KU LEUVEN

Organising & standardising research data that underpin your publication

Author: Veerle Van den Eynden, KU Leuven RDM Competence Centre

Trainer: Bruna Piereck

VIB RDM training, Gent

October 2024

Overview

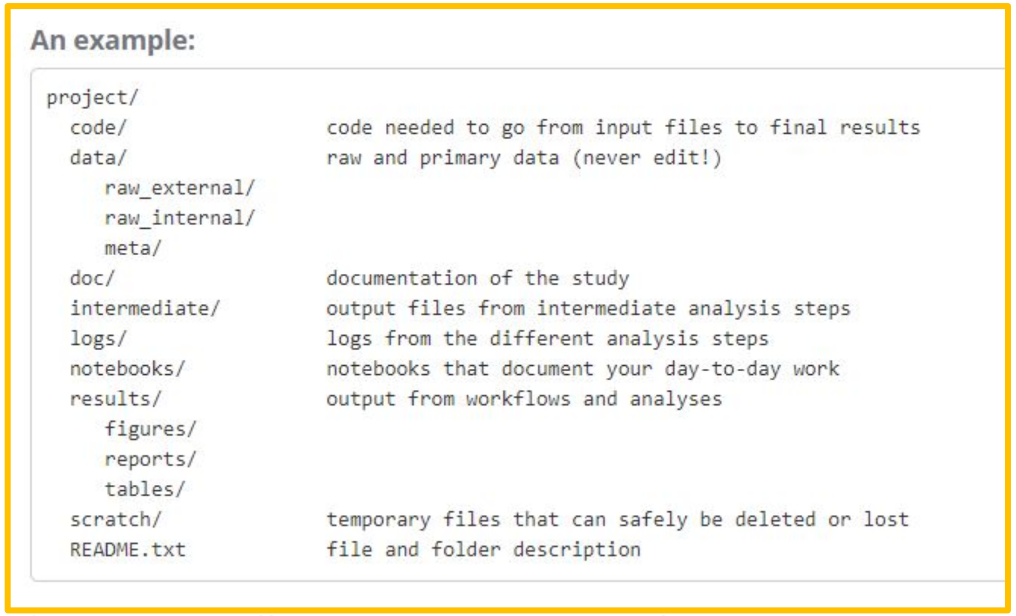
- *Organise and structure data and documentation files*
- *Logical, structured and descriptive file names*
- *Open / standard file formats*
- *File versioning in a project*
- *Data standards*
  - *make data interoperable and reusable*
  - *commonly understandable*

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
- Good practice: ReadMe file in top folder
- Consider read / write access to folders for colleagues / collaborators
- 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 Research Coordination Office at KULeuven, each project has a designated folder. When a new project is started, a new folder is made. There raw data, syntax files, questionnaires, ethical approval, etc are kept. All researchers have access to the shared drive and to all folders.

Record file

- Record file
  - A textual or tabular
  - List all data and documentation files of a project, paper, etc.
  - Specifies 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

In the research group they mostly develop algorithms for simulations. Every researcher has to keep a register (Word file) that lists which code repositories (on GitLab or GitHub) are used with the URL, and where data files are stored.

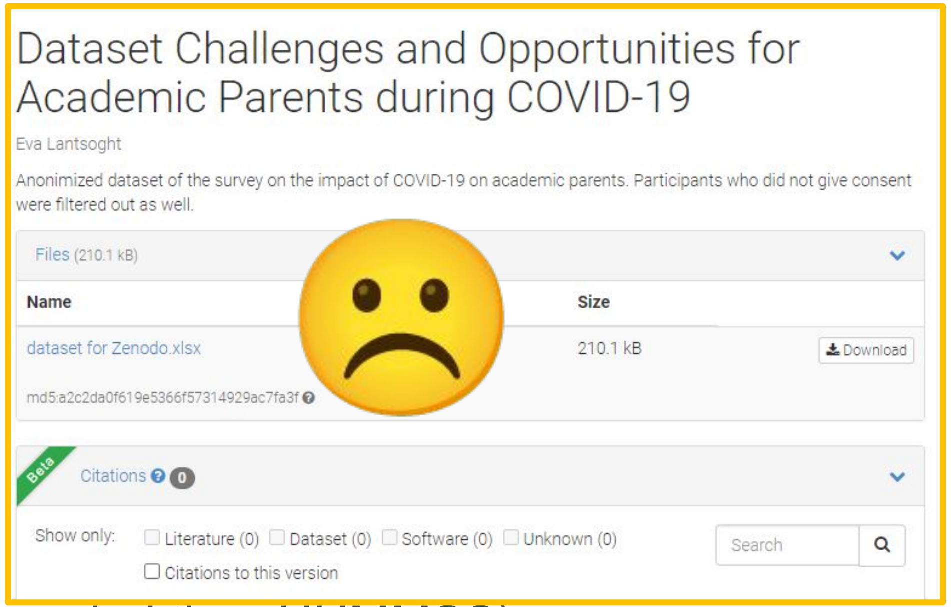
These registers are available with read access for colleagues.

