Bogota DQ Workshop 2014: Impacts on Published Data.

Improving primary biodiversity data shared and published through GBIF network

The main objective of the data quality workshop held in November 18-21 of 2014 (Bogotá, Colombia) was to improve the skills and capabilities on data quality processes that exist in the fellow nodes and Colombian data publishers, in order to surpass their own constraints. This document summarize and describes how the workshop impacted the Colombian data publishers in terms of new resources published through SiB Colombia and it´s improvements in data quality.

A quick overview:

New partner institutions

Our main partner institutions are Universities, NGO's and governmental institutions for environmental management. In the workshop we gain 3 new institutions, two of them are institutions that in the past weren't interested of been part of the community or didn't consider publish primary biodiversity data a priority :

Antea Group Colombia is a private company in civil engineering and environment for the energy sector that provides customer technical support for environmental licenses. Antea attended the workshop with the objective to go beyond reporting data to the environmental authorities, and improve their capture and digitization data procedures to ensure a better data quality.

The National Health Institute works to improve the health of the citizens through knowledge generation and monitoring of public health. They are currently re-organizing the entomology collection of insects of medical importance, they attended the workshop to set a new data quality baseline for the collection in order to take better management actions to control major public health vectors.

Data Quality assessment

To measure the success of the workshop, we evaluated the datasets published by the Colombian participant institutions after the workshop with a modification of the "Apparent Quality Index" (ICA) created by *GBIF.es*. This quality index was developed as an indicator of the database quality (records, observations, specimens of natural history collections) in Darwin Core. The original ICA formula was modified to fulfill the purpose of this document addressing the specific case of Colombian records. The index is divided in three components: 1) Taxonomy, 2) Geography and 3) Structure and semantics**.** We assigned different weights among the aspects of the three components accordingly to the tools taught in the workshop. The aspects that allow better judgment of the fitness for use of a dataset have more weight.

Quality Index Formula

Values are between 1 and 0, were a dataset with high fitness for use will have an ICA value closer to 1.

$$ICA = \frac{ICAt+ICAg+ICAs}{10000}$$

Taxonomy

$$ICAt = 25 \times t1 + 15 \times t2 + 2.5 \times t3 + 2.5 \times t4 = 45\% ICA$$

$t1$ Number of unique scientific names (dwc: kingdom, phylum, class, order family, genu, scientific name) correctly written. Validation was made following 2012 Catalog of life database and the Taxonomic Name Resolution Service.

$t2$ Number of unique scientific names (dwc) consistent with the rest of taxonomic elements documented.

$t3$ Number of records with the "identifiedBy" dwc element documented.

$t4$ Number of records identified below genus category.

Geography

$$ICAg = 10 \times g1 + 20 \times g2 + 10 \times g3 + 5 \times g4 = 45\% ICA$$

$g1$ Number of records with documented coordinates (dwc: decimalLatitude, decimalLongitude)

$g2$ Number of records whose geographic elements (dwc: stateProvince, county, municipality) are consistent with the coordinates.

$g3$ Number of elements documented accordingly to the point-radius method.

$g4$ Number of records with geographic names documented accordingly the Colombian political-administrative division.

Structure and semantics

$$ICAs = 10 \times s1 = 10\% ICA$$

$s1$ Number of elements documented accordingly to DwC, with the adequate format and controlled vocabulary.

Evaluated datasets

From the 15 Colombian participant institutions, so far seven have been published data through SiB Colombia. For those institutions we evaluated the latest published dataset, if they were former publishers we also evaluated the oldest dataset in order to have a valid evaluation of the the workshop. For a better comparison baseline, when possible, we avoided problematic taxa such as Insects which have several endemic or recently described species that are not yet reported in the major databases, we rather selected Birds and Plants. In Appendix 1 is the complete list of the datasets evaluated.

Results and Discussion

SiB Colombia has implemented several strategies in the past years to increase publishers capacities and improve the quality of data, these include workshops and continuous accompaniments before publishing through SiB. These efforts are reflected in the ICA calculations where most of the datasets and ICA components are rated over 0.7. **Figure 1**.

