íctiofauna presente en ...	12	http://doi.org/10.15472/js9iiq
50	Caracterizacion Ecologica de la Fauna de An...	10	http://doi.org/10.15472/i1m2lv
51	Mamíferos de un Área Degradada por Activid ...	8	http://doi.org/10.15468/8jdu8y

- Documentation (see details in the activities section):
    - ***Migrator workflow explanation*** (EN and ES). (https://github.com/VertNet/toolkit/wiki/Migrator-Workflow-(EN), (https://github.com/VertNet/toolkit/wiki/Migrator-Workflow-(EN)).
    - ***Migrator basic use guide*** (EN and ES). https://github.com/VertNet/toolkit/blob/master/README_Instructions%20for%20use_EN.pdf https://github.com/VertNet/toolkit/blob/master/README_Instrucciones%20de%20uso_ES.pdf.
    - Migrator Data Quality Reports explanation (EN and ES) (https://github.com/VertNet/toolkit/tree/master/reports).
    - Data Migrator Toolkit internal functioning. https://github.com/SIB-Colombia/CESP-GBIF-SiB-VertNet/blob/master/Migrator_Functioning_EN.png https://github.com/SIB-Colombia/CESP-GBIF-SiB-VertNet/blob/master/Migrator_Funcionamiento_ES.png

6. Project communications

During the second phase of the project communication among the project members was performed as listed below:

- **Monthly meetings**. Remote meetings in April to inform all participants on the advances in the implementation of the Data Migrator Toolkit within the SiB Colombia data quality workflow, and in May to inform and discuss final results and future steps. The agenda and notes from those meetings can be found at https://goo.gl/HbBywP (mostly in Spanish) and relevant topics were documented as issues in the GitHub repository.
- **GitHub repository**. We documented all issues related to the Data Migrator Implementation in the GitHub repository, as well as all modifications to the toolkit and recommendations regarding the workflow (https://github.com/SIB-Colombia/CESP-GBIF-SiB-VertNet). In this repository, we also keep all the documentation generated throughout the CESP project, including the improvements made to the reports and their translations into Spanish, and the use guides produced during the first phase of the project.

Communications to the broader community included/will include the following:

- 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) [reported in the mid-term report].
- Documentation in the GitHub repositories, detailed above.
- Document describing the data quality process within SiB Colombia, particularly using the Migrator Toolkit, and emphasizing the importance of biodiversity data quality, to be distributed among the data providers that publish their data through SiB Colombia, as well as distributed to the other GBIF Spanish-speaking Nodes. This document is currently under development.
- Talk at the V Colombian Symposium on Biodiversity Informatics, V Colombian Zoology Congress, 3-7 Dec 2018 (http://vccz.aczcolombia.org/informatica-biodiversidad/), about biodiversity informatics and how data quality has become a major concern within the community.

We have reviewed the project webpage (https://www.gbif.org/project/83348/extending-tools-for-improving-biodiversity-data-quality-to-spanish-speaking-communities) and, if possible, we would like to add a link to the Toolkit GitHub repository, and to point to the Spanish versions of the documentation.

7. Evaluation: lessons learned and best practices

Overall we consider that the project was successful, achieving its goals and being able to provide materials that can be reused within the broader community. Although we could automate many steps of the process to avoid errors in the migrator execution, we did identify some challenges. These are related to either the need for an improvement in the tool when faced with particular types of data sets that have not been processed before using the tool, or related to understanding some very specific processes within the migrator. Both of these require the technical support from

VertNet. Therefore, we conclude that although we were successful in learning and implementing the toolkit, it is highly probable that a constant support from VertNet would be needed.

As a result of the project, we have identified a series of advantages and disadvantages for the implementation of the Data Migrator Toolkit. In doing this we consulted the interns for their impressions and experience using the tool. The comments included observations both regarding the tool and the workflow. The advantages and disadvantages are listed below.

Advantages

- The migrator toolkit is very useful and organized to convert data sets to comply to the Darwin Core standard.
- Once its operation and basic steps are understood, the process can be more automated and the errors in the migration process are reduced.
- The step-by-step guides translated during this CESP project are effective for anyone with basic data quality notions to be able to run a migrator.
- The issues tracked in GitHub constitute vital documentation to transfer capacities regarding the tool, as errors tend to be repetitive and the issues become a guide to solve them.
- Automation of the data cleaning process using the migrator, from standardization to Darwin Core to the generation of detailed reports for the providers, has the potential to reduce considerably the time currently dedicated to accompany the data publication process.
- The data quality enhancements performed by the migrator are in line with the expected results from the tool regarding the completeness of the elements associated with the records, date formats, higher geography correction and use of controlled vocabularies for several fields.
- If the migration is being performed on several similar data sets (e.g., same institution, or same taxon), a migrator that has already been configured can be re-utilized, therefore greatly reducing the time needed for processing.
- Similarly, the migrator toolkit facilitates updating resources: if a copy of the migrator performed on a data set is saved, it can be re-utilized and therefore reduce the time needed for processing.
- The vocabularies used to check and improve the geographic information are complete at the county level and usually need no modifications to resolve them.
- The data cleaning allowed getting a better visualization of some rare species occurrence. For example, the “rana dardo dorada”, one of the most poisonous frogs in the world, it lives in the Colombian Pacific region. There is only one occurrence of this species in the datasets selected and his identification was mistaken in the higher taxonomy, changing the class from amphibia to reptilia. With the cleaning process, we are able to filter this species by his correct taxonomic group in our databases.

