



Curso de formação:

Gestão de coleções biológicas utilizando Specify 6

Qualidade e limpeza de dados

Rui Figueira

ruifigueira@isa.ulisboa.pt

Nó Português do GBIF,
Instituto Superior de Agronomia,
Universidade de Lisboa

SUMÁRIO

BLOCO 2 – QUALIDADE E LIMPEZA DE DADOS

2.1. Introdução ao padrão de dados Darwin Core

2.2. Conceitos de qualidade de dados e controlo de qualidade

2.2.1. Dados taxonómicos

2.2.2. Dados geográficos

2.2.3. Pessoas e entidades, datas

2.2.4. Dados sensíveis

2.3. Ferramentas e recursos para a limpeza de dados

Qualidade de dados - referências

PRINCÍPIOS DE QUALIDADE DE DADOS

Arthur D. Chapman¹

Although most data gathering disciplines treat error as an embarrassing issue to be expunged, the error inherent in [spatial] data deserves closer attention and public understanding ... because error provides a critical component in judging fitness for use. (Chrisman 1991).

GBIF
www.gbif.org

¹ Australian Biodiversity Information Services
PO Box 7491, Toowoomba South, QLD, Australia
email: info@gbif.org

<https://www.gbif.org/document/80924>

PRINCIPLES AND METHODS OF DATA CLEANING

PRIMARY SPECIES AND SPECIES-OCCURRENCE DATA

Arthur D. Chapman¹

Error qui non resistitur, approbatur
An error not resisted is approved.
(Ref. Doct. & Stud. c. 770).

Keywords:
Data Cleaning, Data Cleansing.

GBIF

¹ Australian Biodiversity Information Services
PO Box 7491, Toowoomba South, QLD, Australia
email: papers.digit@gbif.org

<https://www.gbif.org/document/80528>

PLOS ONE

RESEARCH ARTICLE

A conceptual framework for quality assessment and management of biodiversity data

Allan Koch Veiga^{1*}, Antonio Mauro Saraiva^{2*}, Arthur David Chapman³, Paul John Morris⁴, Christian Gendreau⁵, Dmitry Schigel⁶, Tim James Robertson⁷

¹ University of São Paulo, Research Center on Biodiversity and Computing, São Paulo, São Paulo, Brazil, ² Australian Biodiversity Information Services, Ballan, Victoria, Australia, ³ Harvard University, Museum of Comparative Zoology, Cambridge, Massachusetts, United States of America, ⁴ Université du Montréal, Institut de Recherche en Biologie Végétale, Montréal, Québec, Canada, ⁵ Global Biodiversity Information Facility, Secretariat, Copenhagen, Denmark

* allan.kv@gmail.com (AKV), saraiva@usp.br (AMS)

Check for updates

Abstract

The increasing availability of digitized biodiversity data worldwide, provided by an increasing number of institutions and researchers, and the growing use of those data for a variety of purposes have raised concerns related to the "fitness for use" of such data and the impact of data quality (DQ) on the outcomes of analyses, reports, and decisions. A consistent approach to assess and manage data quality is currently critical for biodiversity data users. However, achieving this goal has been particularly challenging because of idiosyncrasies inherent in the concept of quality. DQ assessment and management cannot be performed if we have not clearly established the quality needs from a data user's standpoint. This paper defines a formal conceptual framework to support the biodiversity informatics community allowing for the description of the meaning of "fitness for use" from a data user's perspective in a common and standardized manner. This proposed framework defines nine concepts organized into three classes: DQ Needs, DQ Solutions and DQ Report. The framework is intended to formalize human thinking into well-defined components to make it possible to share and reuse concepts of DQ needs, solutions and reports in a common way among user communities. With this framework, we establish a common ground for the collaborative development of solutions for DQ assessment and management based on data fitness for use principles. To validate the framework, we present a proof of concept based on a case study at the Museum of Comparative Zoology of Harvard University. In future work, we will use the framework to engage the biodiversity informatics community to formalize and share DQ profiles related to DQ needs across the community.

1. Introduction

Data Quality (DQ) is a subject that permeates most research. As a result, research on DQ, information quality, or data fitness for use has been conducted and applied in a number of domains, covering multiple aspects and approaches [1–6].

OPEN ACCESS

Citation: Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, et al. (2017) A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE 12(6): e0178731. <https://doi.org/10.1371/journal.pone.0178731>

Editor: Ulrich Meixner, Oklahoma State University, UNITED STATES

Received: July 4, 2016

Accepted: May 18, 2017

Published: June 28, 2017

Copyright: © 2017 Veiga et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data used and generated are available from the API <http://case.gbif.org> or <http://case.gbif.org/br3070/veiga/>. All the information needed for interested researchers to query the available data, using a standard RESTful API interface, is available in the SI Appendix, with the proof of concept.