eLab Notebook

Data management plan

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 md5:a2c2da0f619e5366f57314929ac7fa3f	210.1 kB	Download

Citations 0

Show only: □ Literature (0) □ Dataset (0) □ Software (0) □ Unknown (0)

□ 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

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

File naming examples

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

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

File naming examples

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

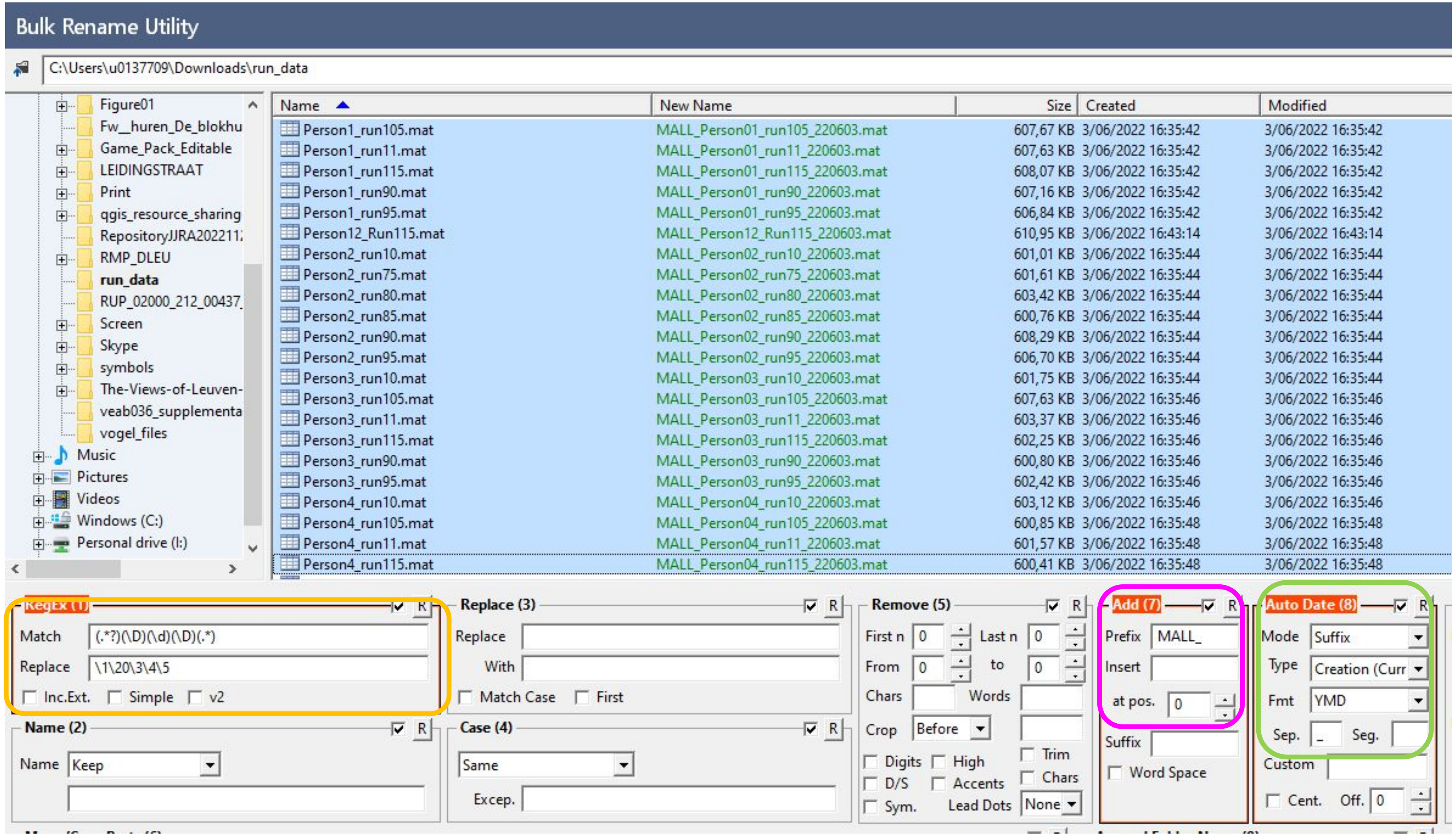
Batch file renaming

- 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



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 (.*?)(\D)(\d)(\D)(.*)
Replace \1\20\3\4\5
□ Inc.Ext. □ Simple □ v2

Name (2) ☑ R
Name Keep

Replace (3) ☑ R
Replace
With
□ Match Case □ First

Case (4) ☑ R
Same
Excep.

Remove (5) ☑ R
First n 0 Last n 0
From 0 to 0
Chars Words
Crop Before
□ Digits □ High □ Trim
□ D/S □ Accents □ Chars
□ Sym. Lead Dots None

Add (7) ☑ R
Prefix MALL_
Insert
at pos. 0
Suffix
□ Word Space

Auto Date (8) ☑ R
Mode Suffix
Type Creation (Curr
Fmt YMD
Sep. _ Seg.
Custom
□ Cent. Off. 0

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

