ICEDIG.EU

Funded by the Horizon 2020 Framework Programme of the European Union

A linked project of

DiSSCo

Distributed System of Scientific Collections

Improved standardization of transcribed digital specimen data

Quentin Groom[1], Mathias Dillen[1], Helen Hardy[2], Sarah Phillips[3], Luc Willemse[4] and Zhengzhe Wu[5]

[1]Meise Botanic Garden, BE
[2]Natural History Museum, London, UK
[3]Royal Botanic Gardens, Kew, UK
[4]Naturalis, NL
[5]Luomus, FI

CETAF DWG/ISTC 2020

Zoom, April 21 2020

Background

- Data may be transcribed from specimens verbatim or interpreted. Interpreted data are considered more useful, but verbatim data still have use cases. These might not be compatible.

- Data may be absent or incomplete for different reasons. Right now, there is no easy way to find out why.

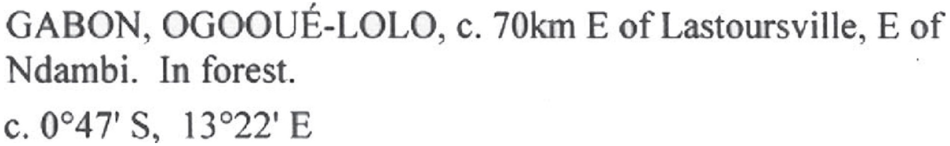
- The core properties of a specimen are what, **when**, **where**, which and who. All of these still have problems.

Verbatim transcription

- Use cases
  - Data cleaning
  - Information associated with original formatting (such as language)
  - Findability
  - Incomplete or unstandardized transcriptions
  - Old (bespoke) standards
  - Training automated transcription methods

Verbatim transcription

- Not necessarily compatible



GABON, OGOOUÉ-LOLO, c. 70km E of Lastoursville, E of
Ndambi. In forest.
c. 0°47' S, 13°22' E

Google Cloud Vision (raw)
GABON, OGOOUÉ-LOLO, c. 70km E of Lastoursville, E of
Ndambi. In forest.
c. 0°47' S, 13°22' E

Transcription platform (to DWC)

dwc:country	GABON
dwc:verbatimLocality	GABON, OGOOUÉ-LOLO, c. 70km E of Lastoursville, E of Ndambi
dwc:stateProvince	OGOOUÉ-LOLO
dwc:habitat	In forest.
dwc:verbatimCoordinates	c. 0°47' S, 13°22' E

Verbatim transcription

- What is verbatim data to us? What should it be?
  - Misfits
  - Raw data
  - Literal (yet classified) rendition

- How do we store and publish these?
  - Different methodologies, use cases
    -> different versions?

Unknown and incomplete

- Missing data:
  - Empty field
  - Negative indicator: S.L., S.D., inconnu, 00
  - Missing property (e.g. column in a CSV file)
  - `dwc:informationWithheld`

- Incomplete data:
  - [...], ?, ...
  - “Niet zichtbaar”, “covered by leaf”