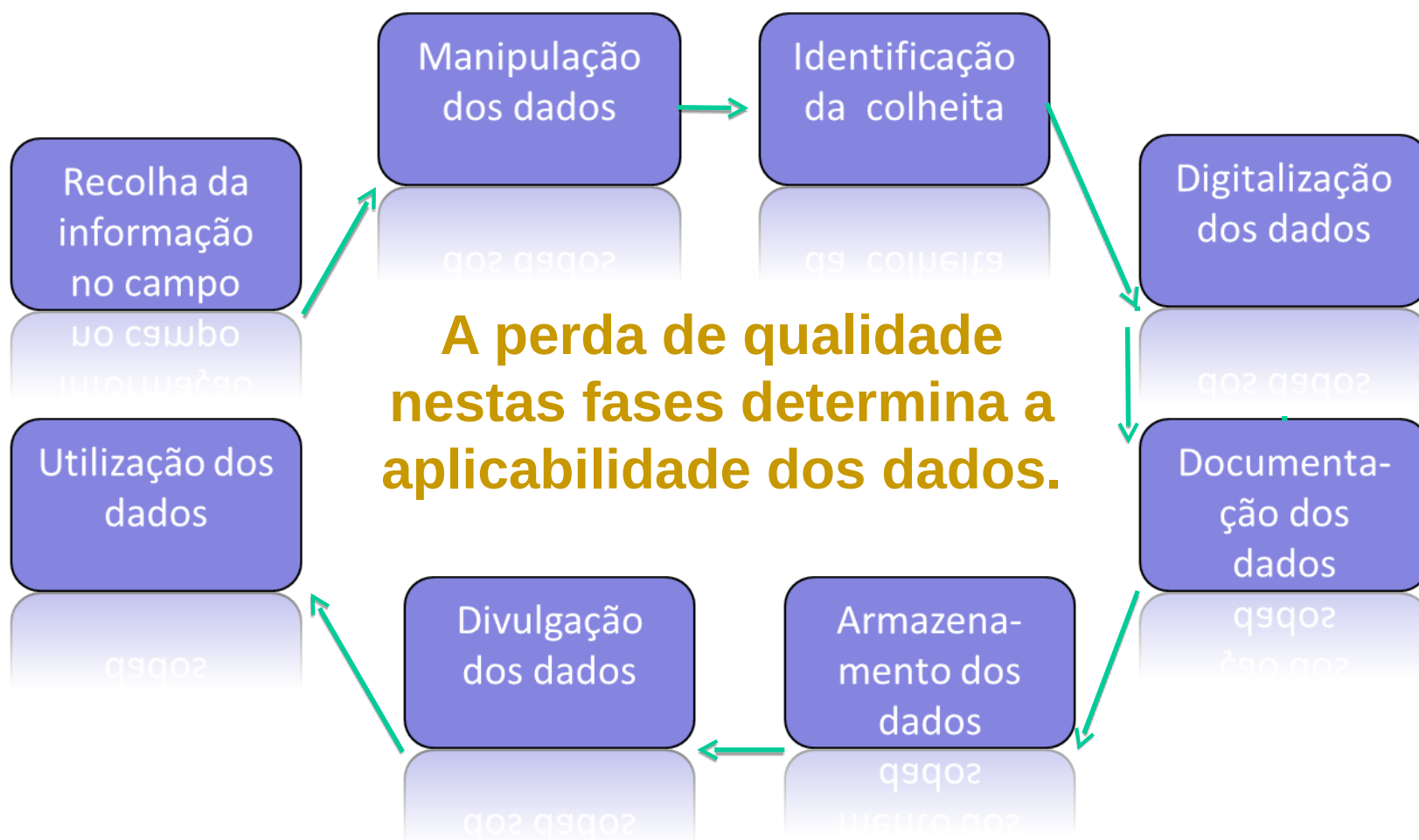
Funding: The study is part of the Doctorate Degree thesis of AKV, who was supported by a Brazilian Governmental National Research Agency - Conselho de Aperfeiçoamento de Pessoal de

