



Principles and tools on data quality and fitness for use of biodiversity occurrence data

Rui Figueira
Nó Português do GBIF
rui.figueira@iict.pt



UNIVERSIDADE
DE LISBOA



nó português do GBIF

Apoio

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

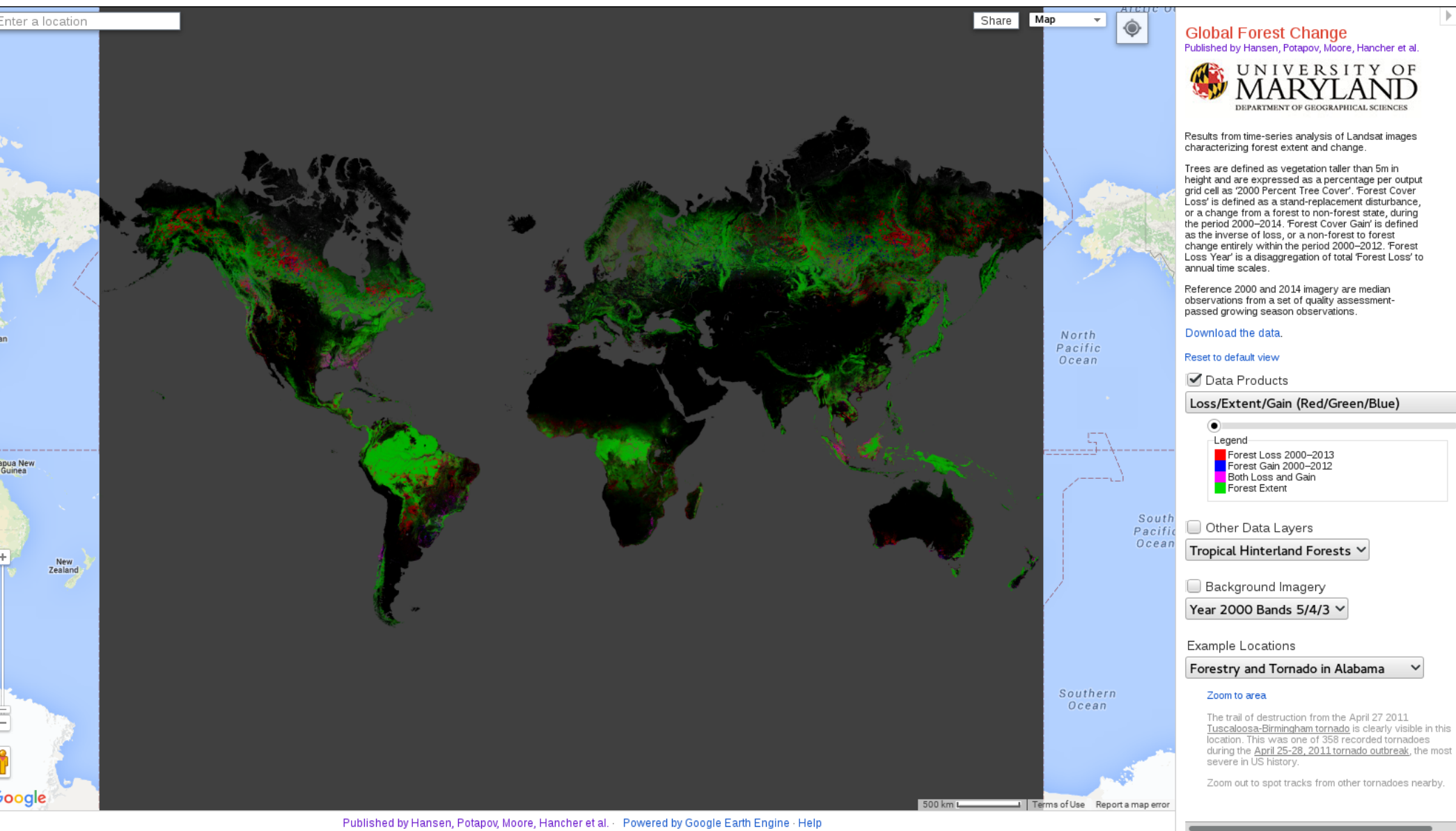
Outline

- Introduction
- Data quality – main concepts
- Using data – “fitness for use”
- Tools to promote data quality