- DONOR, COPD, IPF + number
- Whole, Part
- Mild, moderate, severe
- Lung
- Sample number

Exercise: open / standard file formats

- List all file formats you use in your daily work / research. (group activity)
- After a few minutes you will share with the group
  - Open formats
  - Proprietary formats (standard / not)

Open / standard file formats

- Use oppen/standard file formats
  - Long term access
  - Use of research data
- Good source: fairsharing.org
  - examples:

- Containers: TAR, ZIP
- Databases: XML, CSV, JSON
- Video: MPEG (mp4), AVI
- Sounds: WAVE, AIFF, MP3, FLAC
- Statistics: DTA, POR, SAS, SAV
- Images: TIFF, JPEG 2000, PNG, GIF
- Tabular data: CSV, TXT
- Text: XML, PDF/A, HTML, JSON, TXT, RTF
- 3D: X3D, C3D
- Neuroimaging: DICOM, Nifti
- Mass spectrometry: mzML
- Sequencing data: FASTA, FASTQ
- Microscopy: OME Next Generation File Format, Bio-formats conversions

File versioning

- Manage multiple versions
- Enable reverting to an earlier version
- Easy methods for small demands of versioning:
  - File naming
  - Cloud storage file versioning, e.g. OneDrive

- For automatic management of versioning
  - conflict resolution
  - back-tracing capabilities,
  - proper version control