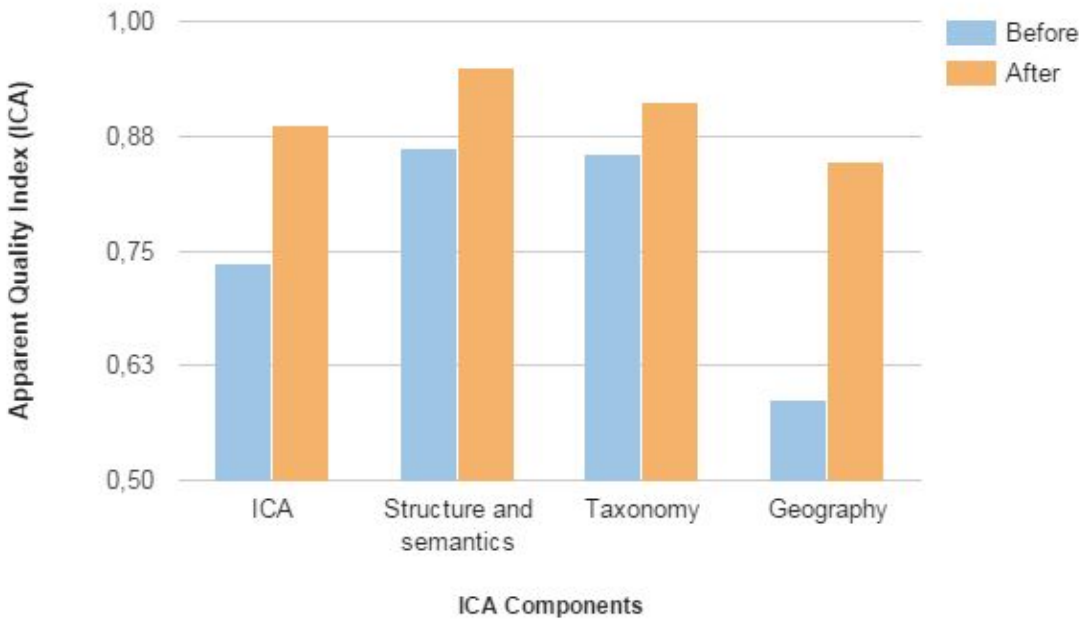

Figure 1. ICA average of datasets evaluated before and after the workshop. Results are also displayed for each of the components of the index.

Comparative ICA values (before and after), show the effect of the workshop on the ability of publishers to improve the quality of their datasets. The three components of data quality shown in

Figure 1 also exhibit the strengths and weakness of publishers structuring primary biodiversity data. While following DwC parameters and controlled vocabularies can be a straightforward task, the management of geographic data can be difficult. Before the workshop the average value of the geographic component was 0.59 and increased almost 30% after the workshop. Although geographic data had the best results it remains as the weaker component into SiB's community publishers; most of our publisher are Biologist with strengths in taxonomy nevertheless with needs of deeper training in geographic data capture.

The ICA value, as stated in its name, is apparent, thus it can be subjective especially when comparing datasets with different purposes, taxa, DwC elements and number of records. A way of overcoming this constraint is to contrast different versions of the same dataset. In order to do so, we took the oldest and latest version of a dataset of one participant institution the *Instituto Humboldt*, which provides the biggest number of datasets in our community and it has been part of SiB's network since the beginning. Measuring the data quality changes of the same dataset gave us a reliably and objective measure of the data quality results achieved with the workshop.

As can be seen in **Figure 2**, the analysis of the data quality follow the same pattern that the one shown in **Figure 1**: a general data quality improvement, particularly in the geographic data. The dataset displayed below. has almost a perfect 1 in taxonomy and, structure and semantic components, this is not only the case for this dataset but for most of them, see Appendix 2. The lowest "after" scores for taxonomy correspond to a dataset of insect for which several names didn't match with Catalog of Life, nevertheless most of them are probably correct but some are endemic and other rarely know.

