

Organising & standardising research data that underpin your publication

Willeke de Haan, KU Leuven Research Coordination Office

VIB RDM training, Leuven
13-14 November 2023

Overview

- *Organise and structure data and documentation files linked to a publication during and after research*
- *Logical, structured and descriptive file names*
- *Open / standard file formats*
- *File versioning in a project*
- *Data standards to make data interoperable, commonly understandable and reusable (the I and R of FAIR)*

Organise / structure files

- Different options exist
- Here examples of good practices that researchers at KU Leuven use
- Find what works for you, in line with technical knowledge / support available and your data collection methods and active data storage system
- Centre organisation around:
 - Research project
 - Research paper

Folder structure

- File Explorer, OneDrive, MS Teams, ...
- Develop a structure organised by:
 - Paper, Project, Researcher, Experiment, Instrument
- Folders should:
 - follow a structure with folders and subfolders that correspond to the project design and workflow
 - have a self-explanatory name that is only as long as is necessary
 - have a unique name
- Consider read / write access to folders for colleagues / collaborators
- Good practice: ReadMe file in top folder
- When paper is published (or end of project): package structure and files into zip bundle and move to archival storage

An example:

project/	
code/	code needed to go from input files to final results
data/	raw and primary data (never edit!)
raw_external/	
raw_internal/	
meta/	
doc/	documentation of the study
intermediate/	output files from intermediate analysis steps
logs/	logs from the different analysis steps
notebooks/	notebooks that document your day-to-day work
results/	output from workflows and analyses
figures/	
reports/	
tables/	
scratch/	temporary files that can safely be deleted or lost
README.txt	file and folder description

Source: https://rdmkit.elixir-europe.org/data_organisation

In our department, each project has a designated folder. When a new project is started, we make a new folder. There we keep raw data, syntax files, questionnaires, ethical approval, etc. All researchers have access to the shared drive and to all folders.

Example

ProjectName

README file that lists the folder structure explaining the content of each folder, incl **date of last update**

- Data:** Separate folder for all fixed **raw data** used in the project that do not change throughout the project
 - Subfolders** for individual datasets + information file explaining methods used and relevant parameters to create the data
- Preprocessing:**
 - Subfolders** for **each step** in preprocessing, in **chronological** order with **date at start of subfolder name** (track sequence of analytical steps)
 - Within each **subfolder** ReadMe file with links to relevant wiki pages, source code files
 - Within each **subfolder** scripts developed and resulting intermediate data together: script describes how this version of data was processed from the previous version.
- Analysis:**
 - Subfolders** for each **experiment**, in chronological order by using the date at the start of a subfolder name
 - Scripts**
 - Outputs**
 - Results**
- Sourcecode:** self-developed code or models + ReadMe file with link to git repository
 - Active:** sandbox for active development of code where best practices are not expected
 - Release:** final version of code
- Doc:** documentation files for the project, incl **papers**
- Datapoints:** folder for **intermediate datasets** for use in future; if during preprocessing / analysis you obtain a useful intermediate dataset you may use in future, copy it here to a subfolder, **with relevant script**; include **ReadMe** file to explain how datapoint was created
- Temp:** Folder for draft versions of data and codes you want to keep. Move them here from “Analysis” if no longer relevant there

Record file

- A textual or tabular record file can list all data and documentation files of a project, paper, etc. This can record standard information for each dataset:
 - Unique ID
 - Dataset name
 - Description
 - Origin
 - Owner
 - Person responsible
 - Purpose, e.g. project name
 - Storage location, e.g. where on server, OneDrive, etc.
 - Contains personal data Y/N
 - Size / volume
 - Access: who has / needs access to the data

Within our research group we mostly develop algorithms for simulations. Every researcher has to keep a register in the form of a Word file that lists which code repositories (on GitLab or GitHub) are used (URL) and where data files are stored. These registers are available on a sub-website and colleagues have read access.

eLab Notebook

- eLab notebook systems (e.g. LabCollector, eLabFTW) can be used as registry to note and point to - or include - data files, protocols, experiments, documentation, samples used, etc.
- Some allow files to be uploaded into the notebook or file paths can be used
- When paper is published: export overview of all experiments, descriptions and links to data files to single PDF file as documentation
- End of a project: export all data, annexes and documents to zip files for archiving

A lab uses eLabFTW as electronic lab notebook. Researchers use tablets in the lab to record their experiments. They use templates for different experiments and included databases, cell lines, reference data and protocols that are frequently used within the lab, so researchers can simply point to those. Researchers can set up 'to do' lists at the start of their experiment, and sign those off as they proceed. Each PhD researcher has to use the electronic lab book to record all steps in his/her experiment, upload or point to protocols, import data files, etc. Supervisors have read access, can include comments and use the lab notebook for PhD progress discussions. When a PhD project ends, the lab exports a zip file (bundle) of the entire lab notebook of that person for archiving. This contains the lab book as single PDF file, with all annexes organised in folders.