- Git
- GitHub
- GitLab
- BitBucket

Data standards

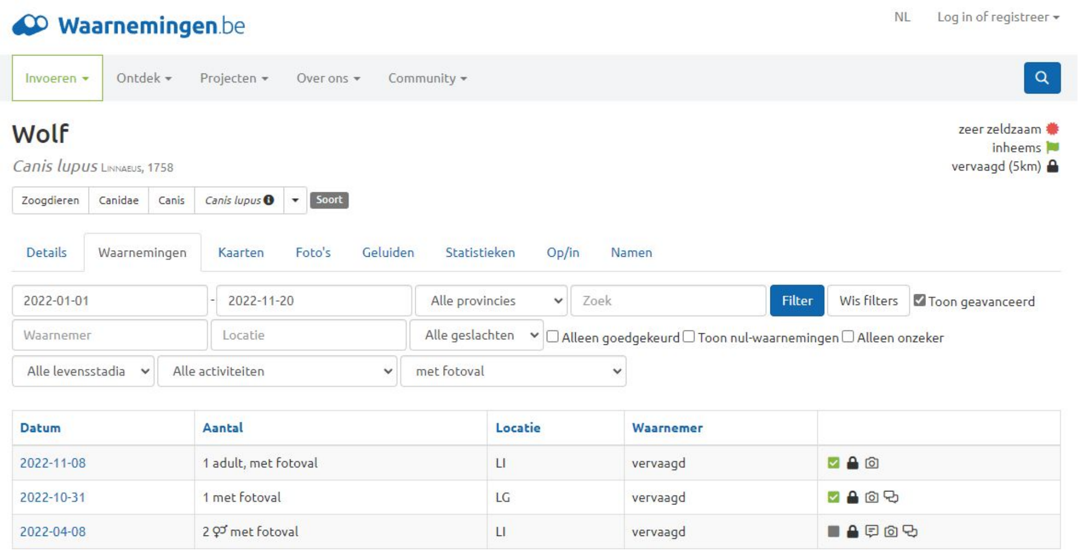
- Make data interoperable
- Easier to understand
  - by multiple communities
- Reusable more widely
  - International, common standards
  - Community standards

Question

- Which standards do you already use in your research?

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

zeer zeldzaam
inheems
vervaagd (5km)

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

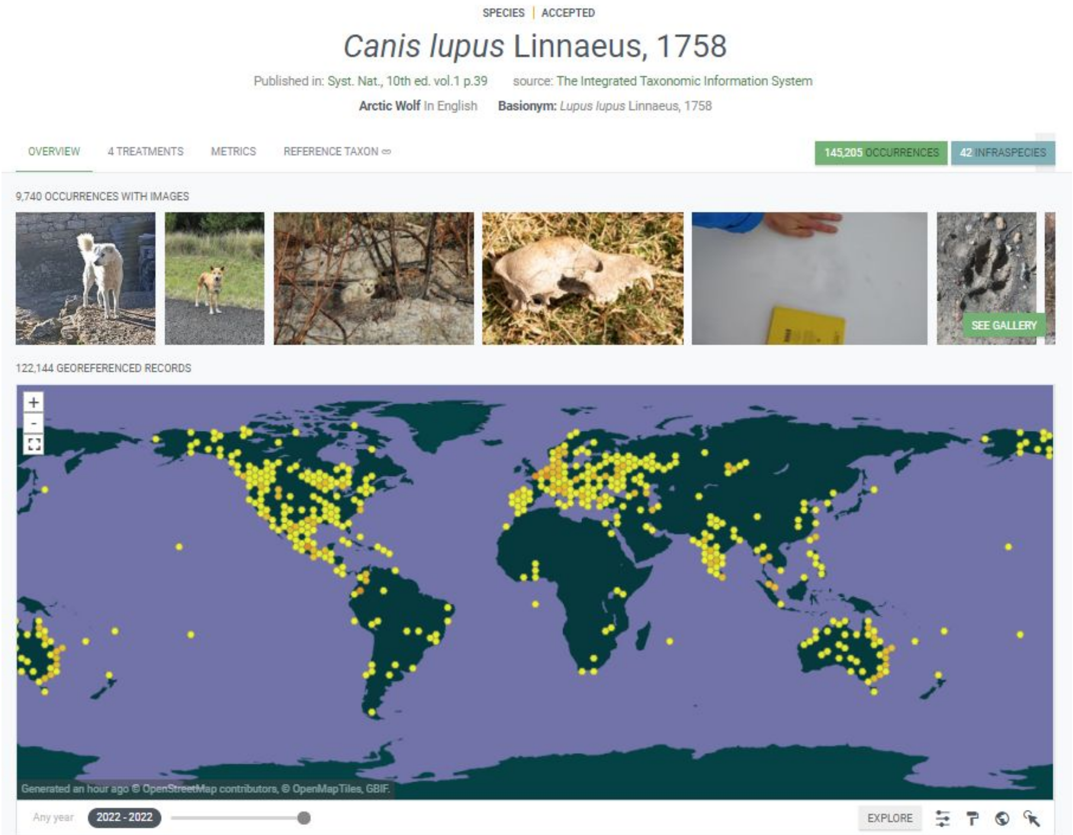
Waarnemer   Locatie   Alle geslachten   □ Alleen goedgekeurd □ Toon nul-waarnemingen □ Alleen onzeker

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	■ 🔒 📷

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.

