Improving biodiversity data quality: a regional need from digitization to data repatriation

Programme: CESP
Project ID: CESP2020-018
Project page: https://www.gbif.org/project/5JJH6ZKCjztKrqT50OAadQ/improving-biodiversity-data-quality-in-latin-america-documenting-best-practices-across-data-workflows-and-life-cycles
Project lead organization: Sistema de Información sobre Biodiversidad de Colombia
GBIF Participant Represented: Colombia
Project implementation period: 30/11/2020 - 30/12/2022
Report approved: 30/11/2023

Narrative Final report

Executive Summary

We aimed as a project to address an identified need for the implementation of good biodiversity data quality practices in different steps of the data workflow. This was achieved through workshops and documentation focused on the Spanish-speaking community. By the time we write this report:

- All partners have created an inventory with existing materials related to data quality from international and local sources, including critical details for use in the development of the main document of this project. This inventory was consolidated using Zotero.
- Based on this compilation we built a draft for the document "Mejores prácticas en calidad de datos para digitalización, publicación y repatriación de datos de biodiversidad" that addresses all the needs identified in the biodiversity data quality processes.
- We organized the central workshop of this project (contents, agenda, format, dates, call, mentors, etc.). 60 participants of 14 different nationalities attended the online workshop.

Initially, the activities were planned to be face-to-face, but due to the pandemic, we were forced to run all of these activities online and change the lead organization of the project.

These conditions were challenging to ensure that all partners were aligned and working on time to fulfill all goals. However, was a good opportunity to reach a higher number of participants in the central workshop. Requesting an extension in the project period was necessary to have time to complete all the many time-consuming scheduled activities and deliverables.

The engagement and willingness for collaboration of partners during all of the project and tasks must be highlighted here as well. We have maintained a good level of communication between project partners to complete all commitments.

Report on Activities

Activity progress summary

Kick-off virtual meeting: A first meeting held in December (2020-12-18) was to kick off the project and set the roadmap. Here we agreed on the use of some basic elements for communication with all project partners, and the tools to document and monitor the progress of the project.

Gathering and assessment of existing materials: A second meeting held in February (2021-02-19) was to evaluate the progress in the gathering and evaluation of existing materials, using the Zotero platform. It was necessary to specify some additional agreements on how to standardize and categorize the bibliography in Zotero so that it can be easily used.

Documentation development in Spanish: Two meetings were held in April (2021-04-08 & 2021-04-19) to define the draft table of contents of the document "Mejores Prácticas en Calidad de Datos para Digitalización, Publicación y Repatriación de Datos de Biodiversidad", set the scope of each chapter and assign leader developers of each one. This activity presents a delay that started with some trauma generated at the family level in some project partners and later the impossibility of synchronizing agendas due to other commitments that arose, the impact on the project is low, considering that the

main project workshop was successfully completed. However, the project partners are committed to finalizing this post-project activity in the first quarter of 2023.

Documentation development in English: As this activity depends on the previous one, it is expected to be completed after the project in the first half of 2023. Its impact is even lower but remains within the agreed commitments.

First Workshop material preparation: In April 2021 we held a couple of working sessions to define the structure of the workshop, the methodology, and the necessary material to be prepared. Nine modules were defined, each one with an instructor selected according to the knowledge of each topic. SiB Colombia presented the terms of the call and selection of participants, as well as the visual identity of the workshop and its material. The material generated was shared freely and openly through different platforms (Github, Youtube, Google Drive).

Call and selection of first workshop candidates: Together with all the project partners, selection criteria were established to allow participants to have basic knowledge in the management and publication of data through GBIF. Additionally, a maximum of 60 places was determined, with a minimum of 10 places per organizing node, plus 20 for participants from other countries. If an organizing node was unable to fill its maximum quota, these would be distributed among the other organizing nodes with the highest demand. The call was open for 1 month and 160 applications were received from 14 different countries.

First Workshop "Publicación de datos de biodiversidad: cómo mejorar la calidad de los datos": The workshop was developed between November 8 and 16, 2021 through the Zoom platform with the support of six facilitators and a couple of other people supporting the logistical activities and the questions and answers.

Second Workshop "Improving biodiversity data quality: a regional need from digitization to data repatriation": This is an internal workshop with the project partners to work on finalizing the documentation, incorporating lessons learned from the application of the draft documentation during the first workshop. This workshop had the first part to take a balance of the previous workshop, but a second part is still pending to take a review of the best practices document. This activity presents a delay for the same reasons explained above. The impact on the project is low, considering that the main and first project workshop was successfully completed. The project partners are committed to finalizing this post-project activity in the first quarter of 2023, once the final version of the best practices document is available.