The Human Brain Project



Big data – Global Forest Change



GBIF BY THE NUMBERS

570,238,233

species occurrence records

15,073

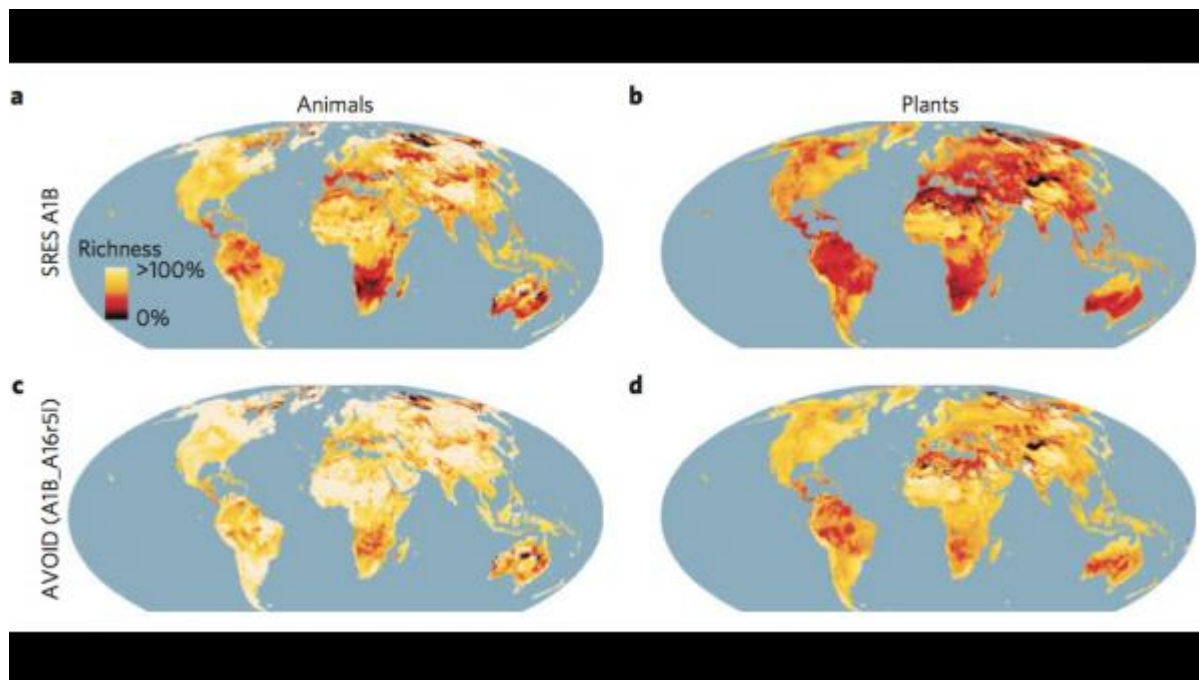
datasets

762

data-publishing institutions



GBIF enables global study of climate impact on species



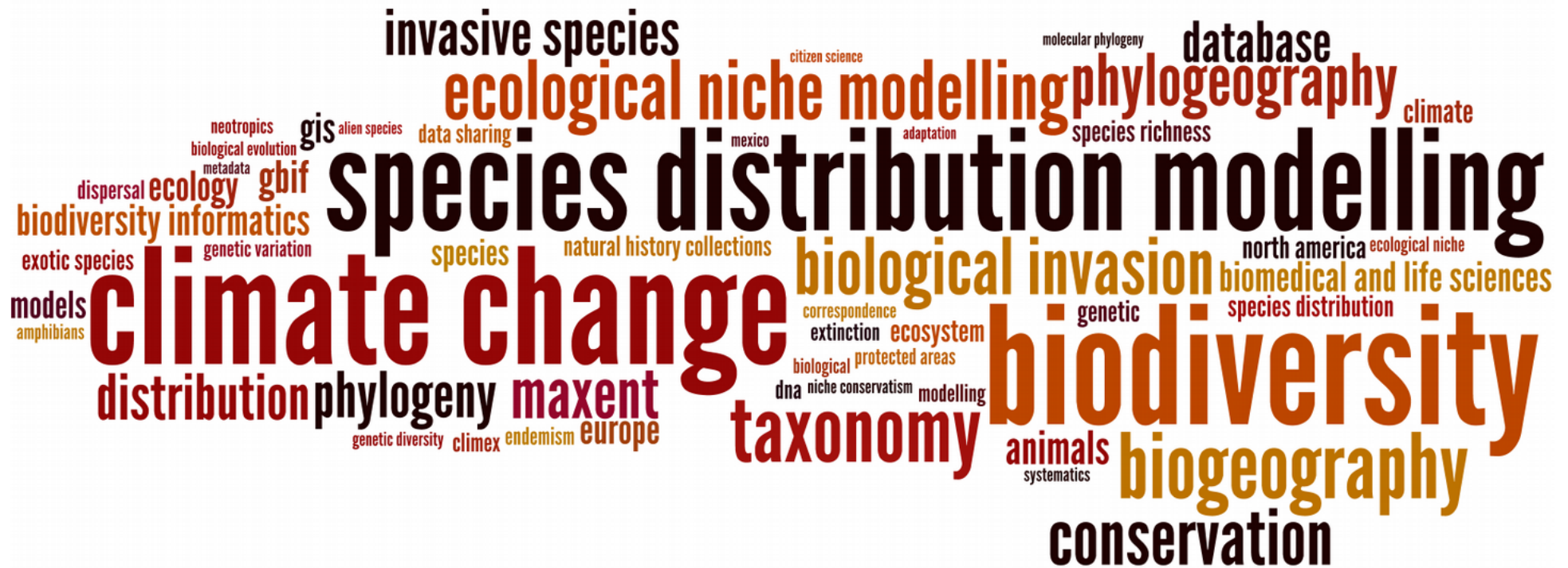
50 000 species
170 million records

More than half of the plants and over a third of animal species could lose more than half of their climatic range by 2080.

Citation Information

Warren, R. et al., 2013. Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss. *Nature Clim. Change*, advance online publication. Available at: [dx.doi.org/10.1038/nclimate1887](https://doi.org/10.1038/nclimate1887).

Data usage



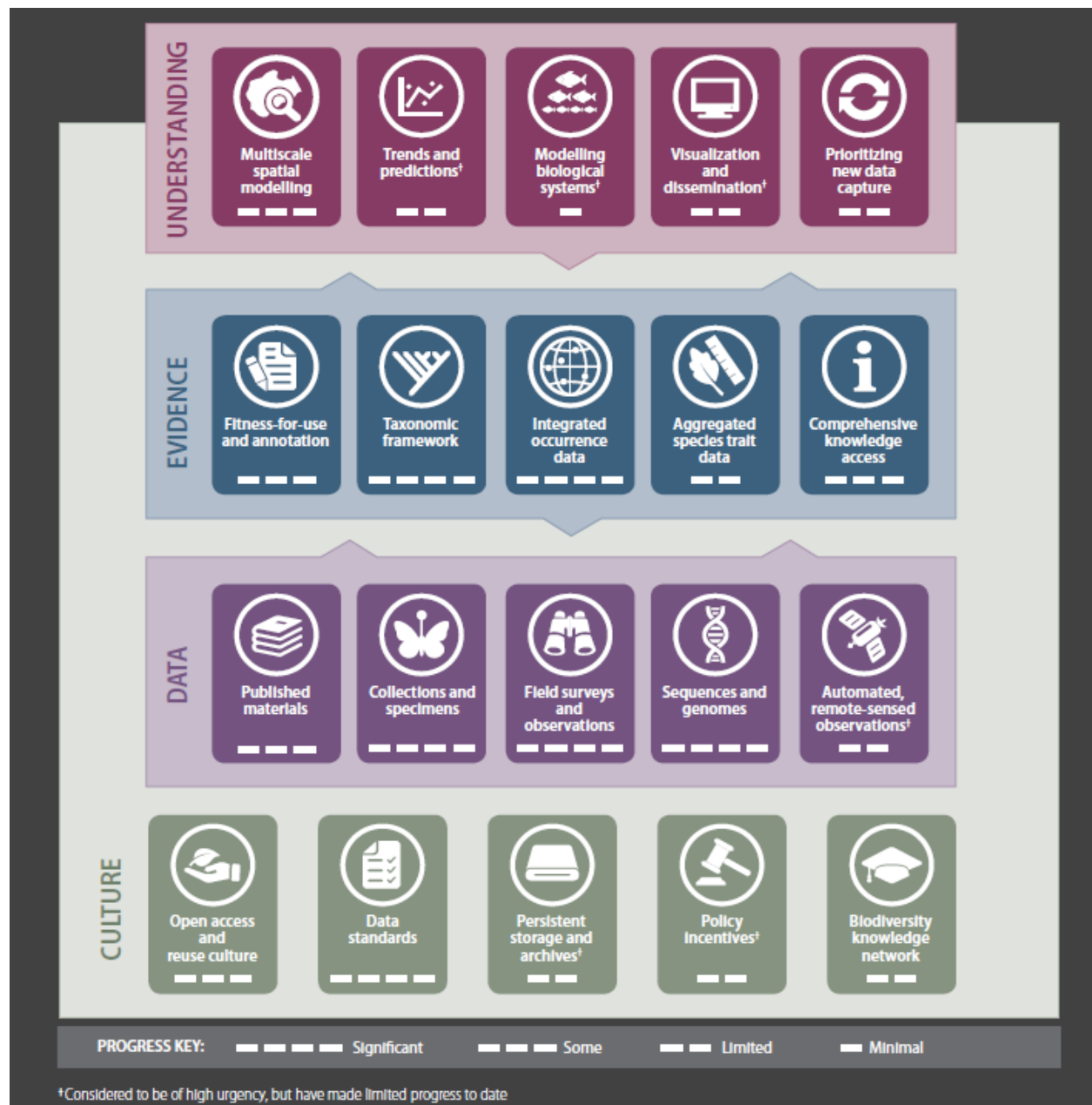
<http://www.mendeley.com/groups/1068301/gbif-public-library/>

keywords in ~1900 research
papers that use or quote GBIF,
from GBIF Public Library