Data management plan

- Records all data files generated and used for a researcher project / paper
- Describes how data are used / generated, who is responsible, where data are held (stored), etc.
- Details all accompanying documentation files and any relevant ethical, legal or compliance details

In our research group we work with existing third party datasets that we acquire. Before a dataset is imported for use, a data management plan has to be written by the researcher that provides information on what the dataset will be used for, where it was obtained, who needs to have access to the dataset, which licence conditions apply, etc.

File naming

- Develop a **logical** structure for **meaningful** file names
- Order 4-7 elements from generic to specific
- Suggested elements:
 - Project / experiment name, acronym or number
 - Creator name or initials
 - Date of creation: use ISO8601 format YYYYMMDD (and if needed time HHMMSS)
 - Type of data: sample ID
 - Version number: v01, v02, 00.01, 01.01 (leading zeros ensure correct sorting of files)
 - Location
- No spaces: use underscore (_), hyphen (-) or Capitalized letters to separate elements
- Avoid special characters such as “/ \ : * ? ” < > [] & \$
- Independent of the location of the file on a computer
- Include a txt-file that explains your naming convention in your documentation

Dataset Challenges and Opportunities for Academic Parents during COVID-19

Eva Lantsoght

Anonimized dataset of the survey on the impact of COVID-19 on academic parents. Participants who did not give consent were filtered out as well.



Files (210.1 kB)		
Name	Size	
dataset for Zenodo.xlsx	210.1 kB	Download
md5:a2c2da0f619e5366f57314929ac7fa3f		
Citations 0		
Show only: <input type="checkbox"/> Literature (0) <input type="checkbox"/> Dataset (0) <input type="checkbox"/> Software (0) <input type="checkbox"/> Unknown (0)		
<input type="checkbox"/> Citations to this version		

File naming examples

Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020

- File name: 20201202_HB_EXP2_HEL_DATA_V03.xls
- Explanation:
Date_ProjectAbbreviation_ExperimentNumber_Location_TypeOfData_VersionNumber

Cropped image of an ant head taken on the third of December 2020 by Meg Megson

- File name: 20201203_MM_HEAD_CROPPED_V1.psd
- Explanation: Date_CreatorData_Type_Modification_Version

Version 4 of the survey procedures for the British Dental Health Survey.

- BDHS_SurveyProcedures_00-04.pdf
- Explanation: Project acronym_Type_version number

Source: https://rdmkit.elixir-europe.org/data_organisation

Batch file renaming

- Need to rename large amounts of file names?
 - Images from digital cameras with automatically assigned files names/numbers
 - Default filenames generated by proprietary software or instruments
 - Removing spaces, odd characters, etc
 - Add meaningful elements to file name, e.g. project acronym, date, etc.
- Use a batch renaming tool for consistent, structured renaming
 - Bulk Rename Utility (Windows)
 - Renamer (Mac)
 - Gnome Commander (Linux)
- Demo: [Renaming Files Using Bulk Rename Utility - YouTube](#)

