BIOINFORMATICS - APPLICATIONS NOTE - Vol. 28 no. 16 2012, pages 2207–2208.

Publisher's version: doi:10.1093/bioinformatics/bts359

BIDDSAT: visualizing the content of biodiversity data publishers in the Global Biodiversity Information Facility network

Javier Otegui[1] and Arturo H. Ariño

Department of Zoology and Ecology, University of Navarra, 31080 Pamplona, Spain

ABSTRACT

Summary: In any data quality workflow, data publishers must become aware of issues in their data so these can be corrected. User feedback mechanisms provide one avenue, while global assessments of datasets provide another. To date, there is no publicly available tool to allow both biodiversity data institutions sharing their data through the Global Biodiversity Information Facility network and its potential users to assess datasets as a whole. Contributing to bridge this gap both for publishers and users, we introduce BIoDiversity DataSets Assessment Tool, an online tool that enables selected diagnostic visualizations on the content of data publishers and/or their individual collections.

Availability and implementation: The online application is accessible at http://www.unav.es/unzyec/mzna/biddsat/ and is supported by all major browsers. The source code is licensed under the GNU GPLv3 license (http://www.gnu.org/licenses/gpl-3.0.txt) and is available at https://github.com/jotegui/BIDDSAT.

Contact: jotellechea@alumni.unav.es

Received on April 16, 2012; revised on June 11, 2012; accepted on June 18, 2012

Advance Access publication June 23, 2012

1 INTRODUCTION

Established in 2001 as an outcome of the OECD 'Mega Science Forum Working Group' with the aim of 'making the world's primary data on biodiversity freely and universally available via the Internet', the Global Biodiversity Information Facility (GBIF, http://www .gbif.org) is currently considered to be the largest initiative in providing access to collections of biodiversity records (Lane, 2003; Saarenmaa, 2005; Scholes *et al*., 2012; Telenius, 2011). It is defined as a network of biodiversity data institutions (data publishers) that own and manage digital collections of primary biodiversity data (PBD) (Johnson, 2007) and make them publicly available on the Internet. Coordinated at the International Secretariat in Copenhagen, Denmark, GBIF acts as a worldwide proxy for PBD owned and shared by participant institutions. In order to optimize information query and retrieval and to serve as a common gateway to the data, GBIF builds a searchable index (http://www.gbif.org/informatics/ infrastructure/indexing/).The institutions that join the GBIF network remain the owners of their own records' intellectual property rights and host the authoritative version of the records. Therefore, GBIF does not modify the original data, it being the data publishers' responsibility for any data quality issue that might appear in their own datasets. However, GBIF flags data for which it encounters issues on a record-by-record basis.

To enable effective interoperability of the heterogeneous data sources that form its network, GBIF has adopted the DarwinCore standard (Wieczorek et al., 2012) for sharing and indexing the biodiversity data made available by the publishers. This standard, available at http://rs.tdwg.org/dwc/, defines elements describing, i.e. the most basic pieces of the PBD—locality and taxonomic identification of the observed organism (Johnson, 2007)—as well as other data such as time of collection, collector and curator names and collection codes, and is used to map each publisher's data structure to a common data model.

In order to explore the reliability and usability of the biodiversity information, recent trends in Biodiversity Informatics are shifting from raw data accrual to developing quality and fitness-for-use assessments on the available data (see for example, Boakes et al., 2010; Yesson et al., 2007), seeking clues on issues that could impact those features. One of the aims of such assessments is to generate sets of recommendations for data holders about where to invest efforts or how to improve their content. To be effective, these recommendations must reach the data managers at the data publisher institution, and this brings the need for both a straightforward assessment of the published data and an effective feedback mechanism. Although being of prime importance, there are still no fully suitable or reliable mechanisms by which data publishers could easily access to the data they share from the users' perspective (Hill et al., 2010; Jetz et al., 2011).

This article presents the development of an online web application that contributes to bridging this gap. BIoDiversity DataSets Assessment Tool (BIDDSAT) is an online PBD visualization environment in which some basic techniques are applied to the content of data publishers and/or collections of the GBIF network. This allows exploring the content of such datasets and find tell-tale patterns or biases that

[1] To whom correspondence should be addressed.

may reduce its quality or usability. With these visualizations, the source of error can also be detected, allowing data publishers to fix the issues at the source.

2 OVERVIEW

2.1 Data sources

Through an agreement with GBIF, we were granted access to images of the full index of the data derived from its global network of data publishers (more information available at www.gbif.org/ informatics/infrastructure/retrieving/). These images are snapshots of the database at the time of release and are structured as a group of MySQL(http://www.mysql.com/) tables. We focused on three tables of many available: 'occurrence_record', which holds the primary records, and 'taxon_concept' and 'taxon_name', which are used to de-codify taxonomic concepts and names that had been codified to save space.

