ENDANGERED WILDLIFE TRUST

Introduction to Metadata and Data Papers

ENDANGERED
WILDLIFE TRUST

What is metadata?

Metadata is an important, but often neglected, part of your dataset

- *Metadata is data about data, it describes data in a human-readable format*

- Non-trivial datasets should always be accompanied by metadata

- Metadata explains the why, how, when, where and who of your data

- It documents the important information that accompanies the dataset outside of your head

- It increases the value and longevity of your data, and is especially important for sharing of data

ENDANGERED
WILDLIFE TRUST
Protecting forever, together.

GBIF metadata

GBIF has 9 different sections on metadata (data about data)

- Description (general introduction)

- Temporal scope (date range)

- Geographic scope

- Taxonomic scope

- Methodology (how was the data collected)

- Contacts (who published the dataset, who are the main contact points for the dataset)

- Data description (languages used)

- GBIF registration (when and through which IPT was it published)

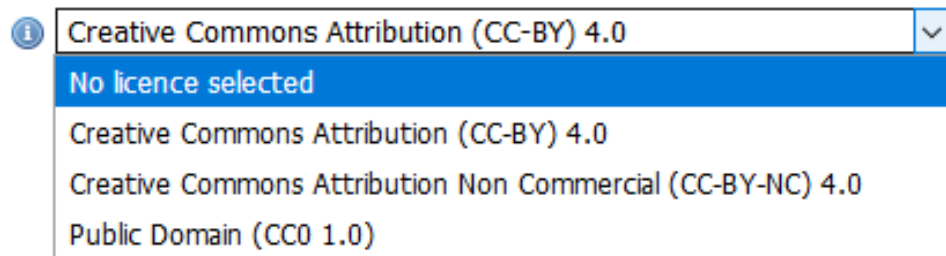- Citation (how the dataset should be cited if it is used)

ENDANGERED
WILDLIFE TRUST
Protecting forever, together.

Licencing

- When publishing a dataset on GBIF, a licence needs to be selected

Data Licence*

ⓘ Creative Commons Attribution (CC-BY) 4.0 ▾

No licence selected

Creative Commons Attribution (CC-BY) 4.0

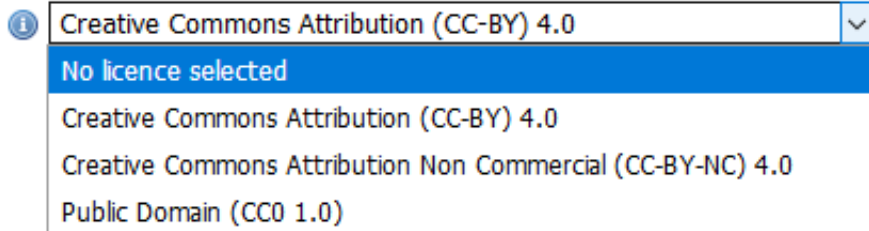Creative Commons Attribution Non Commercial (CC-BY-NC) 4.0

Public Domain (CC0 1.0)

- A Creative Commons (CC) license is a **public copyright license** that enables the free distribution of an otherwise copyrighted work. A CC license is used when an author wants to give people the right to share, use, and build upon a work that they have created, as long as they abide by the conditions that are specified in the license.

ENDANGERED WILDLIFE TRUST
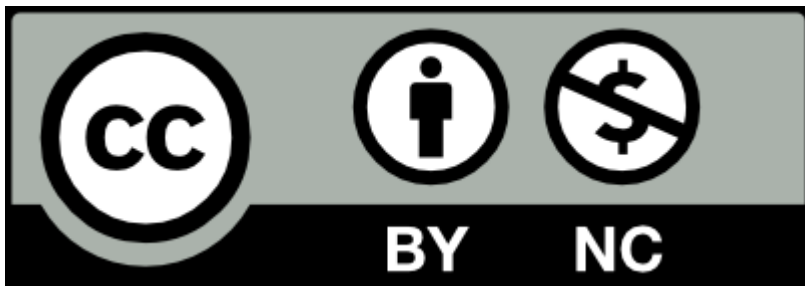Protecting forever, together.

Licencing

Data Licence*

Creative Commons Attribution (CC-BY) 4.0	⌄
No licence selected	
Creative Commons Attribution (CC-BY) 4.0	
Creative Commons Attribution Non Commercial (CC-BY-NC) 4.0	
Public Domain (CC0 1.0)	

- These logos indicate different CC licences

- The most open licence is a CC0 or public domain licence, which allows the data user to do anything with the data

- Typically, however, data publishers like to be acknowledged for their work. They therefore chose a CC-BY, or attribution licence, which requires that a data user acknowledge the data provider

- Lastly, a CC-BY-NC, or attribution non-commercial licence also requires that the data user does not use the data for commercial purposes

ENDANGERED WILDLIFE TRUST
Protecting forever, together.

What is a Data Paper?