Global Biodiversity
Informatics Outlook
www.biodiversityinformatics.org

The GBIO framework



Definition

*An essential and needed feature of data is its “**fitness for use**” .*

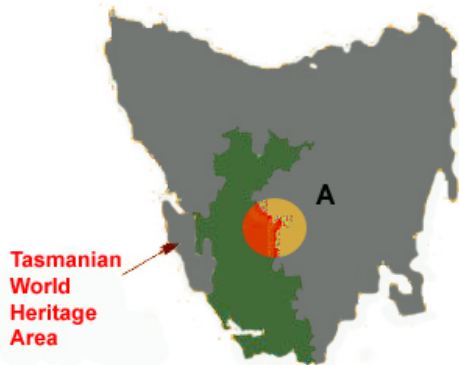
*"The general intent of describing the quality of a particular dataset or record is to describe the **fitness** of that dataset or record for **a particular use** that one may have in mind for the data."*

Chrisman, 1991

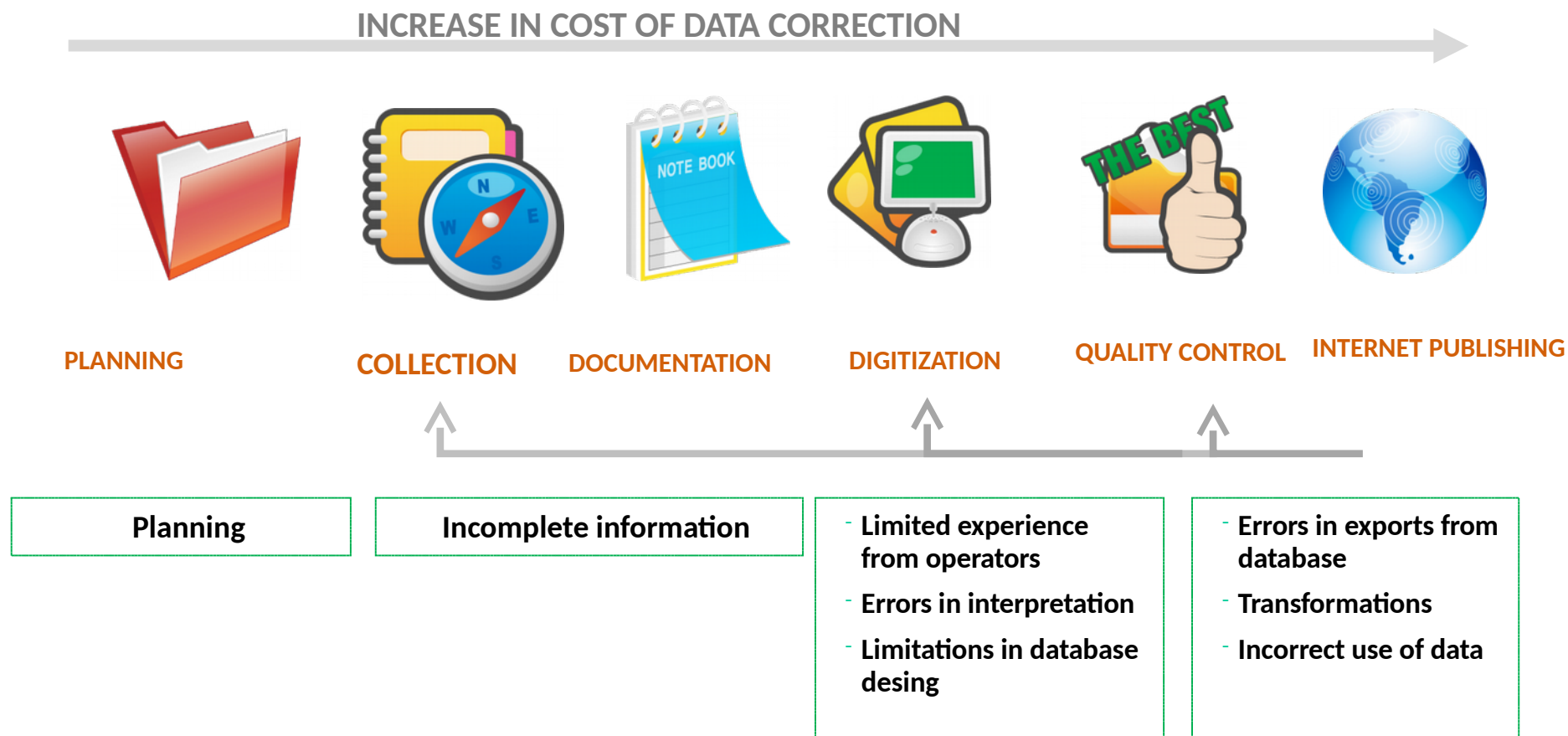


Is species “A” present in Tasmania?

Is species “A” present in the Tasmanian World Heritage Area?



