2014 DATA QUALITY WORKSHOP FINAL ACTIVITY REPORT

Contents

1. Executive summary

The Data Quality Workshop has achieved to convene participants from 6 Ibero-American countries to develop the first regional workshop on biodiversity data quality. The workshop was held from November 18th to 21st of 2014 in Bogota (Colombia). This four-day workshop covered the whole process of data generation, in a quality control environment at every step: data collection, handling prior to systemize, taxonomic identification, systemize of information, documentation, storage and preservation of information, publication and use of data.

Participants from different backgrounds and organizations (universities, NGOs, biological collections and environmental consultants) highlighted the importance of consolidate a continuous program in biodiversity data quality. They consider that data quality is a key topic in order to share useful data for research and decision making in biodiversity. With this workshop, on one hand we certainly refute the old idea that data quality is an “irrelevant topic or is not a priority” for our data publishers. On the other hand, it was clear that issues as time-consuming and unknown tools are the main difficulties to include data quality processes in the policies, generation and publication of biodiversity data by the users.

Results after measuring the technical impact of the datasets that were subjects for data quality improvement, plus taking in count the good reception of the participants from feedback received, the way how data quality concerns were addressed during the workshops and the further activities, and the high interest of the participants of having workshop replications to their institutions and local regions. Give strong arguments to grade this workshop as a successful event.

Furthermore there is already planned a second version of this workshop, that will be held in Montevideo (Uruguay) in order to keep the training of the region on this matters, and especially to strengthen the capacities for the current consolidation of the Uruguay’s node.

2. Contact information

Name of the contact person	Daniel Amariles
Institutional affiliation	Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (IAvH)
Mailing address	Calle 28A # 15 - 09
City and country	Bogotá D.C., Colombia
E-mail	damariles@humboldt.org.co
Telephone	+57 1 3202767 ext 1154
Role(s) in this project	Network contact

3. Project summary

The apparent low quality in data and metadata published through the GBIF network reflect a lack of implementation of a quality processes data by the publishers. It is important to improve and strengthen these practices not only for data publication, but also for something much more inclusive. For this reason, this training aims to involve the whole process of data generation, in a quality control way in all the steps: data collection, handling prior to systemize, taxonomic identification, systemize of information, documentation, storage and preservation of information, publication and use of data.

We are confident that the implementation by the participants and their reproduction of the training increased the quality and fitness for use of the data currently published through the GBIF network. Additionally, it was an important opportunity to create a regional community around data quality processes to enable further capacity-building in the publishers of the nodes.

As a result, the total 143 people achieved by now, enhanced their skills and capacities in order to put into effect data quality practices in the short, medium and long term, not only to improve but to maintain the quality of information shared and published through the network.

Of course, the success of this goals achievement could be made with such quality if the conference wasn't being held by such experienced and schooled experts as the instructors and speakers that participated from the different nodes.

3.1. Activities completed

As scheduled, the Data Quality Workshop were held at Universidad de los Andes in Bogota, Colombia, from the 18th to 21st of November 2014, and 25 participants from 6 Ibero-American countries attended this event. The agenda included conferences headed by 3 speakers from Brazil, 3 from Colombia and 1 from Argentina; for the practical sessions there was the support of 10 instructors from Colombia, 1 from Spain and 1 from Mexico.

3.2. Ongoing and post-project activities

An Spanish version of the book *Principles of Data Quality* by Dr. Arthur Chapman will be developed in order to make available this important document of reference for hispanoamerican audiences. This task will be leaded by SiB Colombia, and will have special support from Prof. Antonio Saraiva and Danny Veléz, current Node Manager from Brazil in SIB BR, whom have participated in the development of the Portuguese version of the same document http://www.gbif.org/resource/80924.

Besides, as mentioned, it's an important activity that will be made within the "Regional capacity enhancement by setting up Uruguay's data portal" project, as stated in the approved proposal: "*a data quality workshop following the regional training activities (which will include both aspects of data quality and data publishing tools). This activity will strengthen the cooperation between Brazil, Canada, and Colombia by consolidating a conceptual and methodological framework for the assessment of biodiversity data quality".*

4. Project objectives

Regarding the objectives stated, there were interesting results with the efforts made to achieve them, some of the aspects to be highlighted are:

Objectives planned vs results:

- **Improve the skills and capabilities of the fellow nodes on data quality processes, to surpass their own constraints as a part of a community.**

  Assuring the assistance of 6 different countries of the regions was the first step in order to achieve this goal. Nevertheless, the most fruitful result was the interaction of having all these experience exchange in just one place, it really helped to set common deals on the way to address the DQ issues, and the opportunity to open spaces in order to keep the discussions.