Disadvantages

- When data are already presented under Darwin Core terms, given that data providers are usually limited in the use of the standard, part of the capabilities of the migrators regarding mapping are underutilized, and only the data quality component of it is used.
- We identified a fault in the current vocabulary for ‘genus’, due to the migrator originally being oriented to vertebrates, where homonymy is less frequent. In SiB Colombia, given the diversity of biological groups it incorporates, the homonyms are much more frequent and generate errors in the migrator cleaning process, which are sometimes difficult to identify.

This is an opportunity to improve the tool, maybe building vocabularies particularly focused on each biological group.

- The vocabularies available to use when running the migrator tool are mainly in English, which means that the final result must be checked and eventually translated when appropriate. The alternative of making changes directly on the vocabularies of the migrator is now only a partial solution, as changes can not persist over time unless there is a commitment to periodically integrate the vocabularies generated by VertNet and by SiB.
- We identified a need to update the taxonomic vocabularies in accordance with local biological taxa and with taxonomic authorities/literature used for each of these particular groups.
- The validation of coordinates against higher geography must be done outside the migrator, as the tool does not include this function.
- The migrator learning curve is pronounced. It is necessary to have VertNet's support to resolve issues, as the documentation still cannot fully help resolve some of the particularities found in some data sets.
- The migration process for some data sets required the migrator to be run several times, mainly due to changes in the taxonomy.

During this project, some of the functions of the migrator and automated cleaning mechanisms were adapted to be performed using other tools that are incorporated in the SiB Colombia workflow, particularly OpenRefine. These new data cleaning mechanisms were tested in parallel with the Migrator and rendered good results, with a much less pronounced learning curve. However, the migration process, particularly for data sets that are not provided using the Darwin Core standard, keeps being of great value for the workflow of the Content Administration team at SiB Colombia.

As pointed out above, the use of controlled vocabularies in Spanish within the tool and integrated with the already established ones in English and other languages is of vital importance for the validations that the migrator performs, as it reduces the re-processing time needed to translate the corrections the migrator suggests into Spanish. Therefore, we conclude that the impact of the use of the migrator would be broader if the translation of the vocabularies could be solved. We are aware of efforts being carried out in the community to construct vocabularies in different languages. We will be participating actively in those fora to share our experiences and contribute to the initiative.

Currently, SiB Colombia is in the process of evaluating the data quality of resources published through the CR-SiB. These resources contain data mainly generated by providers in the private sector, who are required to publish their data to obtain a reporting certificate, which is legally required. These data sets contain important information that complements the data held in natural history collections. However, in order to be useful, they need to be checked for their quality. The Data Migrator Toolkit is the tool that allows testing the quality of these resources for their integration and visualization through the SiB Colombia.

8. Future plans and sustainability

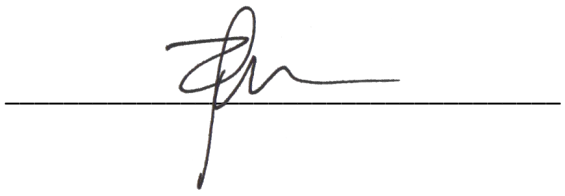
Future plans have been described in Section 3.2. Post-project activities. Some of these include:

- Continue the collaboration between SiB Colombia and VertNet to improve the tool and its implementation. We will continue to register and resolve issues found in the GitHub repositories (the project one and the toolkit one). This might include both troubleshooting and changes to the tool.
- Join efforts to build Spanish versions of the available controlled vocabularies. This will be framed within a broader context in the community, following efforts currently being taken by the Biodiversity Information Standards (TDWG) organization, and in coordination with other Spanish-speaking countries in the region who have already shown interest in participating in such efforts.
- Explore the possibility of expanding the use of the Migrator Toolkit or its components, many of which are already incorporated as Kurator actors and workflows (http://kurator.acis.ufl.edu/kurator-web/) within the region. Conversations about this point are currently being held with colleagues in several countries in the region.

The development of these activities will be dependent on the opportunities available to join efforts within the GBIF and the broader regional and global communities. All parties involved in this project are actively participating in different projects in the community and are actively seeking funding that would foster further collaboration.

9. Signature of the project main contact point

Signed on behalf of the project partners Date



14 SEP 2018

Appendage 1. Details about the interns call

For the implementation of the Data Migrator Toolkit we opened a call for university students to undertake an internship at SiB Colombia. The call was announced on December 21st, 2017 and closed on January 19th, 2018. The selection criteria included the following requirements:

- Be a student in the 8° semester or higher of Biology or related environmental sciences.
- Be a current student in a university holding a current agreement with the Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (Humboldt Institute).
- Be interested in learning biodiversity data and information management and implementation of different tools for improving data quality.

These requirements contemplate that the activity of the interns within the program could be projected within their institutions in the future, strengthening the links between those institutions and SiB Colombia.

As a result of the call, we received **86 applications** from students in different regions of the country, from more than 30 educational institutions.

Of the applications received, 6 candidates were preselected, who underwent a technical test (in Spanish) in which they had to familiarize with the Darwin Core Standard and with data cleaning principles. The test contained:

- Detailed instructions of the activity
- Darwin Core Quick Reference Guide
- Records template used by SiB Colombia (version 3.3.)
- Artificial test data set with 31 mammal records.

We conducted personal interviews with the 6 candidates on January 31st, 2018 after reviewing and scoring the technical test. Finally, **2 candidates** were selected taking into consideration the capabilities identifies during the interviews and the results of the technical test.

The interns were officially incorporated on March 6th, 2018 after administrative procedures were completed between the Humboldt Institute and the respective colombian universities.