Batch rename example

Bulk Rename Utility

C:\Users\u0137709\Downloads\run_data

Name	New Name	Size	Created	Modified
Person1_run105.mat	MALL_Person01_run105_220603.mat	607,67 KB	3/06/2022 16:35:42	3/06/2022 16:35:42
Person1_run11.mat	MALL_Person01_run11_220603.mat	607,63 KB	3/06/2022 16:35:42	3/06/2022 16:35:42
Person1_run115.mat	MALL_Person01_run115_220603.mat	608,07 KB	3/06/2022 16:35:42	3/06/2022 16:35:42
Person1_run90.mat	MALL_Person01_run90_220603.mat	607,16 KB	3/06/2022 16:35:42	3/06/2022 16:35:42
Person1_run95.mat	MALL_Person01_run95_220603.mat	606,84 KB	3/06/2022 16:35:42	3/06/2022 16:35:42
Person12_Run115.mat	MALL_Person12_Run115_220603.mat	610,95 KB	3/06/2022 16:43:14	3/06/2022 16:43:14
Person2_run10.mat	MALL_Person02_run10_220603.mat	601,01 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person2_run75.mat	MALL_Person02_run75_220603.mat	601,61 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person2_run80.mat	MALL_Person02_run80_220603.mat	603,42 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person2_run85.mat	MALL_Person02_run85_220603.mat	600,76 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person2_run90.mat	MALL_Person02_run90_220603.mat	608,29 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person2_run95.mat	MALL_Person02_run95_220603.mat	606,70 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person3_run10.mat	MALL_Person03_run10_220603.mat	601,75 KB	3/06/2022 16:35:44	3/06/2022 16:35:44
Person3_run105.mat	MALL_Person03_run105_220603.mat	607,63 KB	3/06/2022 16:35:46	3/06/2022 16:35:46
Person3_run11.mat	MALL_Person03_run11_220603.mat	603,37 KB	3/06/2022 16:35:46	3/06/2022 16:35:46
Person3_run115.mat	MALL_Person03_run115_220603.mat	602,25 KB	3/06/2022 16:35:46	3/06/2022 16:35:46
Person3_run90.mat	MALL_Person03_run90_220603.mat	600,80 KB	3/06/2022 16:35:46	3/06/2022 16:35:46
Person3_run95.mat	MALL_Person03_run95_220603.mat	602,42 KB	3/06/2022 16:35:46	3/06/2022 16:35:46
Person4_run10.mat	MALL_Person04_run10_220603.mat	603,12 KB	3/06/2022 16:35:46	3/06/2022 16:35:46
Person4_run105.mat	MALL_Person04_run105_220603.mat	600,85 KB	3/06/2022 16:35:48	3/06/2022 16:35:48
Person4_run11.mat	MALL_Person04_run11_220603.mat	601,57 KB	3/06/2022 16:35:48	3/06/2022 16:35:48
Person4_run115.mat	MALL_Person04_run115_220603.mat	600,41 KB	3/06/2022 16:35:48	3/06/2022 16:35:48

RegEx (1) ☒ R Match Replace ☐ Inc.Ext. ☐ Simple ☐ v2

Name (2) ☒ R Name

Replace (3) ☒ R Replace With ☐ Match Case ☐ First

Case (4) ☒ R Same Excep.

Remove (5) ☒ R First n Last n From to Chars Words Crop Before ☐ Digits ☐ High ☐ Trim ☐ D/S ☐ Accents ☐ Chars ☐ Sym. ☐ Lead Dots

Add (7) ☒ R Prefix Insert at pos. Suffix ☐ Word Space

Auto Date (8) ☒ R Mode Type Fmt Sep. Seg. Custom ☐ Cent. Off.

Experiment measuring vertical dynamic running load with 13 treadmill users.

File renaming for 78 files:

- Find single digits in filename and add leading zero
- Add project name 'MALL' as prefix
- Add creation date as suffix, with underscore

Exercise: folder structure & file naming

- The role of basal epithelial cells for small airway loss and epithelial injury in chronic lung disease.
 - Design a suitable folder structure for this research project
 - What would be useful elements for file names?

Exercise: folder structure & file naming

Folders

- Data
 - Scans
 - Processed data
 - Images
 - Measures
- Doc
 - SOP
 - ...
 - Papers
- Code
- Results

Elements

- Whole, Part
- Lung
- Sample number
- Collection date
- Donor pseudoniem
- Replication number
- Control, COPD, IPF
- Mild, moderate, severe

Open / standard file formats

- Use or convert to open / standard file formats when you can
- For long-term access and use of research data
- Good source: fairsharing.org
 - Images: TIFF, JPEG 2000, PNG, GIF
 - Text: XML, PDF/A, HTML, JSON, TXT, RTF
 - Containers: TAR, ZIP
 - Databases: XML, CSV, JSON
 - Video: MPEG (mp4), AVI
 - Sounds: WAVE, AIFF, MP3, FLAC
 - Statistics: DTA, POR, SAS, SAV
 - Tabular data: CSV, TXT
 - 3D: X3D, C3D
 - Microscopy: [OME Next Generation File Format](#), Bio-formats conversions
 - Neuroimaging: [DICOM](#), [Nifti](#)
 - Mass spectrometry: [mzML](#)
 - Sequencing data: [FASTA](#), [FASTQ](#)

File versioning

- Manage multiple versions of the same file / document
- Enable reverting to an earlier version
- Easy methods for small demands of versioning:
 - File naming
 - Version control log in file
 - Cloud storage service with automatic file versioning, e.g. OneDrive
- For automatic management of versioning, conflict resolution and back-tracing capabilities, use a proper version control software such as Git, e.g. GitHub, GitLab, BitBucket: next presentation this afternoon