The information chain and costs in loss of data quality



Dimensions of information

What

Taxonomic/nomenclatural data

Where

Spatial data

Who

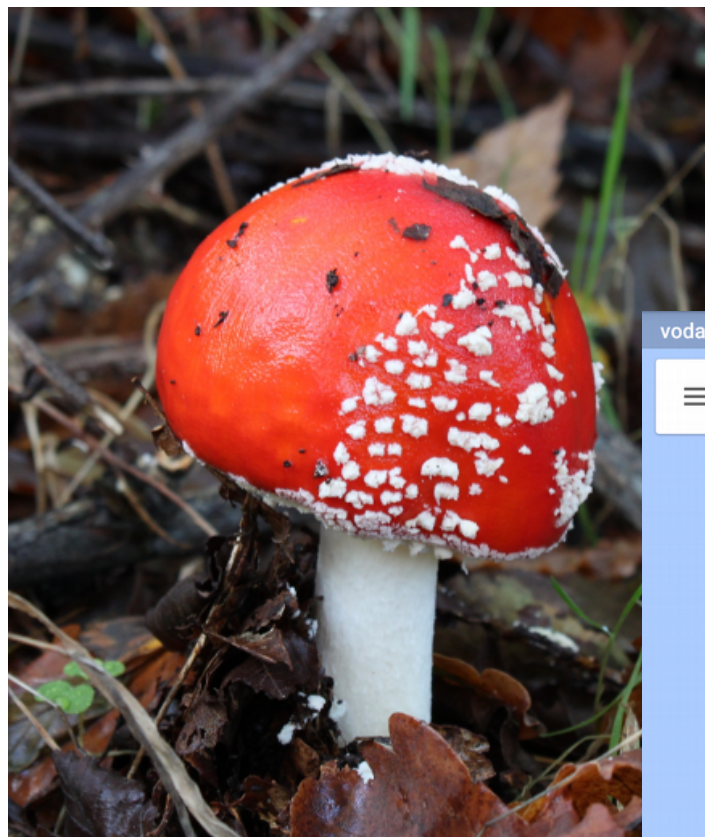
Occurrence data

When

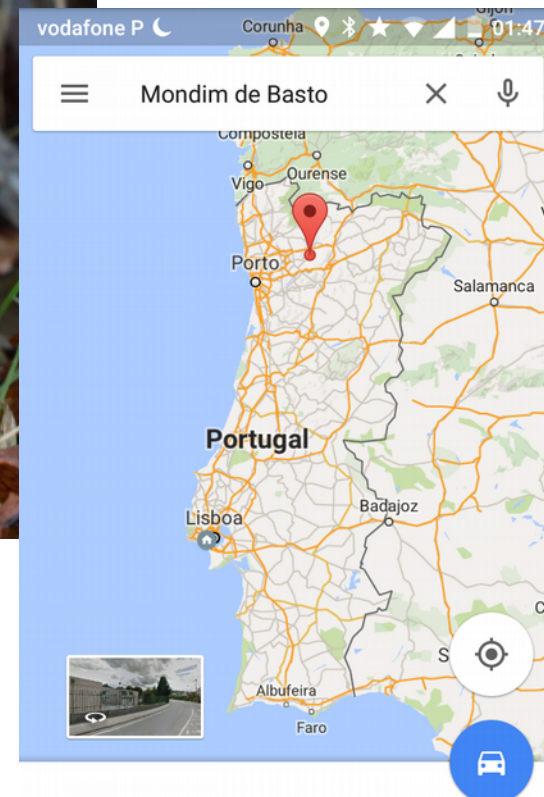
Sampling data

What

Descriptive data



@CésarGarcia, 2010/11/23
Amanita muscaria



Mondim de Basto

Dimensions of information

What

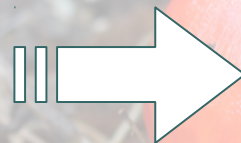
Taxonomic/nomenclatural data



authority files

Where

Spatial data



gazetteers, uncertainty

Who

Occurrence data



lists of collectors

When

Sampling data



database checks

What

Descriptive data



thesauri

@CésarGarcia, 2010/11/23



Mondim de Basto



Resources

Library of documents, tools and other information to support the GBIF community

Resources

Literature

Key information

For publishers

For users

For GBIF delegations

For nodes



Featured resources



GBIF monthly slides (.pptx)

GBIF publishes a set of slides with key information on a monthly basis. The slides...



Apresentação: GBIF Atualização Mensal (.pptx)

GBIF publica um conjunto de slides com as principais informações numa base mensal. Os slides...



GBits Newsletter no. 44

July 2015



GBIF and IPBES Collaboration

This two-page leaflet summarizes the highlights of the collaboration between GBIF and IPBES up to...



Presentation: Using biodiversity data in conservation biology



Presentation: How to use GBIF.org to obtain biodiversity

Filters

[clear all](#)

RESOURCE TYPE

- ☐ Document (397)
- ☐ Link (170)
- ☐ Presentation (137)
- ☐ Tool (124)

PURPOSE

- ☐ Data publishing (108)
- ☐ Data curation & quality (104)
- ☐ Data analysis (98)
- ☐ Data digitization (52)
- ☐ Data access (24)

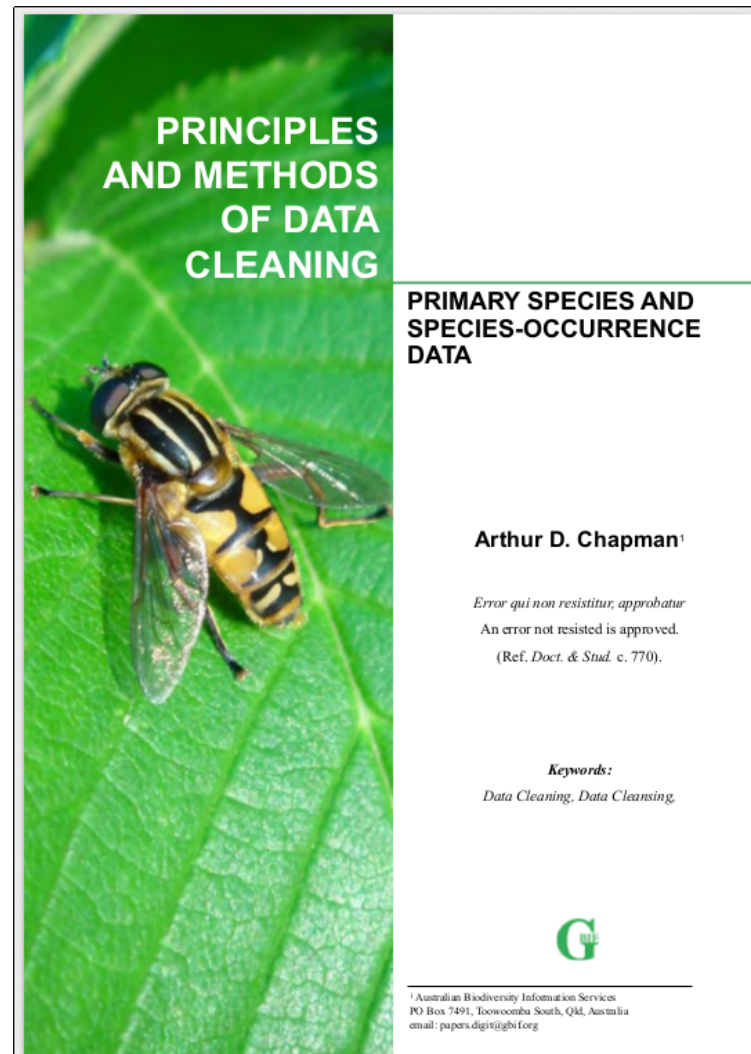
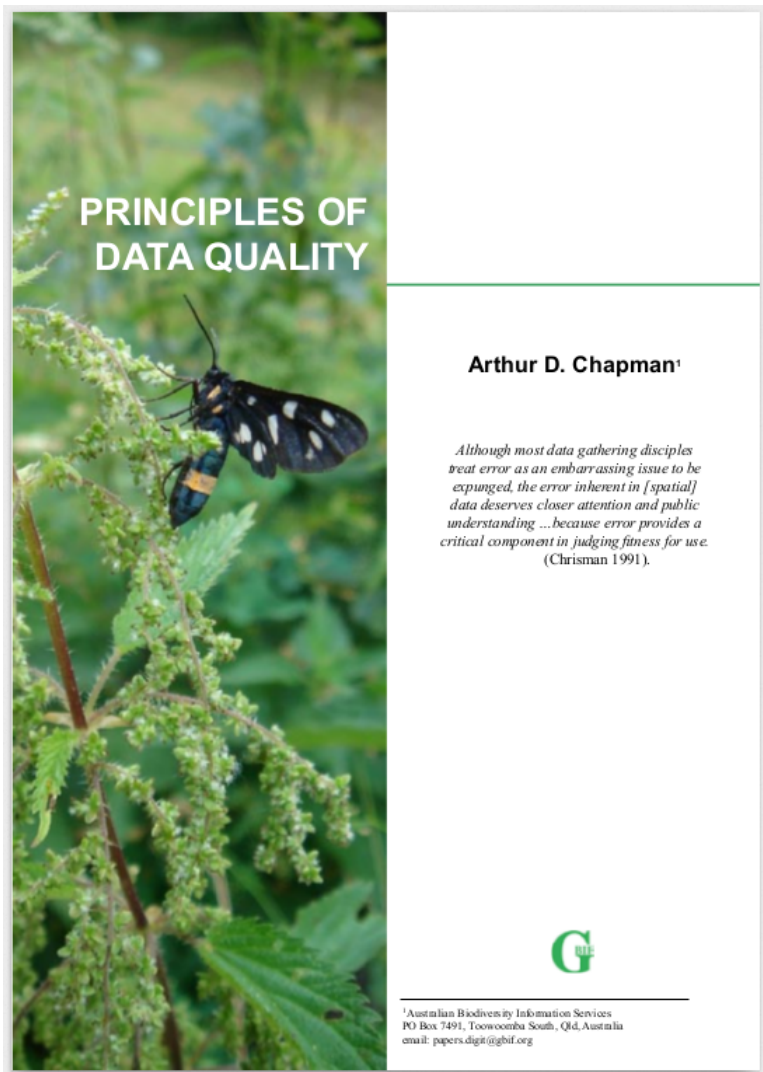
LANGUAGE