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

SEE GALLERY

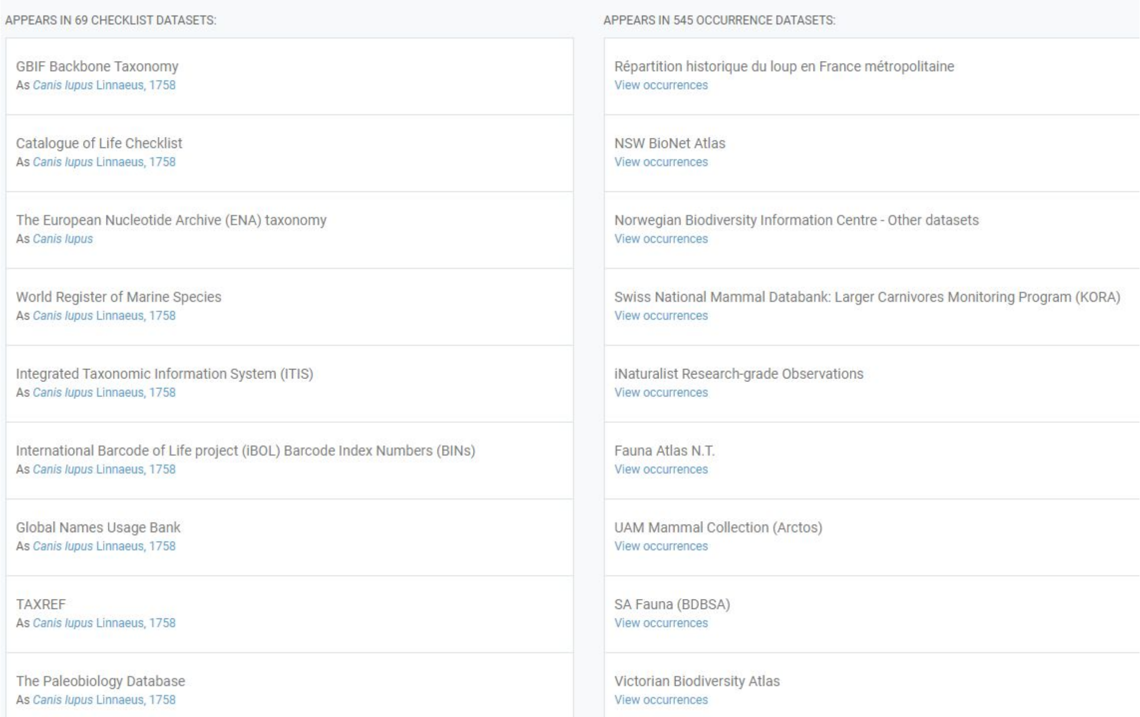
122,144 GEOREFERENCED RECORDS

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

Any year   2022 - 2022   EXPLORE

GBIF & Darwin Core

Appears in Datasets



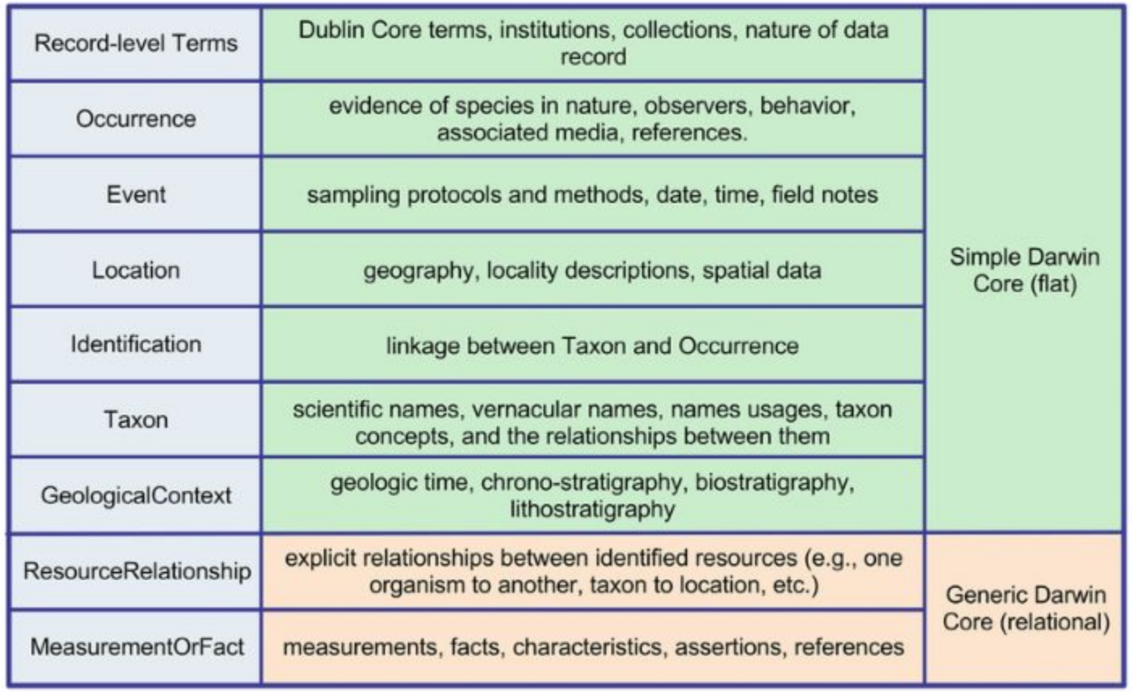
APPEARS IN 69 CHECKLIST DATASETS:

- GBIF Backbone Taxonomy
As *Canis lupus* Linnaeus, 1758
- Catalogue of Life Checklist
As *Canis lupus* Linnaeus, 1758
- The European Nucleotide Archive (ENA) taxonomy
As *Canis lupus*
- World Register of Marine Species
As *Canis lupus* Linnaeus, 1758
- Integrated Taxonomic Information System (ITIS)
As *Canis lupus* Linnaeus, 1758
- International Barcode of Life project (iBOL) Barcode Index Numbers (BINs)
As *Canis lupus* Linnaeus, 1758
- Global Names Usage Bank
As *Canis lupus* Linnaeus, 1758
- TAXREF
As *Canis lupus* Linnaeus, 1758
- The Paleobiology Database
As *Canis lupus* Linnaeus, 1758

APPEARS IN 545 OCCURRENCE DATASETS:

- Répartition historique du loup en France métropolitaine
View occurrences
- NSW BioNet Atlas
View occurrences
- Norwegian Biodiversity Information Centre - Other datasets
View occurrences
- Swiss National Mammal Databank: Larger Carnivores Monitoring Program (KORA)
View occurrences
- iNaturalist Research-grade Observations
View occurrences
- Fauna Atlas N.T.
View occurrences
- UAM Mammal Collection (Arctos)
View occurrences
- SA Fauna (BDBSA)
View occurrences
- 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	