PLOS ONE | <https://doi.org/10.1371/journal.pone.0178731> June 28, 2017

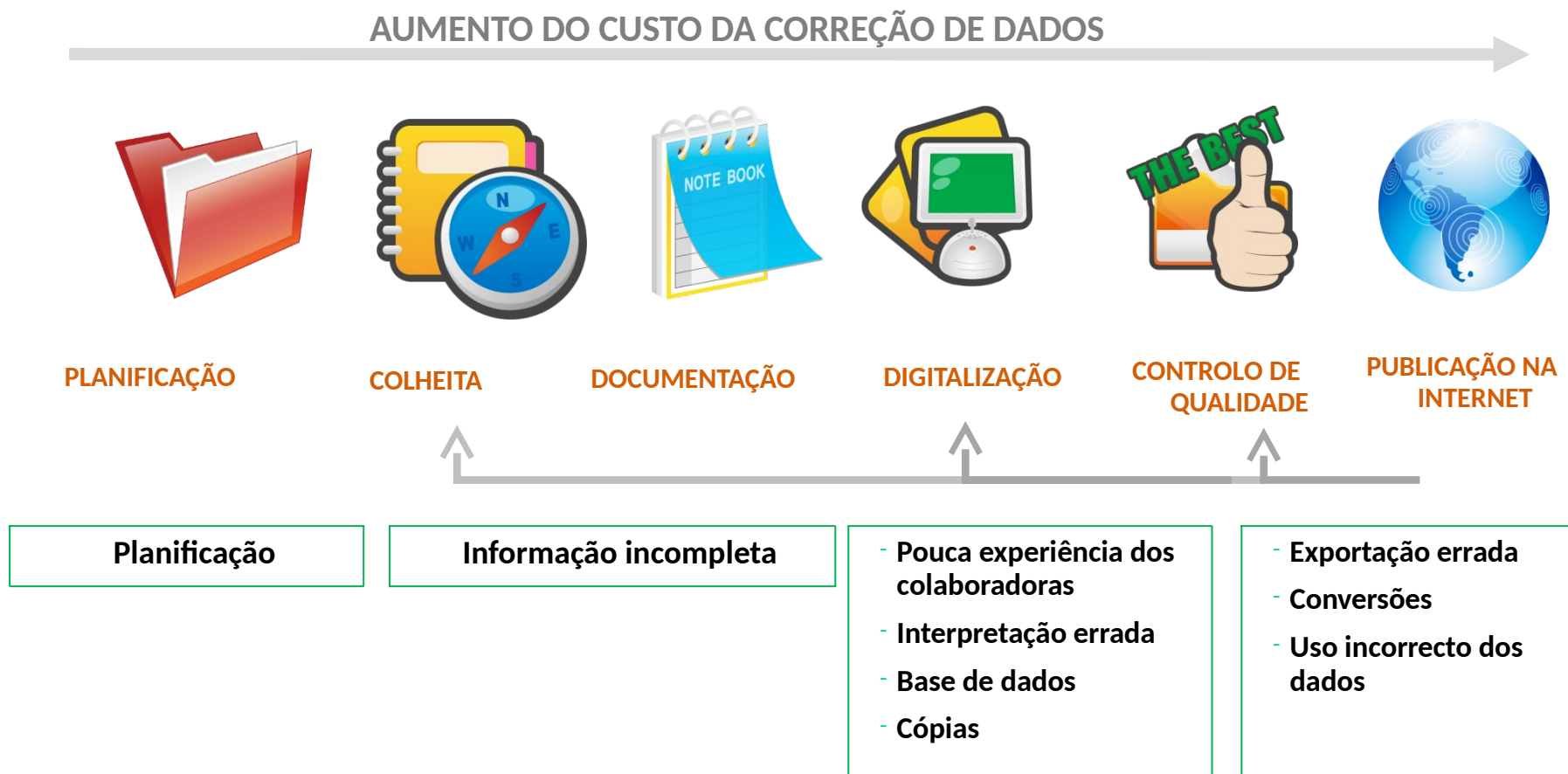
1/20

<https://doi.org/10.1371/journal.pone.0178731>

Ciclo de vida dos dados



A cadeia de informação e a perda de qualidade



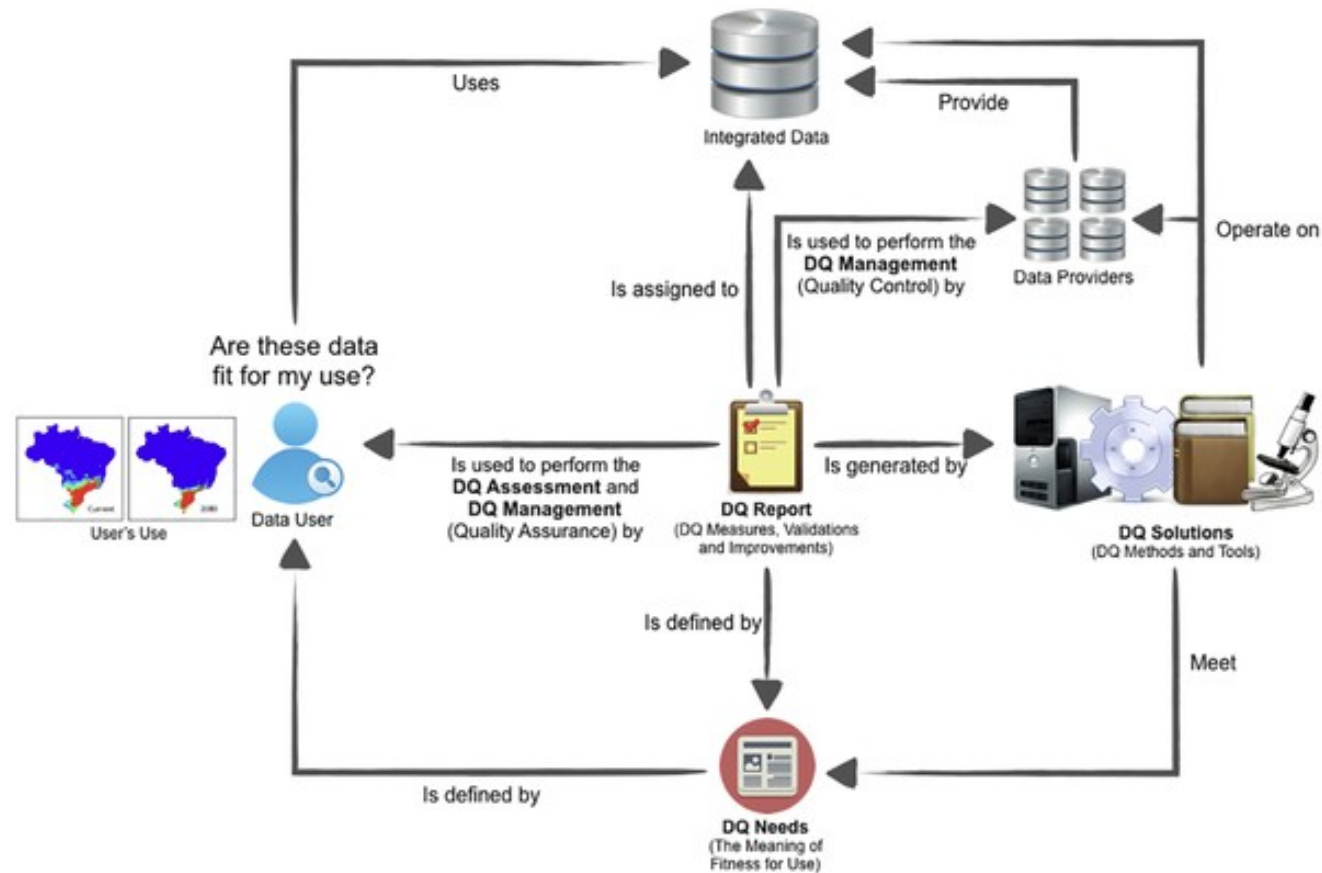
Total Data Quality Management (TDQM)