Data standards

- Use data standards to make data interoperable, easier to understand (by multiple communities) and reusable more widely
 - International, common standards
 - Community standards

Lung Neoplasms MeSH Descriptor Data 2017

Details Qualifiers **MeSH Tree Structures** Concepts

MeSH Heading Lung Neoplasms
Tree Number(s) C04.588.894.797.520
C08.381.540
C08.785.520
Unique ID D008175
Annotation coord IM with histol type of neopl (IM)
Scope Note Tumors or cancer of the LUNG.
Entry Version LUNG NEOPL
Entry Term(s) Cancer of Lung
Cancer of the Lung
Lung Cancer
Neoplasms, Lung
Neoplasms, Pulmonary
Pulmonary Cancer
Pulmonary Neoplasms
NLM Classification # WF 658
See Also Carcinoma, Non-Small-Cell Lung
Carcinoma, Small Cell
Date Established 1966/01/01
Date of Entry 1999/01/01
Revision Date 2012/07/03

Neoplasms [C04]
Neoplasms by Site [C04.588]
Thoracic Neoplasms [C04.588.894.797.520]
Respiratory Tract Neoplasms [C04.588.894.797.520.109]
Lung Neoplasms [C04.588.894.797.520]
Bronchial Neoplasms [C04.588.894.797.520.109]
Multiple Pulmonary Nodules [C04.588.894.797.520.237]
Pancoast Syndrome [C04.588.894.797.520.734]
Pulmonary Blastoma [C04.588.894.797.520.867]
Pulmonary Sclerosing Hemangioma [C04.588.894.797.520.933]
Pleural Neoplasms [C04.588.894.797.640]
Tracheal Neoplasms [C04.588.894.797.760]

Entry terms (related terms) shows a different definitions which may exist for a given term.

MeSH Tree Structures: Broader terms are listed up in a hierarchy. Narrower terms are listed below in a hierarchy.

Standards

International

- ISO 8601 standards for date / time
- ISO 3166 standard for country codes
- Getty Thesaurus for geographical names

Community

- DICOM MRI data
- NACE code: Statistical classification of economic activities in European Community
- Standard International Age Classification, UNStat 1982

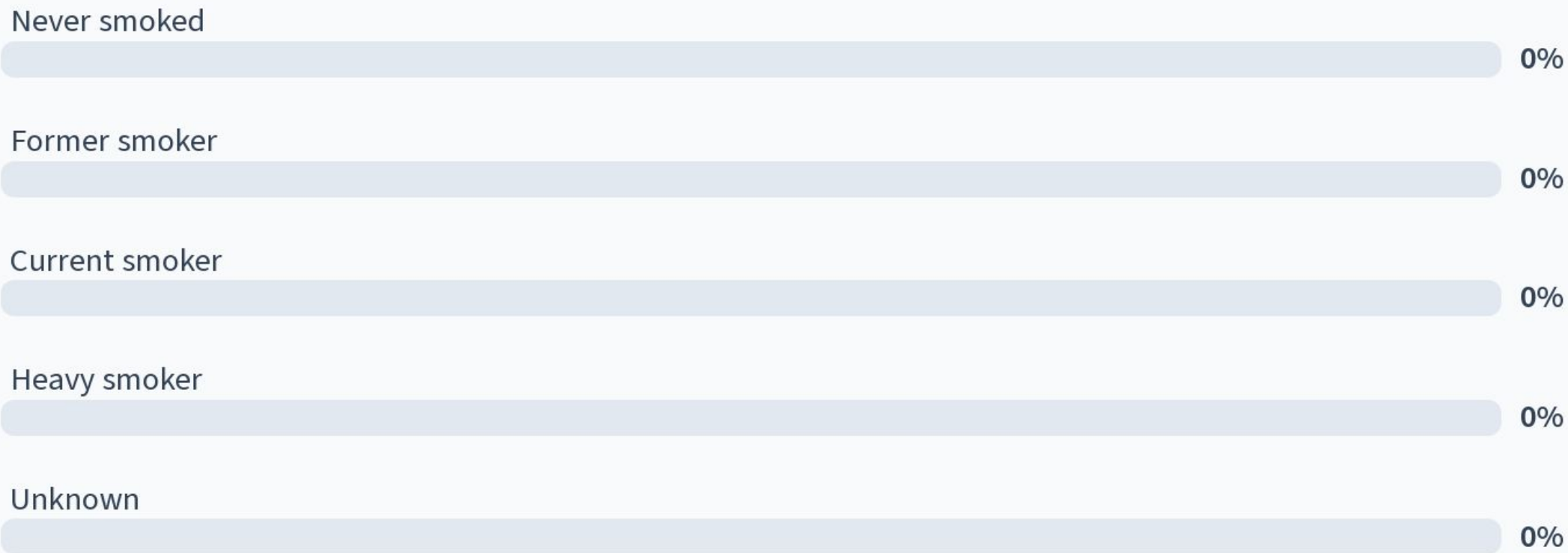
C. Learning and education services	
1. Enrolment in regular and adult education	2-4; 5 y.gr. 5-24; 10 y.gr. 25-64; 65+
2. Educational attainment	5 y.gr. 15-24; 10 y.gr. 25-64; 65+
3. Illiteracy	5 y.gr. 10-24; 10 y.gr. 25-64; 65+
G. Health, health services and nutrition	
1. Morbidity and handicaps (for mortality see I)	u 1; 1-4; 10 y.gr. 5-74; 75+
2. Usage of health services	u 1; 1-4; 10 y.gr. 5-74; 75+
3. Food consumption	u 1; 1-4; 10 y.gr. 5-74; 75+
4. Malnutrition	u 1; 1-4; 10 y.gr. 5-74; 75+
D. Earning activities and the inactive	
1. Labour force participation	u 15; 5 y.gr. 15-24; 10 y.gr. 25-54; 5 y.gr. 55-74; 75-84; 85+
2. Employment/unemployment/underemployment	u 15; 5 y.gr. 15-24; 10 y.gr. 25-54; 5 y.gr. 55-74; 75-84; 85+