Self-assessment meetings: After the first workshop, a self-assessment meeting was held to determine the results of the workshop and to establish the next steps that would allow us to move forward with the best practices document and the certification of the participants. There it was also identified that the development of the content took more time than estimated and required extra effort on the part of the instructors to achieve the milestone. It is recommended that for future projects these activities have more preparation time or a greater number of instructors.

Completed activities

Activity name: Kick-off virtual meeting

Description: Online meeting with project partners to set the roadmap
Start Date - End Date: 18/12/2020 - 18/12/2020
Verification Sources: Minutes of the Kick-off meeting http://bit.ly/minutesGBIF-CESP-2020-018

Activity name: Gathering and assessment of existing materials

Description: Data Quality related materials were gathered from international and local sources and key information was extracted
Start Date - End Date: 19/2/2021 - 4/11/2021
Verification Sources: https://www.zotero.org/groups/2764305/cesp_2020-018/library

Activity name: Documentation development in Spanish

Description: Draft document in Spanish entitled "Buenas prácticas en calidad de datos para digitalización, publicación y repatriación de datos de biodiversidad", to be finalized in the first quarter of 2023
Start Date - End Date: 8/4/2021 - 31/3/2023
Verification Sources: https://bit.ly/DataQualitySpanishGBIF-CESP-2020-018

Activity name: First Workshop material preparation

Description: Planning of the central workshop of this project (contents, draft agenda, format, dates, mentors, etc.). Online and asynchronous working sessions were held to prepare and organize the workshop materials.
Start Date - End Date: 19/4/2021 - 4/11/2021

Verification Sources: https://bit.ly/firstWorkshopGBIF-CESP-2020-018

Activity name: Call and selection of first workshop candidates

Description: A maximum of 60 places was determined, with a minimum of 10 places per organizing node, plus 20 for participants from other countries
Start Date - End Date: 19/8/2021 - 4/10/2021
Verification Sources: https://bit.ly/WorkshopCandidatesSelectionGBIF-CESP-2020-018

Activity name: First Workshop "Publicación de datos de biodiversidad: cómo mejorar la calidad de los datos"

Description: A one week workshop was held online to train key stakeholders in the use of data quality tools and protocols. The workshop was in Spanish with 60 participants and 6 trainers.
Start Date - End Date: 8/11/2021 - 16/11/2021
Verification Sources: https://bit.ly/WorkshopCandidatesSelectionGBIF-CESP-2020-018

Activity name: Self-assessment meetings

Description: Self-assessment meeting was held after the central workshop to determine the advances and success of the project.
Start Date - End Date: 22/11/2021 - 26/11/2021
Verification Sources: http://bit.ly/minutesGBIF-CESP-2020-018

Report on Deliverables

Deliverables progress summary

Minutes of the Kick-off meeting: A first meeting held in December (2020-12-18) was to kick off the project and set the roadmap. Here we agreed on the use of some basic elements for communication with all project partners, and the tools to document and monitor the progress of the project. See the details at http://bit.ly/minutesGBIF-CESP-2020-018

Reports, shared working documents: Report, using the Zotero platform, with the consolidated bibliography about Data Quality from international and local sources and the key information extracted. It was necessary to specify some additional agreements on how to standardize and categorize the bibliography in Zotero so that it can be easily used. 35 items were consolidated in a Zotero group library as the initial bibliography for the construction of the document "Buenas prácticas en calidad de datos para digitalización, publicación y repatriación de datos de biodiversidad".

"Buenas prácticas en calidad de datos para digitalización, publicación y repatriación de datos de biodiversidad": Draf document, in Spanish, focused on best practice protocols for data quality. This deliverable presents a delay due to some trauma generated at the family level in some project partners and later the impossibility of synchronizing agendas due to other commitments that arose. However, the project partners are committed to finalizing this post-project activity in the first quarter of 2023.

Guidelines document in EN: As this deliverable depends on the previous one, it is expected to be completed after the project in the first half of 2023. The guidelines document in English remains within the agreed commitments.

Minutes and Guidelines document: About 10 online meetings were held with all project partners, along with other specific working sessions. Minutes with the report of the online sessions in http://bit.ly/minutesGBIF-CESP-2020-018

Workshop wrap-up, candidates selection process: Together with all the project partners, selection criteria were established to allow participants to have basic knowledge in the management and publication of data through GBIF. Additionally, a maximum of 60 places was determined, with a minimum of 10 places per organizing node, plus 20 for participants from other countries. The call was open for 1 month and 160 applications were received from 14 different countries.