<https://doi.org/10.1371/journal.pone.0178731.g001>

DQ conceptual framework

<https://doi.org/10.1371/journal.pone.0178731.g002>

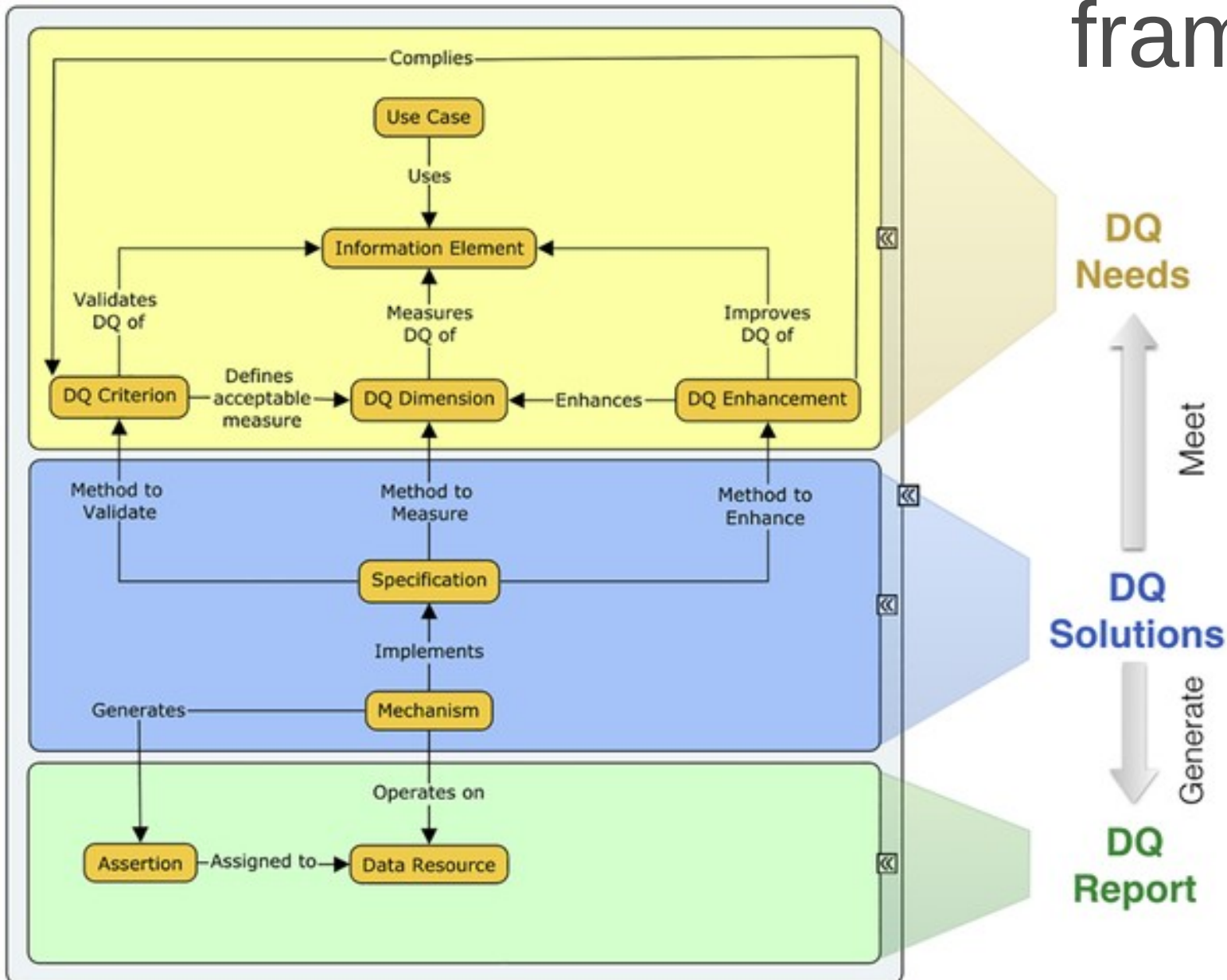


DQ Needs – ajuste para o uso

DQ Report – avaliação, validação e melhorias

DQ Solutions – métodos e ferramentas

DQ conceptual framework



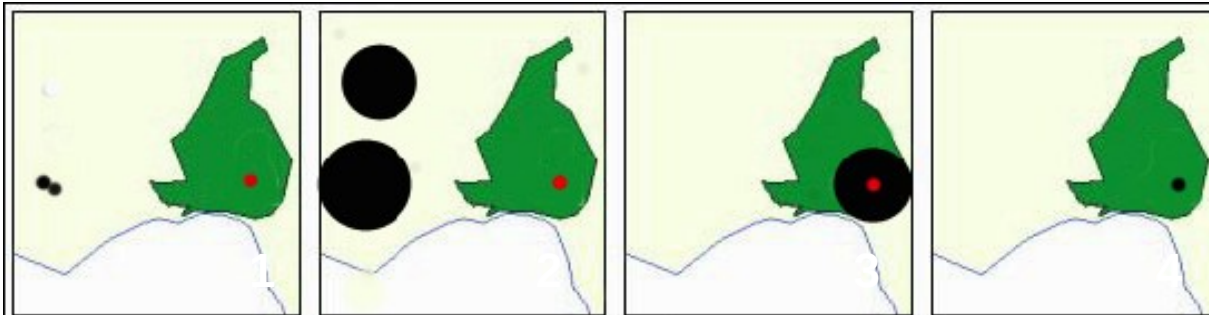
Conceitos de qualidade de dados

Conceitos

Exactidão (accuracy) – proximidade dos valores medidos, observados ou estimados, aos valores reais ou verdadeiros

Precisão ou resolução (precision) –

- *precisão estatística* – proximidade (ou variabilidade) entre várias observações
- *precisão numérica* – resolução, número de dígitos significativos para uma determinada observação



- 1-alta precisão, baixa exactidão