- ☐ English (647)
- ☐ Spanish (53)
- ☐ French (49)
- ☐ Chinese, Traditional (25)
- ☐ Chinese, Simplified (24)

Biodiversity data quality hub

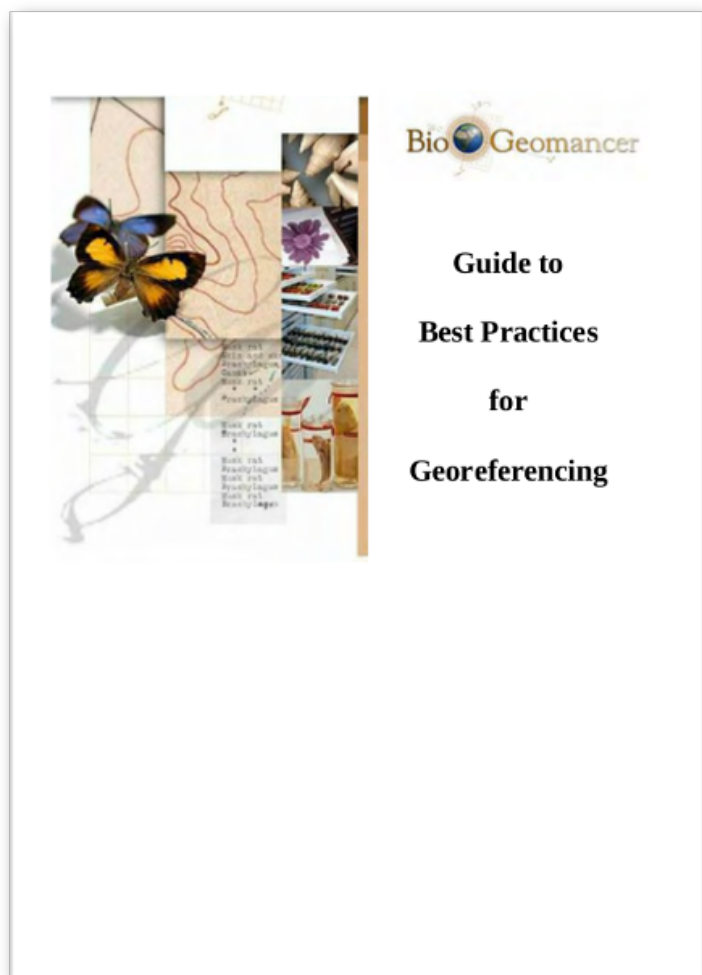
<http://www.gbif.es/BDQ.php>





<http://www.gbif.org/resources/2829>

<http://www.gbif.org/resources/2833>



Biogeomancer, Guide to Best Practices in Georeferencing

A.D. Chapman, J. Wiecek (Eds.)



Spatial, taxonomic, temporal, miscellaneous issues
(missing values, incoherence, no match, ...) - more than 80 checks

Data quality tests

Test name	Result
Basis of record badly formed ?	✖ Failed
Collection code not recognised ?	✖ Failed
Institution code not recognised ?	✖ Failed
Coordinate precision not valid ?	✖ Failed
Data are generalised ?	⚠ Warning
Name not in national checklists ?	⚠ Warning
Basis of record not supplied ?	✔ Passed
Missing catalogue number ?	✔ Passed
Missing taxonomic rank ?	✔ Passed
Name not supplied ?	✔ Passed
Kingdom not recognised ?	✔ Passed
Name not recognised ?	✔ Passed
Invalid scientific name ?	✔ Passed
Decimal coordinates not supplied ?	✔ Passed
Geodetic datum assumed WGS84 ?	✔ Passed
Unrecognized geodetic datum ?	✔ Passed
Coordinates are transposed ?	✔ Passed
Coordinates are out of range for species ?	✔ Passed
Supplied coordinates are zero ?	✔ Passed
Supplied country not recognised ?	✔ Passed



urn:lsid:artportalen.se:Sighting:54509279

Human Observation of *Amanita muscaria* (L.) Hook., 1797 recorded on Jul 20, 2015

from Artdata dataset

Information

Verbatim

INTERPRETATION ISSUES

GBIF found issues interpreting the [verbatim content](#) of this record:

- Geodetic datum assumed WGS84
- Country derived from coordinates
- Taxon match higher rank
- Depth unlikely



Location

LOCALITY

Åsenholmsgrundet, [Sweden](#)

23.4602, 65.8033 ± 1

ELEVATION

-1m ± 0m

DEPTH

1m ± 0m

INTERPRETATION ISSUES

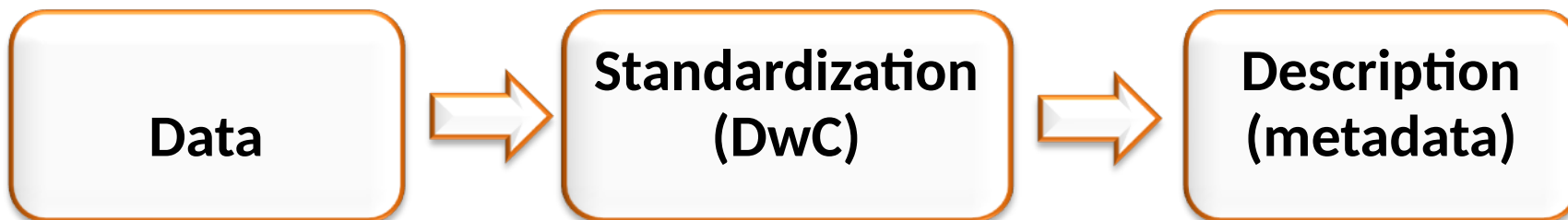
GBIF found issues interpreting the verbatim content of this record:

- Geodetic datum assumed WGS84
- Country derived from coordinates
- Taxon match higher rank
- Depth unlikely

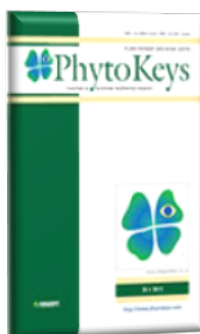
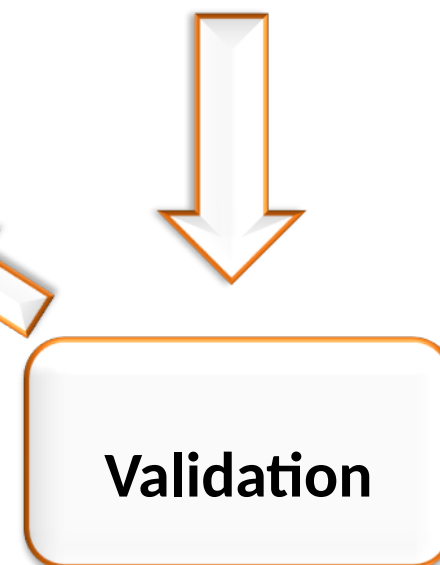