First workshop materials (text, slides, exercises), workshop memories, social media links: This workshop, was held online and trained relevant stakeholders in the use of the data quality best practices developed. The candidate's selection process were target relevant actors that represent a wide range of stakeholders within the participant countries, and which show potential for the actual application of the knowledge acquired during the training in their home institutions. The material generated was shared freely and openly through different platforms (Github, Youtube, Google Drive).

Second workshop report. Document listed above under documentation development: This is an internal workshop with the project partners to work on finalizing the documentation, incorporating lessons learned from the application of the draft documentation during the first workshop. This deliverable presents a delay for the same reasons explained above. The project partners are committed to finalizing this post-project activity in the first quarter of 2023, once the final version of the

best practices document is available.

Minutes, self-assessment report: There it was identified that the development of the content took more time than estimated and required extra effort on the part of the instructors to achieve the milestone. It is recommended that for future projects these activities have more preparation time or a greater number of instructors. See the details at http://bit.ly/minutesGBIF-CESP-2020-018

Completed deliverables

Title: Minutes of the Kick-off meeting

Type: Documentation
Description: kick off the project and set the roadmap
Sources of verification: http://bit.ly/minutesGBIF-CESP-2020-018

Title: Reports, shared working documents

Type: Resources
Description: 35 items were consolidated in a Zotero group library as the initial bibliography for the construction of the document "Buenas prácticas en calidad de datos para digitalización, publicación y repatriación de datos de biodiversidad".
Sources of verification: https://www.zotero.org/groups/2764305/cesp_2020-018/library

Title: Buenas prácticas en calidad de datos para digitalización, publicación y repatriación de datos de biodiversidad

Type: Documentation
Description: Draft document, in Spanish, focused on best practice protocols for data quality.
Sources of verification: https://bit.ly/DataQualitySpanishGBIF-CESP-2020-018

Title: Guidelines document in EN

Type: Documentation
Description: This document will constitute a summary in English of the main document produced (ES). As this deliverable depends on the previous one, it is expected to be completed after the project in the first half of 2023. The guidelines document in English remains within the agreed commitments.
Sources of verification: In progress

Title: Minutes and Guidelines document

Type: Documentation
Description: About 10 online meetings were held with all project partners, along with other specific working sessions.
Sources of verification: http://bit.ly/minutesGBIF-CESP-2020-018

Title: Workshop wrap-up, candidates selection process

Type: Documentation
Description: A maximum of 60 places was determined, with a minimum of 10 places per organizing node, plus 20 for participants from other countries. The call was open for 1 month and 160 applications were received from 14 different countries.
Sources of verification: https://bit.ly/WorkshopCandidatesSelectionGBIF-CESP-2020-018

Title: First workshop materials (text, slides, exercises), workshop memories, social media links

Type: Regional Event
Description: The material generated was shared freely and openly through different platforms (Github, Youtube, Google Drive)
Sources of verification: https://bit.ly/firstWorkshopGBIF-CESP-2020-018

Title: Second workshop report. Document listed above under documentation development

Type: Training Workshop
Description: Internal workshop with the project partners to work on finalizing the documentation, incorporating lessons learned from the application of the draft documentation during the first workshop.

It is expected to be completed after the project in the first half of 2023
Sources of verification: In progress

Title: Minutes, self-assessment report

Type: Documentation
Description: About 10 online meetings were held with all project partners, along with other specific working sessions.
Sources of verification: http://bit.ly/minutesGBIF-CESP-2020-018

Events

Taller Calidad de datos + biodiversidad

Dates: 8/11/2021 - 16/11/2021
Organizing institution: SiB Colombia, VertNet, GBIF Chile, GBIF México, GBIF Argentina
Country: Online
Number of participants: 60
Comments: Central workshop of the project
Website or sources of verification: https://biodiversidad.co/comunidad/formacion/taller-CESP-DQ-2021/

Communications and visibility

All materials related to the workshop were uploaded to and openly available from a GitHub repository that was created for the project.
https://github.com/SIB-Colombia/CESP2020-018

The project webpage available at GBIF was translated into spanish in order to promote the project in the LA community:
https://www.gbif.org/es/project/5JJH6ZKCjztKrqT50OAadQ/mejorando-la-calidad-de-datos-sobre-biodiversidad-en-latinoamerica-documentacion-de-buenas-practicas-para-los-flujos-de-trabajo-y-ciclos-de-vida-de-los-datos

The project also had another alternate access point on the SiB Colombia website:
https://biodiversidad.co/comunidad/proyectos/cesp/cesp2020-018/