- 2-baixa precisão, baixa exactidão
- 3-baixa precisão, alta exactidão
- 4-alta precisão, alta exactidão

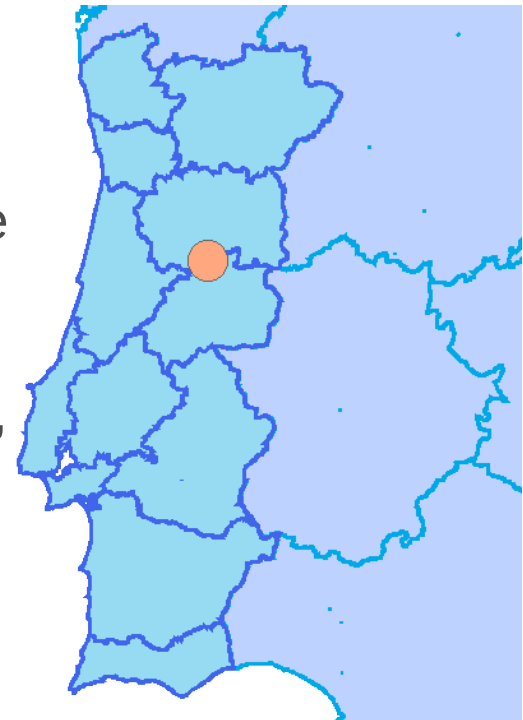
Conceitos de qualidade de dados

Conceitos

Qualidade

A qualidade depende da utilização que será dada aos dados

A qualidade dos dados é multidimensional, envolvendo a gestão dos dados, análise e modelação, controlo de qualidade, armazenamento e apresentação.



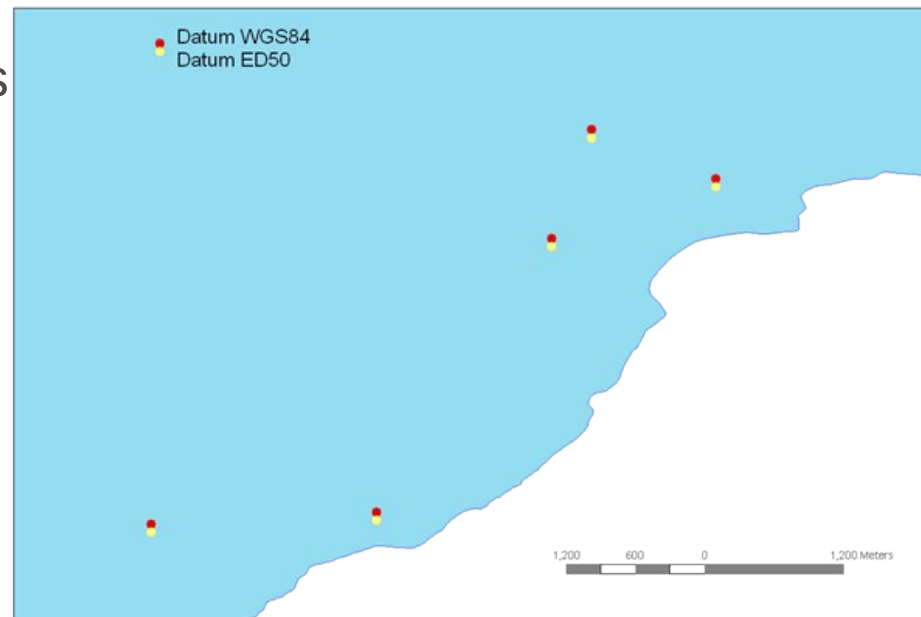
Conceitos de qualidade de dados

Conceitos

Incerteza e Erro

Existe sempre incerteza nos dados, ou seja, na forma como o observador apreende/compreende os dados.

O erro inclui as imprecisões e inexactidões. Pode ser aleatório ou sistemático. Um exemplo de erro sistemático é a definição errada de um datum.