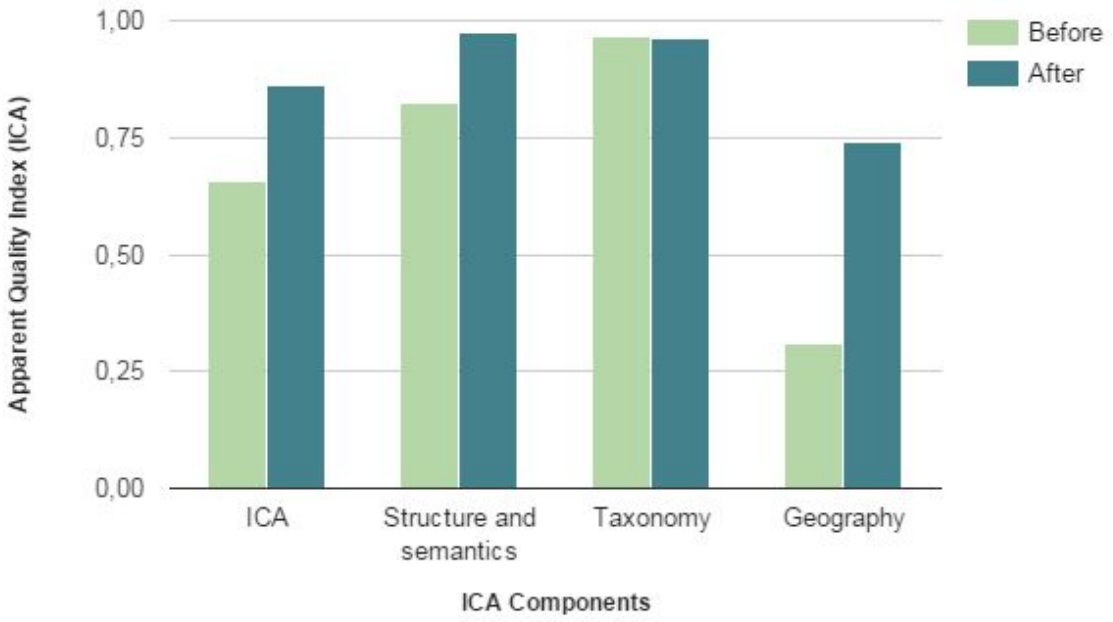

Figure 2. Data quality comparison of the same dataset of the Research Institute of Biological Resources Alexander von Humboldt. Results are also displayed for the three components of the ICA formula.

Final notes

The workshop was a success as it improved Colombian data publishers skills and capabilities on data quality processes. We hope that the other countries improve as well their network and be able to show the results in the short term.

SiB Colombia gained new partner institutions, increasing so far the biodiversity primary data with 48 new resources with 55902 records, furthermore this newly published data has better data quality with an average ICA of 0.89. SiB Colombia's cumulative efforts to improve quality, accompaniments and workshops alike, has shown excellent results. We have reached a point where publishers have good practices using DwC standard and make good use of taxonomy validation tools.

Despite metadata wasn't mentioned in the present document it has the same importance in terms of the resource quality. 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.

Finally, we should consider increasing our efforts towards a deeper training in how to tackle geographic data from the field to DwC. Additionally all the material created for the workshop will become one of the most important tools for guiding new publishers in the path of data quality.

Appendix 1. List of the resources analysed

Institution	Resource URL	Remarks
Universidad del Valle	http://ipt.sibcolombia.net/valle/resource.do?r=insectos-universidad-del-valle http://ipt.sibcolombia.net/valle/resource.do?r=dictyoptera-musenuv	The main group of the resources published after the workshop were Insects.
UIS	http://ipt.sibcolombia.net/sib/resource.do?r=uis-002	UIS started publishing after the workshop, we selected the herbarium resource to avoid problematic taxa.
UNIVERSIDAD ICESI	http://ipt.sibcolombia.net/valle/resource.do?r=herbario-bs-icesi	ICESI started publishing after the workshop, we selected the herbarium resource to avoid problematic taxa.
INSTITUTO HUMBOLDT COLOMBIA	http://ipt.sibcolombia.net/iavh/resource.do?r=aves_iavh&v=15 http://ipt.sibcolombia.net/iavh/resource.do?r=aves_iavh	For the same dataset we estimate the quality of a previous and a later version.
FUNDACIÓN ESTACIÓN BIOLÓGICA GUAYACANAL	http://ipt.sibcolombia.net/sib/resource.do?r=2783_coello_20150403 http://ipt.sibcolombia.net/sib/resource.do?r=guayacanal-parcelasmartos	In both datasets the main taxon is Plantae in order to have a more objective analysis.
abc Creando con-ciencia ambiental	http://ipt.sibcolombia.net/sib/resource.do?r=abc-2009-aves http://ipt.sibcolombia.net/sib/resource.do?r=aves_pautoj	Both datasets main taxon is Aves in order to have a more objective analysis.
anteagroup	http://ipt.sibcolombia.net/cr-sib/resource.do?r=0741_campovelasquez_20141002 http://ipt.sibcolombia.net/cr-sib/resource.do?r=0741_jazmin_20150311	Both datasets main taxon is Aves and Plantae in order to have a more objective analysis.

Appendix 2. Individual ICA values by dataset.

	ICA		Taxonomy		Geography		Structure and semantics	
Universities	Before	After	Before	After	Before	After	Before	After
Univalle	0,59	0,81	0,65	0,69	0,48	0,88	0,80	0,98
UIS		0,90		0,86		0,92		1,00
ICESI		0,96		0,99		0,94		0,94
Research Institutes								
IAVH	0,66	0,86	0,97	0,96	0,31	0,74	0,82	0,97
NGO'S								
Guayacanal	0,85	0,94	0,80	0,94	0,87	0,94	0,97	0,94
ABC	0,89	0,82	0,95	1,00	0,83	0,61	0,90	0,98
Others								
ANTEA	0,69	0,91	0,91	0,94	0,44	0,89	0,81	0,84
Average	**0,74**	**0,89**	**0,86**	**0,91**	**0,59**	**0,85**	**0,86**	**0,95**