- ✓ Zero coordinate
- ✓ Coordinate out of range
- ✓ Coordinate invalid
- ✓ Coordinate rounded
- ✓ Geodetic datum invalid
- ✓ Geodetic datum assumed WGS84
- ✓ Coordinate reprojected
- ✓ Coordinate reprojection failed
- ✓ Coordinate reprojection suspicious
- ✓ Country coordinate mismatch
- ✓ Country mismatch
- ✓ Country invalid
- ✓ Country derived from coordinates
- ✓ Continent country mismatch
- ✓ Continent invalid
- ✓ Continent derived from coordinates
- ✓ Presumed swapped coordinate
- ✓ Presumed negated longitude
- ✓ Presumed negated latitude
- ✓ Recorded date mismatch
- ✓ Recorded date invalid
- ✓ Recorded date unlikely
- ✓ Taxon match fuzzy
- ✓ Taxon match higher rank
- ✓ Taxon match none
- ✓ Depth not metric
- ✓ Depth unlikely
- ✓ Depth min max swapped
- ✓ Depth non numeric
- ✓ Elevation unlikely
- ✓ Elevation min max swapped
- ✓ Elevation not metric
- ✓ Elevation non numeric
- ✓ Modified date invalid
- ✓ Modified date unlikely

How to publish data through GBIF



Data publishing



Metadata publishing (Data paper)

Darwin Core

Biodiversity
Information
Standards
TDWG

Introduction

References

Quick Reference Guide

Term Index

Record-level Terms

Occurrence

Event

Location

GeologicalContext

Identification

Taxon

ResourceRelationship

MeasurementOrFact

Term Definitions

Simple Darwin Core

Type Vocabulary

Namespace Policy

XML Guide

Text Guide

Complete History

Decision History

Mapping to ABCD

Mapping to Old Versions

Translations

Darwin Core Terms: A quick reference gu

Title: Darwin Core Terms: A quick reference guide

Date Issued: 2009-02-12

Date 2011-10-26

Modified:

Abstract: This document is a quick reference for all recommended Darwin Core terms. For complete historical term information, including version cha and pre-standard terms, see [\[HISTORY\]](#). For a comparative table of elements from pre-standar versions of Darwin Core to the current terms in standard, see [\[VERSIONS\]](#).

Contributors: John Wieczorek (MVZ), Markus Döring (GBIF), R De Giovanni (CRIA), Tim Robertson (GBIF), Dave Vieglaiss (KUNHM)

Legal: This document is governed by the standard legal copyright, licensing provisions and disclaimers is by the Taxonomic Databases Working Group.

Part of TDWG Standard: <http://www.tdwg.org/standards/450/>

Creator: Darwin Core Task Group

Identifier: <http://rs.tdwg.org/dwc/2011-10-26/terms/>

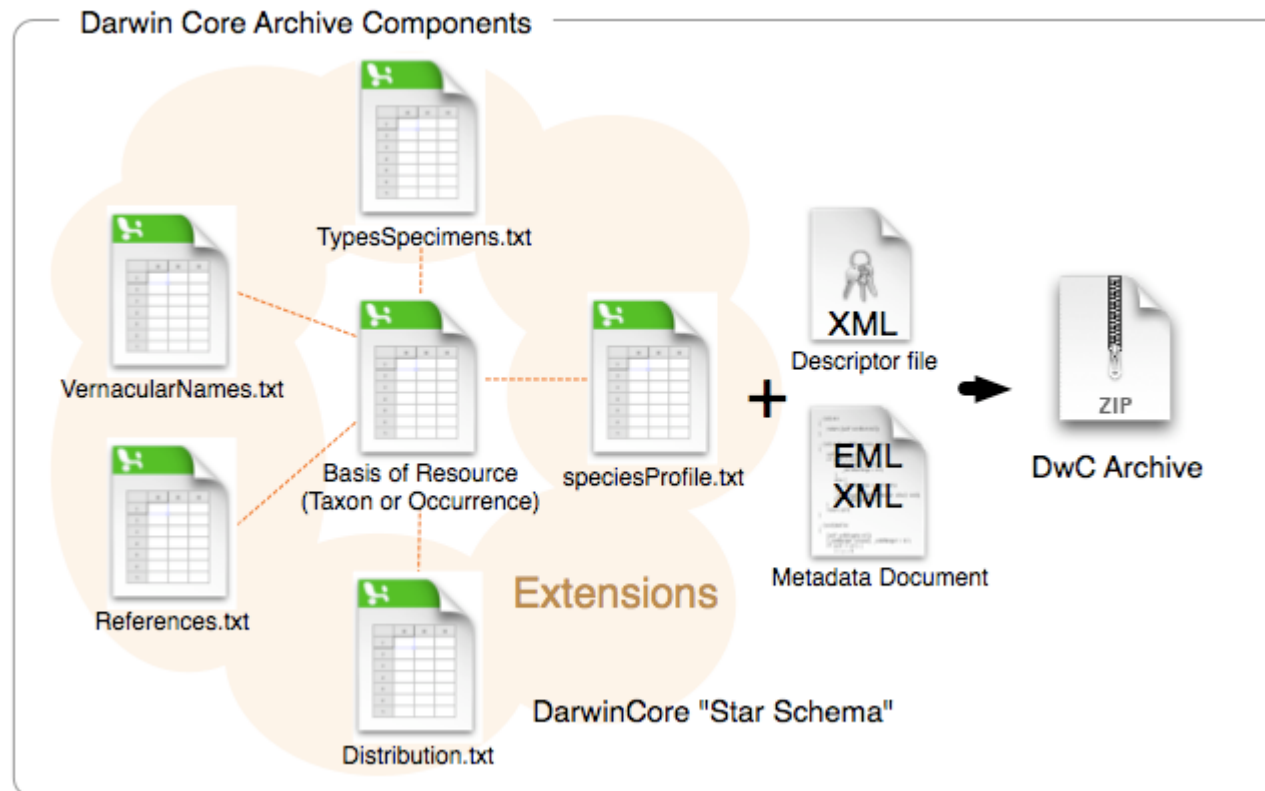
Latest Version: <http://rs.tdwg.org/dwc/terms/>

Replaces: <http://rs.tdwg.org/dwc/2009-12-07/terms/>

Document Status: Current Standard

Darwin Core – Star schema supporting several types of information:

Germoplasm, multiple determinations, types and specimens, common names, alternative identifiers, species profile, references, taxon description, traits, multimedia, among others





ECAT Name Parser

[parser](#) [api](#)

Name Parser

This is a simple html form to make use of the GBIF name parser. The parser is written in java and based on regular expressions to dissect name strings into its components. It does only keep name parts required to reconstruct a full 3-parted name with an optional subgenus, but ignores additional infraspecific parts such as the subspecies given for varieties. Please see our [API documentation](#) for details.

You can copy paste a list of names, one per row, or upload a text file with a name per line. Uploaded files have to be encoded as utf8!