Unknown and incomplete

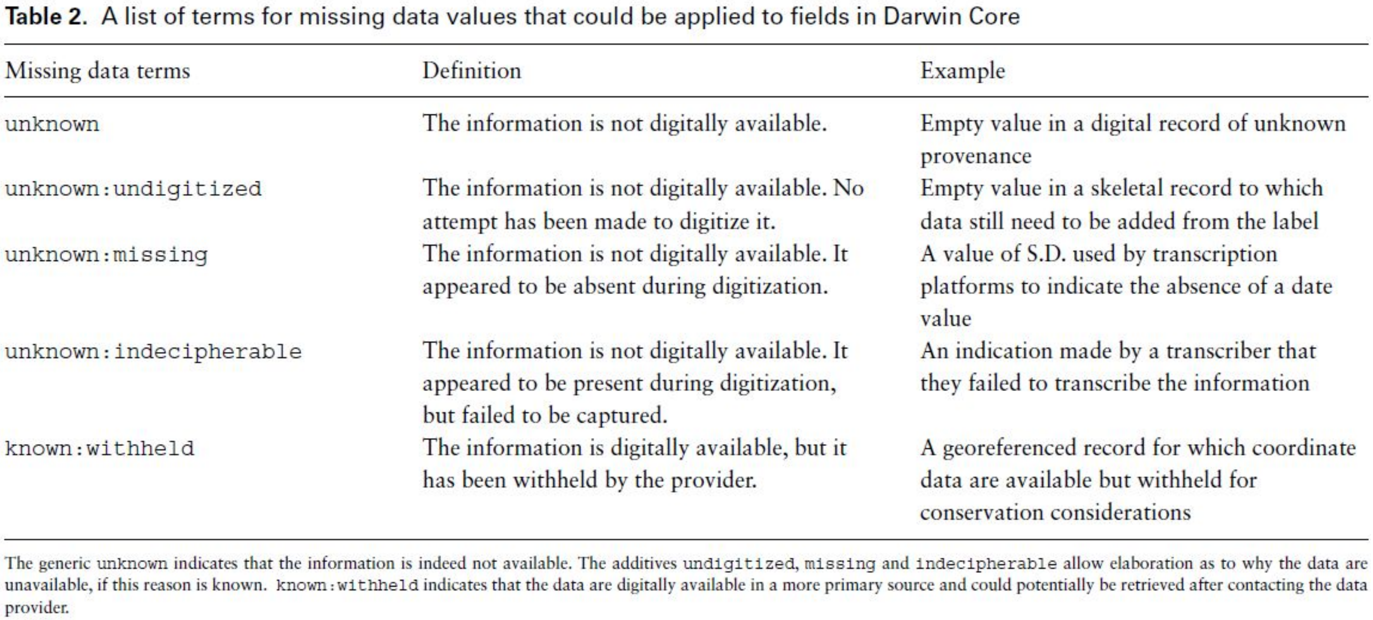


Table 2. A list of terms for missing data values that could be applied to fields in Darwin Core

Missing data terms	Definition	Example
`unknown`	The information is not digitally available.	Empty value in a digital record of unknown provenance
`unknown:undigitized`	The information is not digitally available. No attempt has been made to digitize it.	Empty value in a skeletal record to which data still need to be added from the label
`unknown:missing`	The information is not digitally available. It appeared to be absent during digitization.	A value of S.D. used by transcription platforms to indicate the absence of a date value
`unknown:indecipherable`	The information is not digitally available. It appeared to be present during digitization, but failed to be captured.	An indication made by a transcriber that they failed to transcribe the information
`known:withheld`	The information is digitally available, but it has been withheld by the provider.	A georeferenced record for which coordinate data are available but withheld for conservation considerations

The generic unknown indicates that the information is indeed not available. The additives `undigitized`, `missing` and `indecipherable` allow elaboration as to why the data are unavailable, if this reason is known. `known:withheld` indicates that the data are digitally available in a more primary source and could potentially be retrieved after contacting the data provider.

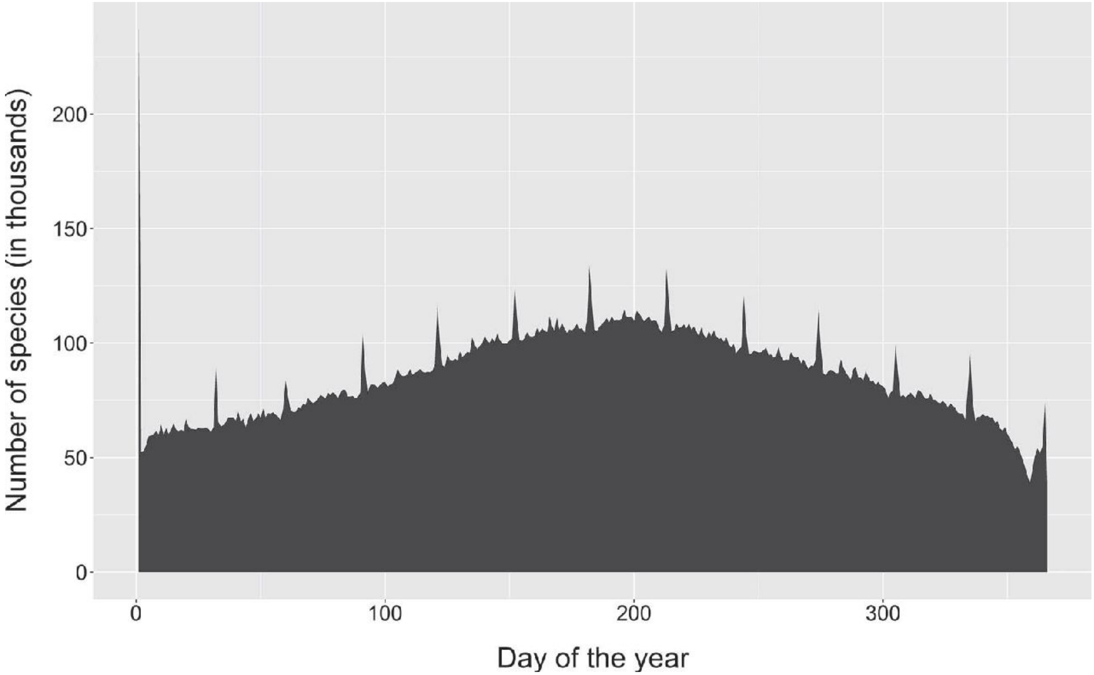
[] and [...] for uncertain and incomplete transcriptions

Core properties: when

- Partial dates
  - Yearless dates
  - ISO standard or DMY?

- Date ranges
  - Not always supported
  - Inferred ranges?

- Common corruption
  - Excel problems?



Core properties: where

- Grid systems
  - Conversion from grid to point/radius

- Provenance of georeferencing
  - Methodology
  - Uncertainty
  - System, datum used

- Application: Data cubes
  - https://github.com/trias-project/occ-cube

Funded by the Horizon 2020 Framework Programme of the European Union

THANK YOU FOR YOUR ATTENTION!

More info:

https://doi.org/10.1093/database/baz129

INNOVATION AND CONSOLIDATION FOR LARGE SCALE DIGITISATION OF NATURAL HERITAGE

www.icedig.eu