  Besides, the interest of having a second version of this event is materialized in the proposal made from Uruguay in the GBIF Capacity Enhancement Call of 2015. That fortunately for the region it was approved to be supported from GBIF as well from the organizing nodes: Argentina, Brazil, Canadensys, Colombia and the host, Uruguay.

- **Consolidate an exchange of information, technologies and experience among fellow nodes in order to be implemented and conveyed for each node to the potential publishers of data.**

  At the regional level, there was a clear exchange of experiences, techniques applied and technologies developed. Countries as Brazil, Mexico, Spain and Colombia have a notable background on this matters, and opportunely it could be presented to the audience in the spaces settled in the agenda.

- **Sharing tools and experience to optimize the data quality workflow from data collection to publication.**

  This objective was achieved since the tools presented in the practical sessions were pretty well welcomed, as reflected in the feedback. And it apparently well learned and used correctly since it was notice a good impact to the data quality.

- **Consolidate a data quality community between nodes.**

  The participants verbally agreed in this consolidation, and to formalize this community there was created the forum "*Grupo Iberoamericano en calidad de datos*" in order to set a collaborative space to support the task of this community. Nonetheless, a strong effort to activate the participation in this matter is needed.

- **Increase the use and reuse of the data (fitness for use) that are currently published through the GBIF Network.**

  Sessions related to species distributions models not only revealed the importance of data quality for these kind of analysis, but also show the importance of the use of open data in order to address the issues of biodiversity research as it happens with the biomodelos platform and its community.

- **Improve documentation and publication of current data and metadata.**

  For the evaluated datasets in the measurement of impacts to the data quality it was stated that **"***Almost 100% of the metadata or at least a very conscientious description of the data was documented for all the resources, due to the subjective approach of the metadata quality a further analysis was not realized*"

- **Enhance the structural and semantic consistency of the published data.**

  At least for the evaluated datasets, there is an measurable evidence of this enhancement stated in Bogota DQ Workshop 2014: Impacts on Published Data document.

- **Maximize the integration and interoperability of information.**

  DarwinCore and PlinianCore were the only two standards considered in any approximation about handling biodiversity data. Of course, since the importance of the interoperability for data use and licensing were strongly stated, it was evident that the advantages of using a common way to document data is clear way to address this concern.

.

5. Project deliverables

Regarding the deliverables committed during, the current status of each one is the described as follow:

- **A best practice guide for data quality on biodiversity data, in order to put together documents, highlights, etc., after the training.**

  This document is currently in progress as described in section 3.2.

  Nonetheless, there are some presented aspects during the workshop that are worth to be highlighted since address critical concerns regarding this matter:

  - A set of strong theoretical foundations. As presented by Prof. Saraiva and A. Koch, from University of Sao Paulo, during the very first session of the workshop about Concepts and Trends Biodiversity Data Quality.

  - List of best practices can be established for six stages of the data: planning, collecting, digitalisation, quality control checking and publishing. As presented by Katia Cezón from GBIF Spain.

  - Basic and advanced tools that are recommended for data cleaning and data refining, and their use. As presented by the instructors of the practical sessions of the event.

- **Step by step tutorials with the key tools and protocols used in the training.**

  The document *Calidad de Datos: Guía de herramientas para mejorar los datos primarios de biodiversidad* is the compilation of the tools and practices used in the workshop; these are intended to facilitate the process of providing quality to the primary biodiversity data through different methodologies. It's expected that this text will serve as a basis for optimizing the processes around biodiversity data quality, using as a basis the Darwin Core standard.

  The guide is divided into different sections, ranging from managing of basic tools for structuring information such as Excel, to web resources for taxonomic and geographic data validation. In addition some web services that optimize the display of information (CartoDB, Vesper) are shown. The last part is a glimpse to some tools more robust and comprehensive as Open Refine or Darwin Test data.

  Each section is up-to-date and introduces the topic and an approach to the origin and purpose of each tool. In all there are cases of use and a step by step procedure, which allows the reader to have a more tangible approach to the tool. The goal is that each case of use eventually become useful to the very own data of each researcher or publisher, hopefully turning the data cleansing into a much simpler task.