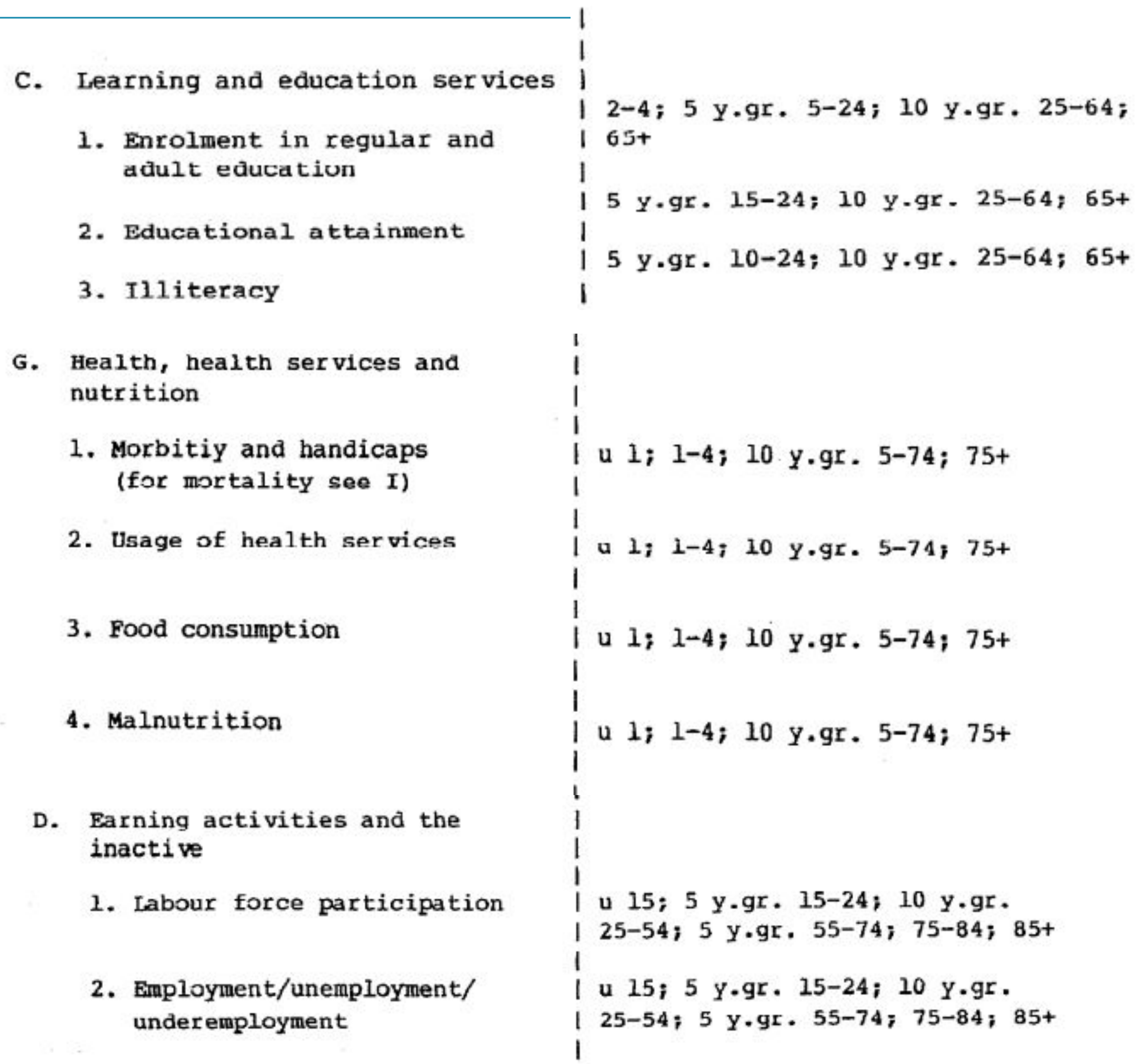
Standards

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. Morbitiy 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 !