The central workshop also had a landing page to maximize its visibility at the time of the call and to consolidate the material: https://biodiversidad.co/comunidad/formacion/taller-CESP-DQ-2021/

Both the call and the material have been widely shared through social networks, mailing, and promotional events of GBIF actions in Latin American nodes. Some examples:
https://twitter.com/search?q=Taller%20Calidad%20de%20datos%20%2B%20biodiversidad&src=typed_query&f=top

https://mailchi.mp/4102f43b2e74/taller-calidad-de-datos-biodiversidad?e=242fd34aab

The playlist of the workshop recordings has had more than 170 views on Youtube. The most viewed video corresponds to the topic of digitization and data storage (206), followed by data quality principles: https://www.youtube.com/playlist?list=PL_KzX3Rxxwb0a-Drp49D314hULds23FWC

Monitoring and evaluation

Monitoring, evaluation and lessons learned
Final Evaluation:
The project activities are implemented 90% by the time we write this report. Some deliverables have been released close to the end of the project timeline (e.g. documentation best practices on data quality), so more time is needed to better evaluate the impact of those outputs. So far we think activities and deliverables have been effective as numbers show: 60 participants from 14 countries were trained in the workshop, the contents of the workshop have been reproduced more than 200 times and have served as input for some nodes to reuse the material as a reference for their data publishers.

The initial evaluation was completed documenting the current status of data quality within each participant node. In this evaluation we identified challenges to data quality assessment and enhancement in three different dimensions: where (location), what (taxonomy), and when (dates).

To monitor the progress of project activities, we use a Trello board that is shared with all project partners. https://trello.com/b/XXCt4Xf3, afterward, we continued to follow up through Google docs.

The project has addressed the need for the implementation of biodiversity data quality best practices and thus contributing to the objectives of the CESP goals. The workshop, with a regional component, allowed the building of the capacity of strategic people from different countries involved in mobilizing data through GBIF. Once the workshop was over, participants' comments were positive, such as Karla from the Autonomous University of Baja California Sur: "I enjoyed and learned a lot in the workshop so I will be looking forward to the next events."

We will follow up on the impact of the documentation we are finalizing on good data quality practices adapted to a regional context and thus contribute to another of CESP's objectives.

Best Practices and Lessons Learned:
One of the main lessons learned is that most of the problems identified in the participants' cases are associated with the underutilization of the Darwin Core standard for data publication and the lack of institutional protocols for the management of the information that is stored before publication through GBIF. The community has a high interest in implementing good practices and incorporating tools such as those used by the nodes to validate data quality. We recommend that the nodes increase their efforts to share with publishers and information managers guidelines on how to use the tools shared by the GBIF network for the organization and validation of data in the early steps of the data flow.

On the other hand, it was satisfactory to find that the contents previously generated by the GBIF network on this topic have been of great help to many of the participants. We recommend generating more spaces on "how to use" each of the existing tools or protocols on "how to report an error", for example: Open Refine, Data Validator, among others.

Post Project Activity(ies):
We are committed to finalizing the documentation with good practices in data quality and to holding a workshop among the project partners to generate a balance of this deliverable and how we can work as a region for its appropriation by the Spanish-speaking community.

In the meantime, all nodes remain committed to constantly generating a culture of best practices in their national communities to reduce the volume of inconsistencies found in the published data.

Sustainability plans

This project has been a unique opportunity to exchange best practice experiences among the participating nodes and learn from each other to improve national processes. These best practices were shared in the regional workshop and are being documented in the best practices document for the Spanish-speaking community. The contents generated to date are already being reused totally or partially in different training spaces at the local level and in the support provided to the publishing organizations. Data quality is an important issue for the entire region and we are sure that this material will be key to continue building a community of good practices from the early stages of data flow. In the case of Colombia, its node constantly promotes the consumption of this material through the support provided to data publishers.

Impact of COVID-19 pandemic on project implementation

The COVID-19 pandemic is one of the most disruptive events the world has faced in its recent history. It has sparked an unprecedented public health and economic crisis. The project has not been immune to these dynamics and the pandemic has had a negative impact.

In the beginning, all the project partners were doing well on a personal and professional level, confident that this would be the scenario throughout the project. However, that changed over time and we underestimated the impacts of the pandemic on the well-being of people and their social groups. Having synchronous sessions became increasingly difficult and we had to implement changes in communications to move forward asynchronously. This further lengthened the proposed timelines for each milestone. We concentrated on the most important milestone, the first workshop, which was aimed at transferring knowledge to the Spanish-speaking community.

The commitment to the project on the part of all the partners remains high and we seek to fulfill the pending post-project activities.