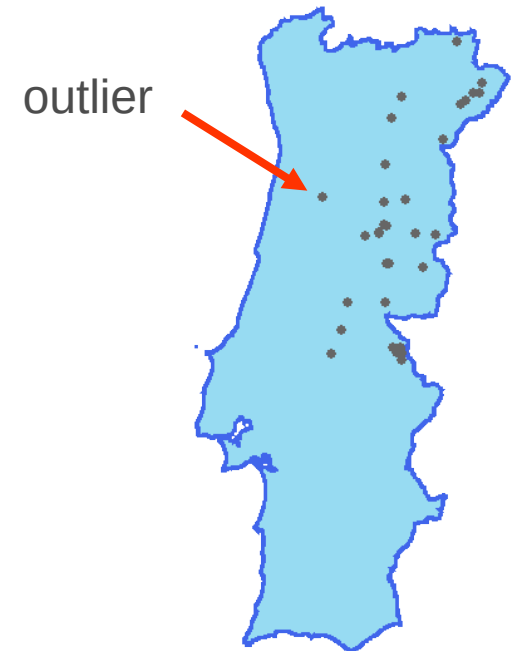
Conceitos de qualidade de dados

Conceitos

Validação e limpeza

A validação consiste em determinar se os dados são inexactos, incompletos ou não razoáveis. Exige a verificação de:

- formatos
- informação completa
- razoabilidade e consistência
- limites
- identificação de *outliers*



Conceitos de qualidade de dados

Conceitos

Validação e limpeza

A limpeza consiste na correcção dos erros. É importante garantir que os dados não são perdidos inadvertidamente. Muitas vezes é melhor reter quer os dados antigos como os corrigidos, lado a lado, na base de dados

Conceitos de qualidade de dados

Princípios de qualidade de dados

Visão, Política, Estratégia

Recomendação: Desenvolver um programa institucional de longo prazo

Ver Chapman, A. D. (2015). Princípios de Qualidade de

Dados: Chapman, A. D. (2015). Princípios de Qualidade de Dados. Versão 1.0 pt em Português lançada em abril 2015 e traduzida para pelo Nó Português do GBIF (www.gbif.pt) e pelo representante brasileiro do GBIF, SiBBr (Sistema de Informação sobre a Biodiversidade Brasileira, www.sibbr.gov.br). Versão original em Inglês lançada em jul 2005. Copenhagen: Global Biodiversity Information Facility. 81 pp. ISBN: 87-92020-58-5. Disponível on-line em

<http://www.gbif.org/document/80924>)

Conceitos de qualidade de dados

São princípios de **boa gestão de dados**

- não reinventar a roda
- observar a eficiência na recolha de dados e procedimentos de controlo de qualidade
- partilha de dados, informação e ferramentas
- utilizar padrões, ou desenvolver novos
- procurar parcerias e criação de redes
- reduzir a duplicação na recolha de dados e controlo de qualidade
- assegurar a criação de boa documentação e de metadados

Conceitos de qualidade de dados

Princípios de qualidade dos dados

As seguintes características devem ainda ser observadas:

- conjunto de **informação mínima completa**
- **consistência** – dados representados sempre da mesma forma – utilizar listas para a entrada de dados
- **flexibilidade** para comportar o dinamismo do processo
- **transparência** – os erros, se não corrigidos, devem ser evidenciados
- definir **controles**
- **minimizar a duplicação e reedição** dos dados

Controlo de qualidade de dados

Princípios de qualidade dos dados

As seguintes características devem ainda ser observadas:

- **manter os dados originais**
- a definição de categorias pode levar à perda de qualidade
- **documentação** – é um princípio chave, permite ao utilizador verificar o ajustamento dos dados ao seu objectivo
- **feedback** – definir mecanismos para promover o *feedback* dos utilizadores, e fazer com que estes se reflectam na qualidade dos dados
- **formação e treino** – pode aumentar largamente a qualidade dos dados. Deve estender-se desde os colectores até aos operadores de inserção de dados e gestores da base de dados.

Verificação de qualidade ao nível do portal nacional

Exemplo: <http://dados.gbif.pt/generic-hub/occurrences/3acad1e8-11b8-4439-bcb7-f5c661e72895>

Data quality tests

Test name	Result
Data are generalised [?]	⚠ Warning
Geodetic datum assumed WGS84 [?]	⚠ Warning
Basis of record not supplied [?]	✅ Passed
Basis of record badly formed [?]	✅ Passed
Missing name of person who identified the specimen/observation [?]	✅ Passed
Collector name unparseable [?]	✅ Passed
Missing catalogue number [?]	✅ Passed
Collection code not recognised [?]	✅ Passed
Institution code not recognised [?]	✅ Passed
Missing taxonomic rank [?]	✅ Passed
Name not supplied [?]	✅ Passed
Kingdom not recognised [?]	✅ Passed
Name not recognised [?]	✅ Passed
Invalid scientific name [?]	✅ Passed
Name not in national checklists [?]	✅ Passed
Decimal coordinates not supplied [?]	✅ Passed
Coordinates are transposed [?]	✅ Passed
Coordinates are out of range for species [?]	✅ Passed

Indexação dos dados pelo portal internacional

Exemplo: <https://www.gbif.org/occurrence/1404437769>

Taxon			
Term	Interpreted	Original	Remarks
Kingdom	Animalia	Animalia	
Phylum	Chordata	Chordata	
Class	Aves	Aves	
Order	Passeriformes	Passeriformes	
Family	Tyrannidae	Tyrannidae	
Genus	Sayornis	Sayornis	
Specific Epithet	phoebe	phoebe	
Scientific Name	Sayornis phoebe (Latham, 1790)	Sayornis phoebe	Altered
Rank	SPECIES		Inferred