When age classification categories are applied consistently at an international level, datasets can be easily linked, combined and compared. But: different disciplines / purposes will need different categories !



PollEv.com/willekedehaan890

To aks participants about their history of using tobacco, which is probably not a value in the Common Data Element (CDE) "Smoke History Status"?



Which is a unique identifier for aspirine?

DB00945 (Drugbank online)

0%

CHEBI:15365 (Chemical Entities of Biological Interest dictionary)

0%

CO1405 (KEGG Compound Database)

0%

Acetylsalicylic acid

0%

What is a standard for temperature?

20 °C

0%

293 K

0%

68 F

0%

Room temperature

0%

Which answer options are suitable to ask about how much participants exercise

not; once in a while; sometimes; often; all the time

0%

<1 hour; 1-2 hours; 2-4 hours; >4 hours per week

0%

.... hours per week

0%

.....

0%

light exercise, moderate exercise, vigorous exercise

0%

Which standards do you use in your research?

Nobody has responded yet.

Hang tight! Responses are coming in.

GBIF & Darwin Core

Appears in Datasets

APPEARS IN 69 CHECKLIST DATASETS:	APPEARS IN 545 OCCURRENCE DATASETS:
GBIF Backbone Taxonomy <i>As Canis lupus Linnaeus, 1758</i>	Répartition historique du loup en France métropolitaine View occurrences
Catalogue of Life Checklist <i>As Canis lupus Linnaeus, 1758</i>	NSW BioNet Atlas View occurrences
The European Nucleotide Archive (ENA) taxonomy <i>As Canis lupus</i>	Norwegian Biodiversity Information Centre - Other datasets View occurrences
World Register of Marine Species <i>As Canis lupus Linnaeus, 1758</i>	Swiss National Mammal Databank: Larger Carnivores Monitoring Program (KORA) View occurrences
Integrated Taxonomic Information System (ITIS) <i>As Canis lupus Linnaeus, 1758</i>	iNaturalist Research-grade Observations View occurrences
International Barcode of Life project (IBOL) Barcode Index Numbers (BINs) <i>As Canis lupus Linnaeus, 1758</i>	Fauna Atlas N.T. View occurrences
Global Names Usage Bank <i>As Canis lupus Linnaeus, 1758</i>	UAM Mammal Collection (Arctos) View occurrences
TAXREF <i>As Canis lupus Linnaeus, 1758</i>	SA Fauna (BDBSA) View occurrences
The Paleobiology Database <i>As Canis lupus Linnaeus, 1758</i>	Victorian Biodiversity Atlas View occurrences