2.2 Data extraction

A series of routines have been developed as batch SQL extractions, to get summary files for certain aspects of the content of the index for each data publisher and each collection: spatial, temporal and taxonomic distribution of records, as well as some other metadata. In order to check the evolution of the information (the global correction and data acquisition patterns), we have been mining several (though not all) index images since May 2008 and have extracted these summaries for each version of the database.

2.3 Data visualization interface

The visualizations are built on the summary files. A series of PHP (http://www.php.net/) webpages implement the scripts that perform the data visualizations: some of them make use of the Google Charts API (http://code.google.com/apis/chart/) to enable interactive graphic representations of the summary files' content, while others transform the content of such files to prepare two special graphic types: maps and chronhorograms (Ariño and Otegui, 2008), built upon the ImageMagick IMagick PHP extension (http:// www.imagemagick.org/). A form gives access to the individual visualizations after selecting the desired version of the GBIF index, the data publisher to be assessed and, optionally, the collection within that publisher. Since the application is designed to check individual collections or individual publishers as a whole, it is not possible to select more than one collection or publisher at a time. The fields for publisher and collection selection have an autocomplete feature written with AJAX.

3 APPLICATION

This application visualizes and assesses several aspects of the content of the data publishers, which are based on recent advances in data quality and fitness-for-use analyses as well as on previous work developed by the authors. The application is focused on assessing the status of the three aspects of the PBD: geospatial, temporal and taxonomic information, but some metadata issues are also addressed, such as the volume of the collection/publisher or the distribution of types of record. The visualizations make it possible to detect patterns—either 'good'(natural) or 'wrong'(artifactual)— and issues, and in most cases the discovery of 'wrong' trends leads to unveiling methodological issues in the datasets.

When processed by GBIF's internal mechanisms, a data publisher's records pass through a standardization process and some information is codified for performance reasons. If the original data are not structured according to the standard schema, some inconsistencies might appear in the process, leading to interoperability issues. The content of a publisher is assessed as it is once the records have gone through these processes, thus detecting the final status of the records as available to end users. Although there are other tools to assess the quality of a set of data, the collection-centric perspective of this tool allows for the effective detection of patterns that would not arise when records are checked one by one.

4 CONCLUSION

Our online application tries to contribute to bridge biodiversity data quality researchers and data managers in the GBIF network, by allowing visual exploration of a data publisher's collection of records. Being a web-based, openly available application, everyone can access the assessments and, thus, anyone can contribute spotting issues to the improvement of the available biodiversity data quality. Our future roadmap points to the development of new quality assessments and the establishment of an effective communication channel with the data publishers themselves.

5 ACKNOWLEDGEMENTS

We are grateful to the GBIF Secretariat Staff at Copenhagen for granting us the access to the dumps of the full GBIF index. Special thanks to Dr Samy Gaiji and Tim Robertson for comments and support.

Funding: This work has been supported by the 'Friends of the University of Navarra' Association.

Conflict of Interest: none declared.

REFERENCES

Ariño, A.H. and Otegui, J. (2008) Sampling biodiversity sampling. In Weitzman, A.L. and Belbin, L. (eds), *Proceedings of TDWG (2008)*. Biodiversity Information Standards (TDWG), Fremantle, AU, p. 107.

Boakes, E.H. *et al.* (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.*, **8**, e10000385.

Hill, A.W. *et al.* (2010) *GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network*. Global Biodiversity Information Facility, Copenhagen, p. 29.

Jetz, W. *et al.* (2011) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.*, **27**, 151–159.

Johnson, N.F. (2007) Biodiversity informatics. *Annu. Rev. Entomol.*, **52**, 421–438.

Lane, M.A. (2003) The global biodiversity information facility. *Bull. Am. Soc. Inform. Sci. Technol.*, **30**, 22–24.

Saarenmaa, H. (2005) Sharing and accessing biodiversity data globally through GBIF. In *ESRI User Conference*, San Diego.

Scholes, R.J. *et al.* (2012) Building a global observing system for biodiversity. *Curr. Opin. Environ. Sustain.*, **4**, 1–8.

Telenius, A. (2011) Biodiversity information goes public: GBIF at your service. *Nordic J. Botany*, **29**, 378–381.

Wieczorek *et al.* (2012) Darwin core: an evolving community-developed biodiversity data standard. *PLoS One*, **7**, e29715.

Yesson, C. *et al*. (2007) How global is the global biodiversity information facility? *PLoS One*, **2**, e1124.