Names to
parse:

*One per line
or delimited
by the pipe
symbol "|"*

Phaeographis smithii (Leight.) de Lesd.
Gyalidea madeirensis Kalb
Thelopsis isiaca Stizenb.
Thelopsis isiaca Stizenb.
Thelopsis rubella Nyl.
Diploschistes actinostomus (Ach.) Zahlbr.
Diploschistes caesioplumbeus (Nyl.) Vain.

Upload File: No file selected.

GBIF

[News](#)
[Datasets](#)
[Species](#)
[Occurrences](#)

Tools

[Name Finder](#)
[Name Parser](#)
[DwC Archive Validator](#)
[DwC Archive Assistant](#)

Developer

[Developer Blog](#)
[API](#)

Contact

[Help Desk](#)
[Directory of contacts](#)

© 2013 The Global Biodiversity Information Facility (GBIF)
All software licended under Apache 2.0

<http://tools.gbif.org/nameparser/>



ECAT Name Parser

[parser](#) [api](#)

Parsed Names

7 name parsed. 7 wellformed, 0 hybrid formulas and 0 doubtful names. See legend for [parsing types](#).

[Show](#) extended parsing

Original	Genus	Infrageneric	Specific	Rank	Notho	InfraSpecific	Authorship	Year	(Authorship)	(Year)
Phaeographis smithii (Leight.) de Lesd.	Phaeographis		smithii	SPECIES			de Lesd.		Leight.	
Gyalidea madeirensis Kalb	Gyalidea		madeirensis	SPECIES			Kalb			
Thelopsis isiaca Stizenb.	Thelopsis		isiaca	SPECIES			Stizenb.			
Thelopsis isiaca Stizenb.	Thelopsis		isiaca	SPECIES			Stizenb.			
Thelopsis rubella Nyl.	Thelopsis		rubella	SPECIES			Nyl.			
Diploschistes actinostomus (Ach.) Zahlbr.	Diploschistes		actinostomus	SPECIES			Zahlbr.		Ach.	
Diploschistes caesioplumbeus (Nyl.) Vain.	Diploschistes		caesioplumbeus	SPECIES			Vain.		Nyl.	

Parser Result Types

sciname

... a scientific name which is not well formed

<http://tools.gbif.org/nameparser/>



Darwin Core Archive Validator

To validate a [Darwin Core Archive](#) file either provide a url to an archive or upload an archive including data files for validation.

Please note that we limit the size of uploaded files to 100MB, so reduce your data files if necessary. We will happily pull bigger archives from a url provided.

Validate archive URL:

Upload local archive:

No file selected.

GBIF

[News](#)
[Datasets](#)
[Species](#)
[Occurrences](#)

Tools

[Name Finder](#)
[Name Parser](#)
[DwC Archive Validator](#)
[DwC Archive Assistant](#)

Developer

[Developer Blog](#)
[API](#)

Contact

[Help Desk](#)
[Directory of contacts](#)



Darwin Core Archive Validator

To validate a [Darwin Core Archive](#) file either provide a url to an archive or upload an archive including data files for validation.

Please note that we limit the size of uploaded files to 100MB, so reduce your data files if necessary. We will happily pull bigger archives from a url provided.

Validate archive URL:

- ✓ structure of the files
- ✓ extensions
- ✓ metadata
- ✓ completeness
- ✓ controlled vocabularies

Upload local archive:

 No file selected.

GBIF

[News](#)
[Datasets](#)
[Species](#)
[Occurrences](#)

Tools

[Name Finder](#)
[Name Parser](#)
[DwC Archive Validator](#)
[DwC Archive Assistant](#)

Developer

[Developer Blog](#)
[API](#)

Contact

[Help Desk](#)
[Directory of contacts](#)

© 2013 [The Global Biodiversity Information Facility \(GBIF\)](#)

All software licended under Apache 2.0

<http://tools.gbif.org/dwca-validator>

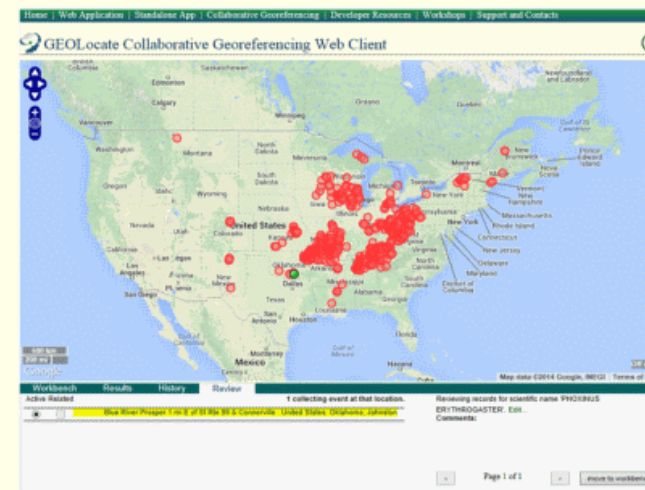
GEOLocate

A Platform for Georeferencing Natural History Collections Data

For Users:

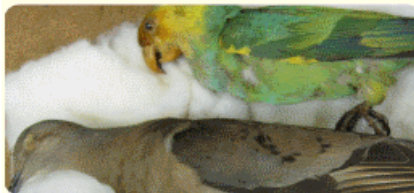
- Overview
- GEOLocate Web Application
- Collaborative Georeferencing
- GEOLocate 3.xx (standalone)
 - Global Expansion
- Education & Outreach

Brief overview (video) of the GEOLocate Project.



For Developers:

- SOAP Services
- JSON/GeoJSON
- Embeddable Web Client



Web Application

Georeference collections data using your web browser. Quick and easy georeferencing.



Web Services

Integrate georeferencing into your own databases and applications using GEOLocate webservice.



Desktop Application

The original standalone desktop application.



Collaborative Georeferencing

Build communities, share data, relate records across collections and improve verification efficiency.

Copyright © 2015



<http://www.museum.tulane.edu/geolocate/>



Open Refine (ex-Google Refine) is a desktop application (although it runs inside the browser), powerfull to clean messy data, transform in other formats, extend with online resources.

The introductory videos are the best way to see full potential and get familiar to Open Refine (also available at <http://openrefine.org/>):

Google Refine 2.0 – Introduction (http://www.youtube.com/watch?v=B70J_H_zAWM)

Google Refine 2.0 - Data Transformation (http://www.youtube.com/watch?v=cO8NVCs_Ba0)