- A searchable metadata document, describing a particular dataset or a group of datasets

- Published in the form of a peer-reviewed article in a scholarly journal

- Unlike a conventional research article, the primary purpose of a data paper is to describe data and the circumstances of their collection, not to report hypotheses, data analysis or provide discussion and conclusions

The African Crane Database (1978-2014): Records of three threatened crane species (Family: Gruidae) from southern and eastern Africa

Tanya Smith[‡], Samantha Page-Nicholson[§], Kerryn Morrison[‡], Bradley Gibbons[§], M. Genevieve W. Jones[|], Mark van Niekerk[¶], Bronwyn Botha[#], Kirsten Oliver[¤], Kevin McCann[¤], Lizanne Roxburgh[§]

‡ International Crane Foundation / Endangered Wildlife Trust Partnership, Johannesburg, South Africa
§ Endangered Wildlife Trust, Johannesburg, South Africa
| Percy FitzPatrick Institute of African Ornithology, DST/NRF Centre of Excellence, University of Cape Town, Cape Town, South Africa
¶ Jean Monnet University, Saint Etienne, France
# Flower Valley Conservation Trust, Gansbaai, South Africa
¤ Wildlands Conservation Trust, Hilton, South Africa

Abstract

Background

The International Crane Foundation (ICF) / Endangered Wildlife Trust's (EWT) African Crane Conservation Programme has recorded 26 403 crane sightings in its database from 1978 to 2014. This sightings collection is currently ongoing and records are continuously added to the database by the EWT field staff, ICF/EWT Partnership staff, various partner organizations and private individuals. The dataset has two peak collection periods: 1994-1996 and 2008-2012. The dataset collection spans five African countries: Kenya, Rwanda, South Africa, Uganda and Zambia; 98% of the data were collected in South Africa. Georeferencing of the dataset was verified before publication of the data. The

---

Tracking data from nine free-roaming Cheetahs (*Acinonyx jubatus*) collared in the Thabazimbi area, Limpopo Province, South Africa

Kelly Marnewick[‡,§], Samantha Page-Nicholson[‡], Lizanne Roxburgh[‡], Michael J. Somers[§]

‡ Endangered Wildlife Trust, Johannesburg, South Africa
§ Centre for Wildlife Management, University of Pretoria, Pretoria, South Africa

Abstract

Background

In partnership with the University of Pretoria, the Endangered Wildlife Trust's Carnivore Conservation Programme collared six male and three female free-roaming Cheetahs (*Acinonyx jubatus*) in the Thabazimbi area in Limpopo Province, South Africa. This study was undertaken to determine the spatial ecology of free-roaming Cheetahs that occur outside of formal protected areas on private ranchland, where they frequently come into conflict with, and are sometimes killed by, private landowners. The data were collected between September 2003 and November 2008, resulting in a total of 3165 location points (65 points from VHF collars and 3100 from GPS collars) for nine individual Cheetahs.

New information

This dataset provides distribution information about this Vulnerable species occurring outside of protected areas within South Africa. The dataset has been published to the

What are its purposes and benefits?

- to provide a citable journal publication that brings scholarly credit to the data publishers

- to describe the data in a structured human-readable form

- to bring the existence of the data to the attention of the scholarly community

- to increase the visibility, usability and credibility of the data resources you publish

Who publishes Data Papers?

- Pensoft is one of the main publishers that has been working with GBIF e.g. Biodiversity Data Journal

- All their papers are peer-reviewed, open access, rapidly published, available online and in print

- Pensoft journals charge for publishing, but rates are low relative to other open access journals (€200+, but there are discounts for African authors)

- Additional data paper publishers listed on the GBIF website (www.gbif.org/data-papers) – 20+

ENDANGERED WILDLIFE TRUST
Protecting forever, together.

Journals that publish data papers

African Biodiversity and Conservation (ABC) Journal
http://www.abcjournal.org
Data papers are a new article type, and are not advertised on their website yet
No publication charges

Journals that publish data papers

Scientific Data

(published by Springer Nature)

www.nature.com/sdata/

Scientific Data is a peer-reviewed, open-access journal for descriptions of scientifically valuable datasets, and research that advances the sharing and reuse of scientific data. They aim to **promote wider data sharing and reuse**, and to **credit those that share**.

Data descriptor = Data paper

Publication costs: €1,305

Types of datasets suitable for data papers

- Bigger is **probably** better
- The dataset should be of some **significance,** e.g. long time scale, first dataset of its kind, entire database of an institution or a project
- Bear in mind that publishers of data papers will also be looking to improve their **publication ratings**
- Like any scientific publication, the decision to publish or not lies with the journal editors
- Look at other **examples** of published data papers for further guidance

Types of datasets suitable for data papers

- Data papers can describe multiple datasets that were collected by more than one institution or project

- Datasets should have a geographic or taxonomic focus, or another thematic focus, e.g., threatened species, invasive species

- Datasets should form the basis of a research story, so that a coherent scientific publication can be developed