Darwin Core standard

Record-level Terms	Dublin Core terms, institutions, collections, nature of data record	Simple Darwin Core (flat)
Occurrence	evidence of species in nature, observers, behavior, associated media, references.	
Event	sampling protocols and methods, date, time, field notes	
Location	geography, locality descriptions, spatial data	
Identification	linkage between Taxon and Occurrence	
Taxon	scientific names, vernacular names, names usages, taxon concepts, and the relationships between them	
GeologicalContext	geologic time, chrono-stratigraphy, biostratigraphy, lithostratigraphy	
ResourceRelationship	explicit relationships between identified resources (e.g., one organism to another, taxon to location, etc.)	Generic Darwin Core (relational)
MeasurementOrFact	measurements, facts, characteristics, assertions, references	

Community standard: biodiversity data

Wolf observations Flanders 2022

Waarnemingen.be NL Log in of registreer ▾

Invoeren ▾ Ontdek ▾ Projecten ▾ Over ons ▾ Community ▾ 🔍

Wolf

Canis lupus LINNAEUS, 1758

Soort Zoogdieren Canidae Canis *Canis lupus* ▾ Soort

Details Waarnemingen Kaarten Foto's Geluiden Statistieken Op/in Namen

2022-01-01 - 2022-11-20 Alle provincies Zoek Filter Wis filters ☒ Toon geavanceerd

Waarnemer Locatie Alle geslachten ☐ Alleen goedgekeurd ☐ Toon nul-waarnemingen ☐ Alleen onzeker

Alle levensstadia ▾ Alle activiteiten ▾ met fotoval ▾

Datum	Aantal	Locatie	Waarnemer	
2022-11-08	1 adult, met fotoval	LI	vervaagd	✅ 🔒 📷
2022-10-31	1 met fotoval	LG	vervaagd	✅ 🔒 📷 🗨
2022-04-08	2 ♂ met fotoval	LI	vervaagd	🔒 🗨 📷 🗨

GBIF platform wolf data 2022

SPECIES | ACCEPTED


Canis lupus Linnaeus, 1758

Published in: Syst. Nat., 10th ed. vol.1 p.39 source: The Integrated Taxonomic Information System
Arctic Wolf in English Basionym: *Lupus lupus* Linnaeus, 1758


OVERVIEW 4 TREATMENTS METRICS REFERENCE TAXON ∞

145,205 OCCURRENCES 42 INFRASPECIES

9,740 OCCURRENCES WITH IMAGES



122,144 GEOREFERENCED RECORDS



Generated an hour ago © OpenStreetMap contributors, © OpenMapTiles, GBIF.

Any year 2022-2022 EXPLORE

Because biodiversity data are collected worldwide using the same data standards, collecting the same attributes and variables, they can be combined into large comparable datasets on the GBIF platform.

Compatible dates: Linking 5 minute weather data with time of sunrise / sunset

TimeStamp in both datasets facilitates interoperability

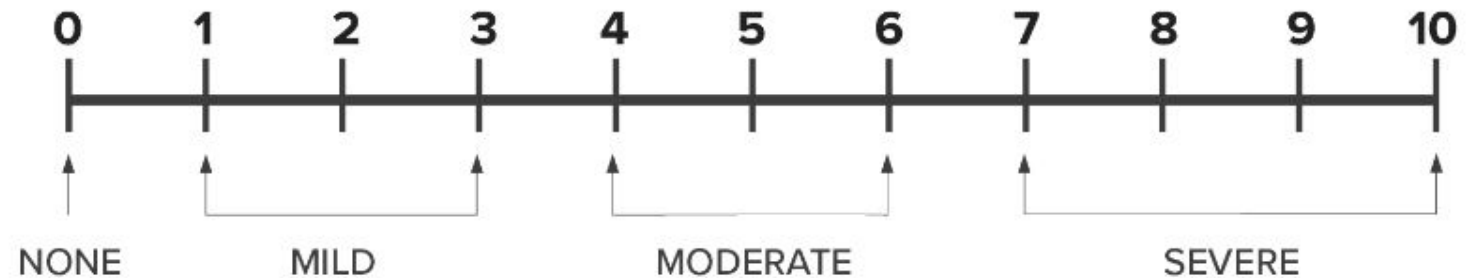
	A	B	C	D	E	F	G	H	I	J	K	L
1	TimeStamp	Time	Td	Tw	RH	Sdur		Sdur_accum_der	TSoil5	TSoil10	TSoil20	TSoil30
2	UTC	hhmm	degC	degC	%	sec		h	degC	degC	degC	degC
3	20180308	0005	4.78	4.14	93.5	0	0.00	0.00	5.64	5.77	5.84	5.87
4	20180308	0010	4.84	4.15	92.8	0	0.00	0.00	5.63	5.76	5.84	5.87
5	20180308	0015	4.83	4.08	92.0	0	0.00	0.00	5.63	5.75	5.83	5.87
6	20180308	0020	4.86	4.06	91.2	0	0.00	0.00	5.62	5.74	5.82	5.86
7	20180308	0025	4.72	3.95	90.8	0	0.00	0.00	5.62	5.73	5.81	5.87
8	20180308	0030	4.53	3.77	90.1	0	0.00	0.00	5.62	5.73	5.81	5.86
9	20180308	0035	4.52	3.74	90.1	0	0.00	0.00	5.62	5.72	5.80	5.86
10	20180308	0040	4.50	3.71	90.4	0	0.00	0.00	5.61	5.72	5.79	5.86
11	20180308	0045	4.51	3.71	90.4	0	0.00	0.00	5.61	5.71	5.79	5.86
12	20180308	0050	4.49	3.69	90.2	0	0.00	0.00	5.61	5.70	5.78	5.86
13	20180308	0055	4.49	3.67	90.0	0	0.00	0.00	5.62	5.70	5.77	5.85
14	20180308	0100	4.48	3.67	90.2	0	0.00	0.00	5.61	5.70	5.77	5.85