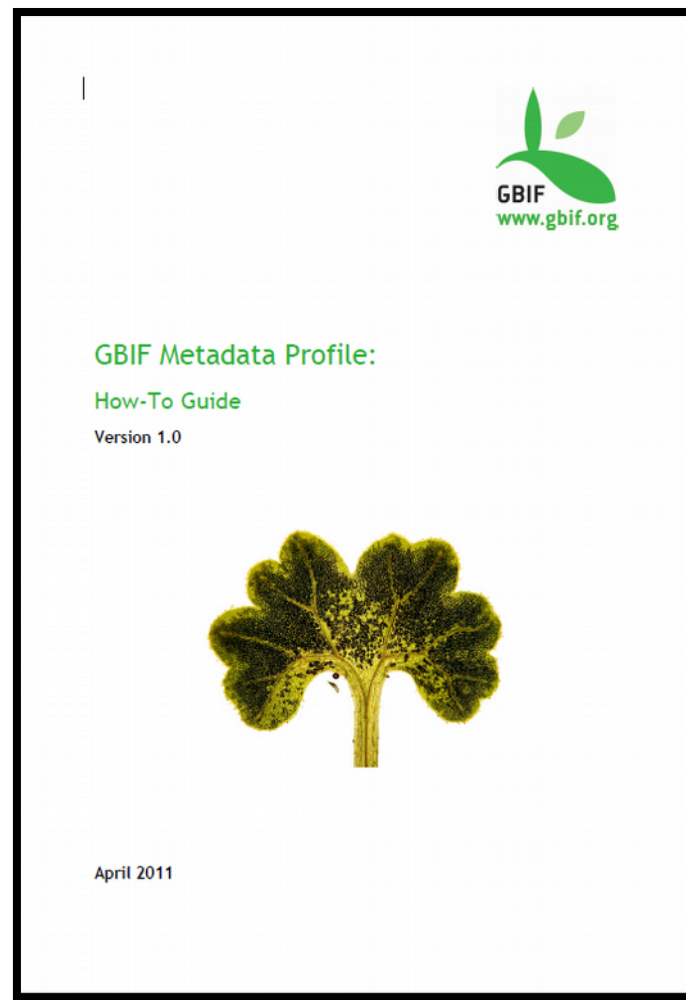
Google Refine 2.0 - Data Augmentation (<http://www.youtube.com/watch?v=5tsyz3ibYzk>)

Metadata


GMP (GBIF Metadata profile):

EML (Ecological Metadata Language)

- Dataset (Resource)
 - Project
 - People and Organisations
 - Keyword Set (General Keywords)
 - Coverage
 - Taxonomic Coverage
 - Geographic Coverage
 - Temporal Coverage
 - Methods
 - Intellectual Property Rights
 - Additional Metadata + NCD (Natural Collections Descriptions Data)
- Related



Integrated Publishing Toolkit


GBIF PORTUGAL - INTEGRATED PUBLISHING TOOLKIT
 acesso aberto e gratuito a dados de biodiversidade (IPT)

[ENGLISH](#)


[Home](#)
[About](#)

Hosted resources available through this IPT

Filter:

Logo	Name	Organisation	Type	Subtype	Records	Last modified	Last publication	Next publication
--	Bryophyte collection of Porto Herbarium (PO)	Museu de História Natural da Universidade do Porto	Occurrence	--	7,621	2014-07-25	2014-07-25	--
--	Checklist da Flora de Portugal (Continental, Açores e Madeira)	GBIF Portugal	Checklist	--	3,994	2014-11-18	2014-11-18	--
--	Moluscos Marinhos de Augusto Nobre	Museu de História Natural da Universidade do Porto	Occurrence	Specimen	881	2013-12-17	2013-12-17	--

Showing 1 to 3 of 3 resources ◀ previous next ▶

The most recently updated resources are also available as an [RSS feed](#). 

IPT Version 2.1.1-r4640
 [About the IPT project](#)
[User manual](#)
[Report a bug](#)
[Request new feature](#)

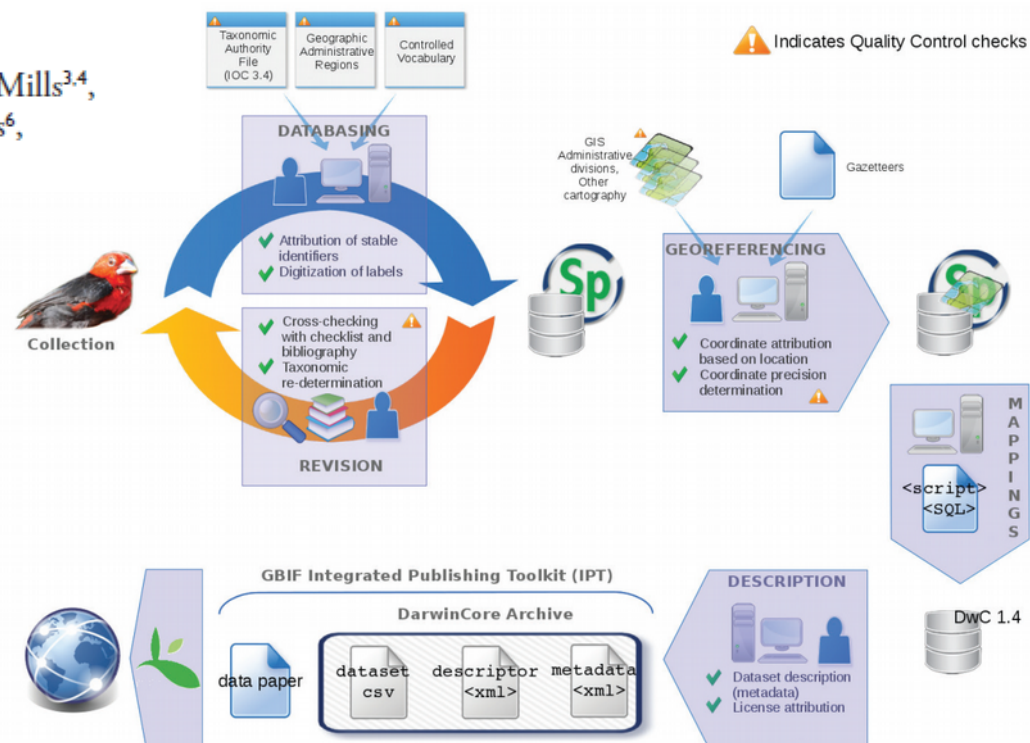
ZooKeys 387: 89–99 (2014)
doi: 10.3897/zookeys.387.6412
www.zookeys.org

DATA PAPER

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

The collection and database of Birds of Angola hosted at IICT (Instituto de Investigação Científica Tropical), Lisboa, Portugal

Miguel Monteiro^{1,2}, Luís Reino², Pedro Beja², Michael Stuart Lyne Mills^{3,4},
Cristiane Bastos-Silveira^{5,6}, Manuela Ramos¹, Diana Rodrigues⁶,
Isabel Queirós Neves^{5,6}, Susana Consciência¹, Rui Figueira^{1,2}



Machine readable information – it is important that it provides:

- License type
- Data quality issues
- Data accuracy
- Missing data



Final remarks

- All steps in data cycle are critical to ensure data quality – best practices;
- The GBIF network provides the facility, but mobilizing is still the most important task to overcome gaps;
- Tools can assist to identify issues and improve data quality;
- Adhering to a global framework should help to overcome barriers, while helping to identify local needs;

Quality of data enhances its multiple use in the future

Credits / References

Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-03-8. Available online at http://www.gbif.org/orc/?doc_id=1229.

Chapman, A. D. 2005. Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available online at http://www.gbif.org/orc/?doc_id=1262.

Chapman, A.D. and J. Wiecek (eds). 2006. Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility. Available online at http://www.gbif.org/orc/?doc_id=1288

Thank you!

Rui Figueira
Instituto de Investigação Científica Tropical
Nó Português do GBIF
Rua da Junqueira, 86-1º
1300-344 Lisboa, Portugal
rui.figueira@iict.pt
www.gbif.pt



<http://creativecommons.org/licenses/by-sa/3.0/deed.pt>