Location			
Term	Interpreted	Original	Remarks
Country	United States	United States	
Country Code	US		Inferred
County	Chelan	Chelan	
Decimal Latitude	47.592446	47.5924463	Coordinate rounded
Decimal Longitude	-120.663443	-120.6634426	Coordinate rounded
Geodetic Datum	WGS84	WGS84	
Locality	Blackbird Island - Leavenworth	Blackbird Island - Leavenworth	
State Province	Washington	Washington	

Inventário de verificações de qualidade

<https://docs.google.com/spreadsheets/d/1tAfQUslzUNfa6Tn5Ezq2cfbMVB0QeXlyp7N0We2IBAY/edit#gid=0>

Data Quality Checks						
File Edit View Insert Format Data Tools Add-ons Help						
100% View only						
Code	Name	Creator	Description	Wiki		
Geospatial						
1	NEGATED_LATITUDE	GBIF	Record appears to be referencing a location in the wrong hemisphere	Wiki	Fix and	
2	NEGATED_LONGITUDE	GBIF	Record appears to be referencing a location in the wrong hemisphere	Wiki	Fix and	
3	INVERTED_COORDINATES	GBIF	Latitude and longitude have been transposed accidentally (typically bad database mapping)	Wiki	Fix and	
4	ZERO_COORDINATES	GBIF	Coordinates given as 0,0. Typically a result of bad default values for empty database fields	Wiki	Exclude	Make vi
5	COORDINATES_OUT_OF_RANGE	GBIF	Latitude >90 or <-90 and Longitude >180 or <-180	Wiki	Exclude	Make vi
6	UNKNOWN_COUNTRY_NAME	GBIF	Unrecognised or unparseable country name	Wiki	Report	
7	ALTITUDE_OUT_OF_RANGE	GBIF	Altitude greater than 10000m, or less than -100m	Wiki	Report	
8	BADLY_FORMED_ALTITUDE	GBIF	Free text string provided as altitude	Wiki	Report	
9	MIN_MAX_ALTITUDE_REVERSED	GBIF	Typically a column mapping issue	Wiki	Fix and	
10	DEPTH_IN_FEET	GBIF	Darwin core specifies metres should be used	Wiki	Fix and	
11	DEPTH_OUT_OF_RANGE	GBIF	Depth greater than 10000	Wiki	Report	
12	MIN_MAX_DEPTH_REVERSED	GBIF	Typically a column mapping issue	Wiki	Fix and	
13	ALTITUDE_IN_FEET	GBIF	Darwin core specifies metres should be used	Wiki	Fix and	
14	ALTITUDE_NON_NUMERIC	GBIF	Should be a numeric value in metres	Wiki	Report	
15	DEPTH_NON_NUMERIC	GBIF	Should be a numeric value in metres	Wiki	Report	
16	COUNTRY_COORDINATE_MISMATCH	GBIF	Coordinates outside the range for the reported country	Wiki	Report	
18	STATEPROVINCE_COORDINATE_MISMATCH	DM	Coordinates dont match the supplied state	Wiki	Report	
19	COORDINATE_HABITAT_MISMATCH	DM	Marine species reported in terrestrial area. Detection is also dependent on high-resolution coastline at the time of recording, e.g., estuaries can change quickly (LB).	Wiki	Exclude	Make vi
20	DETECTED_OUTLIER_ENVIRONMENTAL	DM	Record marked as outlier because it is outside the accepted environmental range/envelope of the species	Wiki	Optional	other re

Tipos de erros

ERROS TÉCNICOS

- **Completitude dos dados**
 - verificação de valores ausentes
- **Valores dentro de limites admissíveis**
 - data (21-10-2099?), altitude (3200 m em Portugal?), profundidade, latitude (98°N?) e longitude (210°E?)
- **Tipo de dado**
 - datas ou números incluídas como texto
- **Formato**
 - separadores decimais, formato de data

Tipos de erros

ERROS DE CONSISTÊNCIA

- **taxonómicos**
 - Se indica um registo ao nível da espécie, o nome científico contém o género e restritivo específico?
- **continuidade**
 - sequência temporal das datas de uma expedição
- **valores não admissíveis**
 - altitude de 1500 m no distrito de Lisboa?
- **geográficos**
 - localidade em Portugal com coordenada $-12,4^{\circ}$, $34,5^{\circ}$

Ferramentas para o controlo de qualidade

NOMES CIENTÍFICOS

Termo DwC	valor
scientificName	Acacia dealbata Link
kingdom	Plantae
phylum	Spermatophyta
class	Magnoliopsida
order	Fabales
family	Fabaceae
genus	Acacia
specificEpithet	dealbata
infraspecificEpithet	
taxonRank	species
scientificNameAuthors hip	Link

Verificações

Validade dos nomes, confrontando com cheklists:

GBIF Species Name Matching
<https://www.gbif.org/tools/species-lookup>

Global Names Resolver
<http://resolver.globalnames.org/>

Catalog of Life List Matching Service
<http://www.catalogueoflife.org/listmatching/>

iPlant TNRS
<http://tnrs.iplantcollaborative.org/>

Atomização de nomes

GBIF Name Parser
<http://tools.gbif.org/nameparser/>

Ferramentas para o controlo de qualidade

DADOS GEGRÁFICOS

Termo DwC	valor
country	Portugal
countryCode	PT
stateProvince	Coimbra
county	
municipality	Coimbra
locality	São Martinho do Bispo
decimalLatitude	40.21406
decimalLongitude	-8.44807
geodeticDatum	
coordinateUncertaintyInMeters	
coordinatePrecision	