	A	B	C	D
1	TimeStamp	Sunrise	Sunset	DayLength
2	UTC	hhmm	hhmm	hrs
3	20180301	0816	1858	10.70
4	20180302	0814	1900	10.77
5	20180303	0811	1902	10.83
6	20180304	0809	1904	10.92
7	20180305	0806	1906	10.98
8	20180306	0804	1908	11.07
9	20180307	0801	1910	11.13
10	20180308	0759	1912	11.22
11	20180309	0756	1914	11.28
12	20180310	0754	1916	11.37
13	20180311	0751	1918	11.43
14	20180312	0749	1920	11.52
15	20180313	0746	1922	11.58
16	20180314	0744	1924	11.67
17	20180315	0741	1926	11.73

NIH Common Data Elements

A Common Data Element (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection.

0 - 10 Numeric Pain Rating Scale

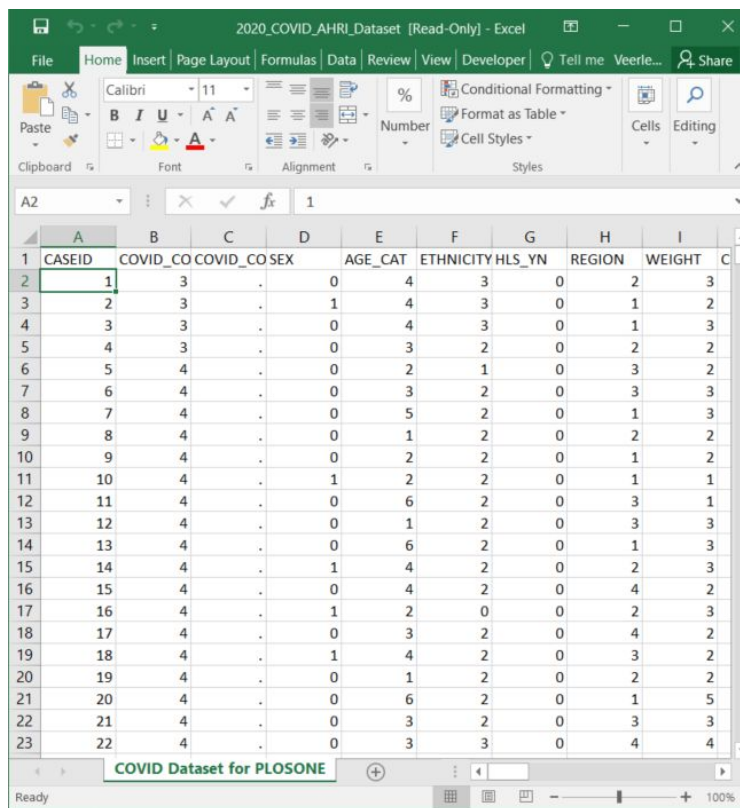


Categorical Scale



Codebook to understand data

Dataset



1	A	B	C	D	E	F	G	H	I	J
1	CASEID	COVID_CO	COVID_COSEX		AGE_CAT	ETHNICITY	HLS_YN	REGION	WEIGHT	
2	1	3	.	0	4	3	0	2	3	
3	2	3	.	1	4	3	0	1	2	
4	3	3	.	0	4	3	0	1	3	
5	4	3	.	0	3	2	0	2	2	
6	5	4	.	0	2	1	0	3	2	
7	6	4	.	0	3	2	0	3	3	
8	7	4	.	0	5	2	0	1	3	
9	8	4	.	0	1	2	0	2	2	
10	9	4	.	0	2	2	0	1	2	
11	10	4	.	1	2	2	0	1	1	
12	11	4	.	0	6	2	0	3	1	
13	12	4	.	0	1	2	0	3	3	
14	13	4	.	0	6	2	0	1	3	
15	14	4	.	1	4	2	0	2	3	
16	15	4	.	0	4	2	0	4	2	
17	16	4	.	1	2	0	0	2	3	
18	17	4	.	0	3	2	0	4	2	
19	18	4	.	1	4	2	0	3	2	
20	19	4	.	0	1	2	0	2	2	
21	20	4	.	0	6	2	0	1	5	
22	21	4	.	0	3	2	0	3	3	
23	22	4	.	0	3	3	0	4	4	

Codebook

1	VAR	Description	Response Options (if applicable)
2	CASEID	Case ID	
3	COVID_COHORTS_4	Cohorts (4): Pos, Neg, Untested, Not suspected	1:Positive; 2:Negative; 3:Untested; 4:COVID Not-Suspected
4	COVID_COHORTS_2	Cohort (2): Pos, Neg	1:Positive; 2:Negative
5	SEX	Sex	0:Female; 1:Male
6	AGE_CAT	Age categorized	1:18-24 years; 2:25-34 years; 3:35-44 years; 4:45-54 years; 5:55-64 years
7	ETHNICITY	Ethnicity	0:Other or more than 1; 1:Black or African American; 2:White; 3:Hispanic or Latino
8	HLS_YN	Hispanic, Latino, or Spanish origin	0:No; 1:Yes
9	REGION	Region	1:Northwest; 2:Midwest; 3:South; 4:West
10	WEIGHT	Weight categorized (self-reported)	1:Under weight; 2:Normal weight; 3:Slightly overweight; 4:Morbidly obese
11	CCI_SCORE	Charlson Comorbidity Score	
