

Addressing biodiversity data quality is a community-wide effort

June 3 2013

cies Locations Collections	Mapping & analysis	Data sets B	logs Get in	volved About th	e Allas
e : Clashboard 1. You can rearrange topics by citizing and dray	yyhy hen			Show raw data	Download as C3V
Occurrence records	Data sets	515	Basis of re	ecords	
39,061,118 records in total.	Harvested websites 168 Ocourrence record sets 244 Document sets 30 Uploaded record sets 73 Most recently added dataset is Australian Tropical Rainforest Plants information system		Human observation 28.61M Preserved specimen 8.37M Machine observation 321,514 Genomic DNA 154,529		
We estimate the number of potential duplicate records to be 5,098,423.			Living specimen 59,890 Fossil specimen 22,114 Image 19,839		
Collections 147	Records by life	eform		Records by	date
Vertex New	Algae 42	1,189 FernsAndAllie: 1,658 Fish	329,879 1,221,399 368,851	Latest record Last image adde	14 May 2013 d 2 October 2012
Other Same Insects	and the second	1,285 Gymnosperms		1600s 1700s 1800s	25 4,305 345,405
	Birds 21,707	1,731 Mammals 1,362 Mollusos	1,379,839	1900s	18.63M 16.53M
	Chromista 47	7,139 Monecets 7,120 Plants 1,836 Protezoa	2,212,219 9,610,561 151,685	-	and the state
	 Institution and the setunder of t	.839 Reptiles	736.476	National Spe	

This image shows a small part of the screen of the dashboard from the ALA. It provides a little indication of what the Atlas has. Credit: Atlas of Living Australia, ALA

Improving data quality in large online data access facilities depends on a



combination of automated checks and capturing expert knowledge, according to a paper published in the open-access journal *Zookeys*. The authors, from the <u>Atlas of Living Australia</u> (ALA) and the <u>Global</u> <u>Biodiversity Information Facility</u> (GBIF) welcome a recent paper by <u>Mesibov (2013)</u> highlighting errors in millipede data, but argue that addressing such issues requires the joint efforts of 'aggregators' and the wider expert community.

The paper notes that aggregations of data openly exposed in facilities such as the ALA and GBIF will contain errors, and both organisations are fully committed to improving the quality of these data. Errors will arise in a multitude of ways. For example, an observation of a species may be misnamed, the name could have changed or the pre-GPS location could be in error. The card entry of this observation could then have been incorrectly transcribed into a digital record by a museum or herbarium. When the record was translated into a standard form for communication with the ALA or GBIF, other errors could have been introduced. At each step of the process, errors can be detected, introduced or corrected.

The authors argue that one of the most powerful outcomes of publishing digital data is that such problems are revealed, providing an opportunity for the whole community to detect and correct them. The paper points out that Mesibov's detection of data issues was only possible with convenient public exposure of a large volume of <u>biological data</u> through the ALA and GBIF.

The ALA and GBIF also run a comprehensive range of automated data checks, for example flagging records whose coordinates lie outside the stated country of the observation or specimen. Such automatic checks will not detect all errors. Specialist expertise therefore remains necessary to detect and correct a wide range of data issues.



Agencies such as the GBIF and the ALA have infrastructure that simplifies error detection and correction. Aggregating many records of a species improves the chances of errors being detected. For example, one observation may be geographically isolated from other records. In the ALA, anyone can <u>annotate</u> an issue exposed in a record. Such annotations are sent to the data provider for evaluation and correction. It then depends on the resources of the provider to ensure that record is updated.

The ability to identify and correct data issues is the responsibility of the whole community and not any one agent such as the ALA. There is the need to seamlessly and effectively integrate expert knowledge and automated processes, so all amendments form part of a persistent digital knowledge base about species. Talented and committed individuals can make enormous progress in error detection and correction (as seen in Mesibov's paper) but how do we ensure that when an individual project like that on millipedes ceases, the data and all associated work are not lost? This implies standards in capturing and linking this information and maintaining the data with all amendments uniquely documented. To achieve this, the biodiversity research community needs to be motivated and empowered to work in a collaborative fashion.

Data should be published in secure locations where they can be preserved and improved in perpetuity. The ALA and GBIF are moving beyond storage of data by individuals or institutions using stand-alone computers that do not have a strategy for enduring digital data integration, storage and access.

More information: Belbin L, Daly J, Hirsch T, Hobern D, Salle JL (2013) A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. Title. *ZooKeys* 305: 67–76, <u>doi:</u> 10.3897/zookeys.305.5438



Provided by Pensoft Publishers

Citation: Addressing biodiversity data quality is a community-wide effort (2013, June 3) retrieved 23 April 2024 from https://phys.org/news/2013-06-biodiversity-quality-community-wide-effort.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.