- **A total of 143 assistants trained directly and indirectly after training replication all the organizations that publish through the fellow nodes.**

  Twentyfive (25) assistants from six (6) countries trained directly on *Bogotá Data Quality Workshop*. Additionally, 118 assistants participated on ten (10) replication workshops.

  A narrative description of the results of the replications already made and further replications thoughts, can be found in Bogota DQ Workshop 2014: Replication Results document.

- **A comprehensive site with presentations and courses available online.**

  ○ Currently all the presentations are available in the workshop repository, as well as the guides for the practical sessions. http://j.mp/CO_DQ_Workshop

  ○ Information related to the the event can be found on the SIB Colombia's web site http://www.sibcolombia.net/web/sib/taller-de-calidad-de-datos

  ○ It is also available in our YouTube Channel a playlist with 17 videos about the recorded presentations.

- **Information published through the GBIF network with increased structural and semantic consistency.**

  Please refer to the document Bogota DQ Workshop 2014: Impacts on Published Data.

- **Data resources with complete metadata in order to improve data interpretation.**

  Please refer to the document Bogota DQ Workshop 2014: Impacts on Published Data.

- **A forum hosted in the I3B infrastructure available to the community and focused in three main topics: future events, results of past events and a "new resources" centre to show new tools or updated workflows about data quality.**

  A comprehensive forum *Grupo Iberoamericano en calidad de datos* was created in Google Groups due to two main reasons: first, the I3B infrastructure will be shut down later this year and previous experiences with the forum tool weren't as successful as expected. Second, the use of Google Groups as a temporal platform in first instance will allow us to create a community, also in the short term will be easier to migrate the contents to a more robust and friendly interface (in developing), a kind of “help center for publishers” hosted in the SiB Colombia Infrastructure.

6. Evaluation: findings and conclusions

The impact of this event into the datasets evaluated, evidenced by the improvement of the “Apparent Quality Index” (ICA), demonstrate this kind of capacity enhancements activities do work in order to increase the quality of the biodiversity data and hence its fitness-for-use.

The strategy of replication training had a positive impact on SiB Colombia network. Students, teachers and professionals who manage data and information on biodiversity, found useful, applicable and necessary tools worked on the additional workshops on data quality. Regarding this, there are some findings and conclusions to be highlighted:

- Biodiversity primary data georeferencing was a common matter that was especially addressed in all the workshop, it means that besides it is an issue of big concern is also well considered to be attended.
- *Open Refine* is a featured as a useful tool for data management, since it is a very powerful tool. Most of the assistants didn’t know about it and its possible uses, so, presented this tool was very amaze.
- The data quality capacity building promotes the re-engineering in order to improve the processes of biodiversity data management.
- The compliance of the legal obligations for the organizations that has collection permissions, is a big concern to take in count.
- The capacity building in biology teachers is a good strategy since students from their courses can receive capacity as well for each semester.
- In two of the universities approached, CES and UIS, came out proposals regarding courses addressing the biodiversity data quality matters. It is expected to have this proposals as academic offers for students and biology teachers.

Additionally, knowing that there is a lack of resources in order to address all the work required to achieve the institutional, national and global goals for biodiversity management. The use of powerful tools are a shortcut to speed up the results regarding this matter. Moreover, the collaboration between the few resources that every node has to work for common outcomes, is an exercise of synergy that has to be done if these goals are wanted to be achieved on time.

7. Recommendations and lessons learned

- Extending this workshop to a regional level gave a really big improvement to the workshop content, as well to the impact it can make.
- This kind of workshops helps to identify new key actors that later can be doing a relevant work for the network.
- Commit from absolute all the partners is always hard to achieve. Nonetheless, the partners that indeed shows commitment, always made efforts that goes beyond the expected.
- It’s always important to identify new needs from the partners of the network.
- Personal meetings always led to a better understanding of the common problems, issues solving and partnership strength.

8. Future plans

- Continue to develop and strengthen capacities of the network partners on data quality, mainly through educational tools and self-learning.
- Find a way to achieve to more universities in order to have more courses regarding data quality matters.
- Create a self-sustainable model of biodiversity informatics capacity building in which the workshops are fund by the organizations that has wild species collection permissions.
- Create data quality check equation, in order to publish an indicator number given by this calculation that suggest the level of quality of the datasets published.
- Support the plan and execution of the second version of this event.

9. Signature of the project main contact point

Daniel Amariles

20/08/2015

Signed on behalf of the project partners

Date