12	CCI_HBP	Hypertension	0:No; 1:Yes
13	CCI_DB_NO_COMPL	Diabetes (Type I or Type II) without complication	0:No; 1:Yes
14	CCI_COPD_BR_EMP	Chronic lung disease (COPD), chronic bronchitis	0:No; 1:Yes
15	CCI_HA	Heart attack	0:No; 1:Yes
16	CCI_CHF	Congestive heart failure	0:No; 1:Yes
17	CCI_STROKE_TIA	Stroke or transient ischemic attack (TIA)	0:No; 1:Yes
18	CCI_LV_MILD	Mild liver disease, hepatitis, cirrhosis	0:No; 1:Yes
19	CCI_DB_COMPL	Diabetes (Type I or Type II) with chronic complication	0:No; 1:Yes
20	CCI_PVD	Peripheral vascular disease	0:No; 1:Yes
21	CCI_LV_MOD_SEV	Moderate or severe liver disease, hepatitis, cirrhosis	0:No; 1:Yes
22	CARE_PHY_YN	Physician visit or call (telemedicine)	0:No; 1:Yes
23	CARE_ER_YN	Emergency room visit	0:No; 1:Yes
24	CARE_HOSP_YN	Hospital visit (stay of one or more nights)	0:No; 1:Yes
25	AB_RESULT	Antibody test result positive or negative	0:Negative; 1:Positive; 2:Not sure
26	SYM_FEVER	Fever	0:No; 1:Yes
27	SYM_COUGH	Dry cough	0:No; 1:Yes

Lego replication game

Lego replication game: discussion

- Structured templates help to write out instructions
 - Standardises the process
- Brick lists help to write out instructions
 - Reduces ambiguity
 - Standardises naming
 - Brick lists could have unique numbers / codes for each brick
 - = controlled vocabulary / community standard
- Visuals help: drawing or pictures of vehicle

Standardisation ...



6x2 brick



4x2 brick



3x2 brick



2x2 brick



4x1 brick



flag



4x2 brick with slope



2x2 brick with slope



3x2 brick with slope



tall 2x1 brick with slope

DETAILED INSTRUCTIONS		
Step	Parts required	Instructions
1		
2		

DETAILED INSTRUCTIONS			
Step	Part shape	Part colour	Instructions
1			
2			
3			



Take away messages

- When you start a project, design your folder structure and file naming system
- When you end your project / publish your paper, check your folder structure / file names are still in order (or fix), then zip and archive your data
- Use open / standard file formats when you can to make your data FAIR
- Use data standards where you can, to make your data interoperable and FAIR