Verificações

Verificação de coordenadas atribuídas:

- SIG: QGIS, ArcGIS
- online: Google Maps, OpenStreetMaps
- software de gestão de colecções: Specify

Georreferenciação

GeoLocate

<http://www.museum.tulane.edu/geolocate/web/WebGeoref.aspx>

Google Maps

<https://www.google.pt/maps>

Georreferenciação reversa

(obter nome da localidade e divisões administrativas a partir das coordenadas)

SpeciesLink InfoXY

<http://splink.cria.org.br/infoxy>

OpenStreetMap Nominatim

<http://nominatim.openstreetmap.org/reverse.php?format=html>

Conversão de coordenadas

Canadensys

<http://data.canadensys.net/tools/coordinates>

IGeoE

<http://www.igeoe.pt/coordenadas/>

Ferramentas para o controlo de qualidade

OUTROS DADOS

Termo DwC (exemplo)	valor
recordedBy	Rui Figueira
identifiedBy	R. Figueira
establishmentMeans	nativo
eventDate	2010-11-23

Verificações

Uso de vocabulários controlados

Ex: sex – male, female, hermaphrodite

Uso de listas autoritárias ou de referência

- lista de colectores/colaboradores da instituição
- lista de botânicos (Brummitt & Powell's (1992)
Authors of Plant Names, IPNI
<http://www.ipni.org/index.html>)

Datas num formato padrão ISO 8601:2004(E)

Ex.

year only: 2010

year and month: 2010-01

year, month, and day: 2010-01-17

year, month, day, and UTC time: 2010-01-17T09:26Z

year, month, day, and local time: 2010-01-17T09:33:59-0300

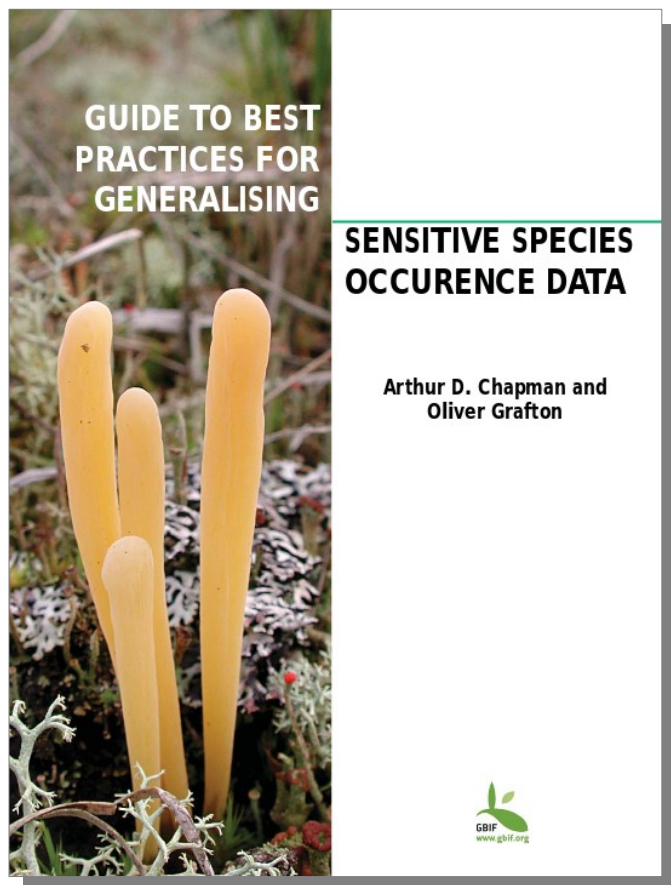
...

Conversão de datas

Canadensys date parsing

<http://data.canadensys.net/tools/dates>

Dados sensíveis



Para espécies sensíveis, devido à sua raridade, estatuto de ameaça ou valor económico, pode justificar-se remover ou generalizar, p.e., os dados de localização.

A remoção ou generalização pode ser aplicada a **campos de texto**, p.e.

- nomes de pessoas
- nome da localidade
- data de colheita
- nº de colector
- habitat
- proprietário
- nomes taxonómicos

Para as **coordenadas geográficas**, é possível diminuir a precisão por arredondamento, ou utilização de uma grelha maior

<https://www.gbif.org/document/80512>

Em qualquer dos casos, as omissões ou generalizações devem ser documentadas, nos dados e nos metadados. Os termos do DwC **informationWithheld** e **dataGeneralizations** servem para essa finalidade.

Gestão de coleções biológicas utilizando Specify 6

Obrigado pela atenção

Nó Português do GBIF
Instituto Superior de Agronomia
Herbário
Tapada da Ajuda
1349-017 Lisboa, Portugal

Tel: (+351) 213653165
email: node@gbif.pt
<http://www.gbif.pt>

O Nó Português é acolhido no ISA com o apoio da FCT.



INSTITUTO
SUPERIOR DE
AGRONOMIA
Universidade de Lisboa

FCT

Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Esta apresentação é publicada
segundo a licença CC-BY-SA