Steps in the publishing process

1
- Clean and standardize the dataset
- Format the dataset using DarwinCore

2
- Prepare a metadata document, which can later be converted to a data paper

3
- Upload the data and metadata to GBIF*, using a GBIF node, or other openly accessible platform

4
- Convert metadata to a data paper, submit to a journal, and publish

Publishing the dataset

- Most challenging part of the process is preparing the dataset itself for publication

- Data should be published and publically accessible e.g., on GBIF, via the Integrated Publishing Toolkit or IPT (through SANBI or EWT)

- Once the dataset is published, it will have a DOI (Digital Object Identifier), to which the data paper will refer e.g. 10.15468/cadrut

Writing the metadata

- The metadata that is written when a dataset is published to GBIF is the basis of a data paper

- A Word document template extracted from the Pensoft journals is available (DataPaperTemplate.docx)

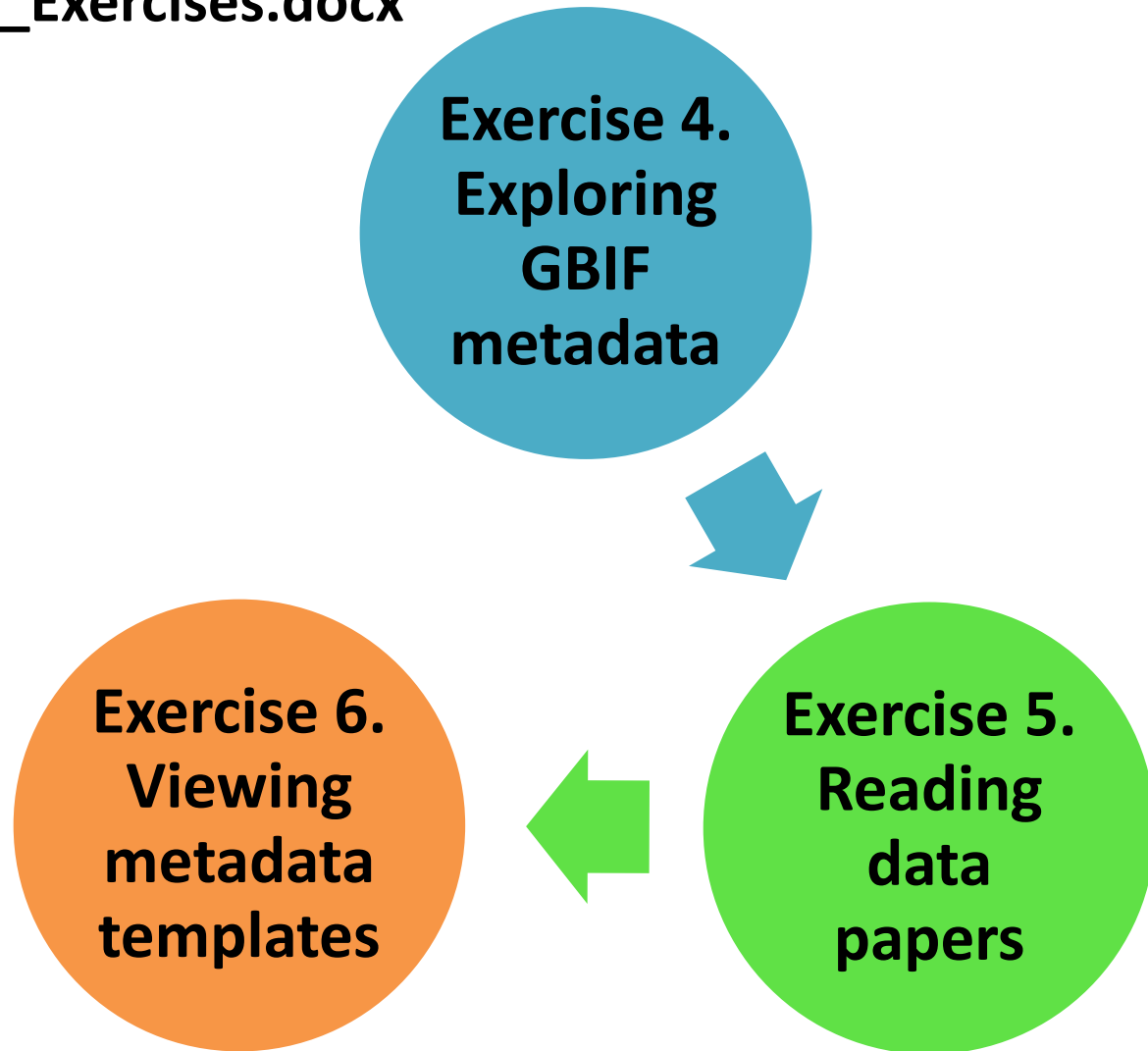- Scientific data also has a downloadable template for data descriptors https://www.nature.com/sdata/publish/submission-guidelines

Practical exercises

File: Day1_GBIF_Exercises.docx

Exercise 4. Exploring GBIF metadata

Exercise 5. Reading data papers

Exercise 6. Viewing metadata templates