Incompatible dates

☹

Terrorism attacks on buildings

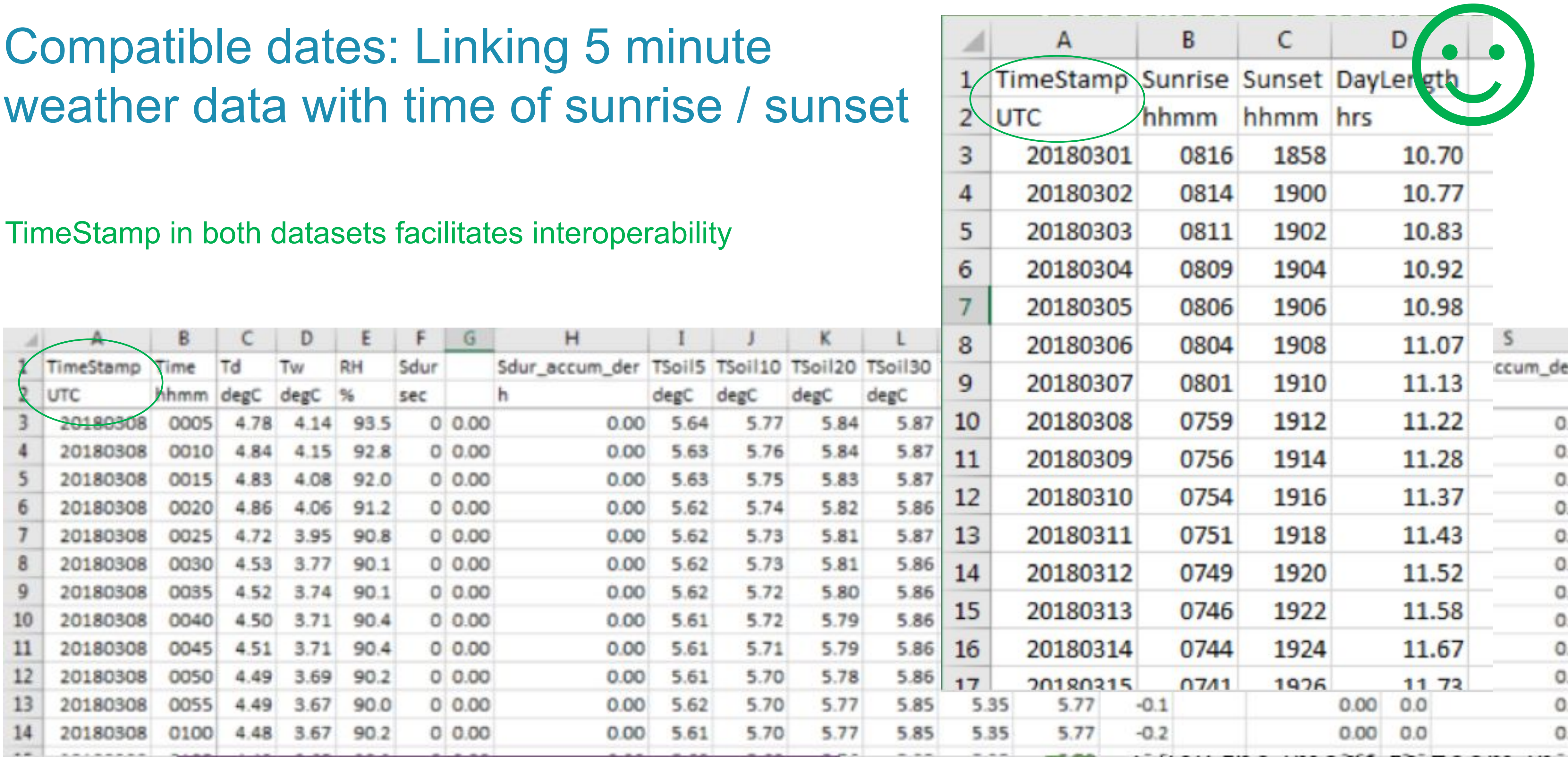
Date	Country	Target	Place
09/11/2001	USA	WTC	New York
13/11/2015	France	Bataclan	Paris
12/10/1984	UK	Grand Hotel	Brighton

Tweets

Tweet ID	Date-Time	Tweet text
320217690004393984	2001-09-11T16:53:41Z	Lorem ipsum dolor sit amet
320206007982755840	20001-09-11T16:07:16Z	Lorem ipsum dolor sit amet
320205389780090880	2001-09-11T16:04:48Z	Lorem ipsum dolor sit amet
320202492031930368	2001-09-11T15:53:17Z	Lorem ipsum dolor sit amet
320197516371062784	2001-09-11T15:33:31Z	Lorem ipsum dolor sit amet
320197511107211265	2015-11-13T15:33:30Z	Lorem ipsum dolor sit amet
320195708835749889	2015-11-13T15:26:20Z	Lorem ipsum dolor sit ame
320193833260425216	2015-11-13T15:18:53Z	tLorem ipsum dolor sit amet

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

TimeStamp in both datasets facilitates interoperability

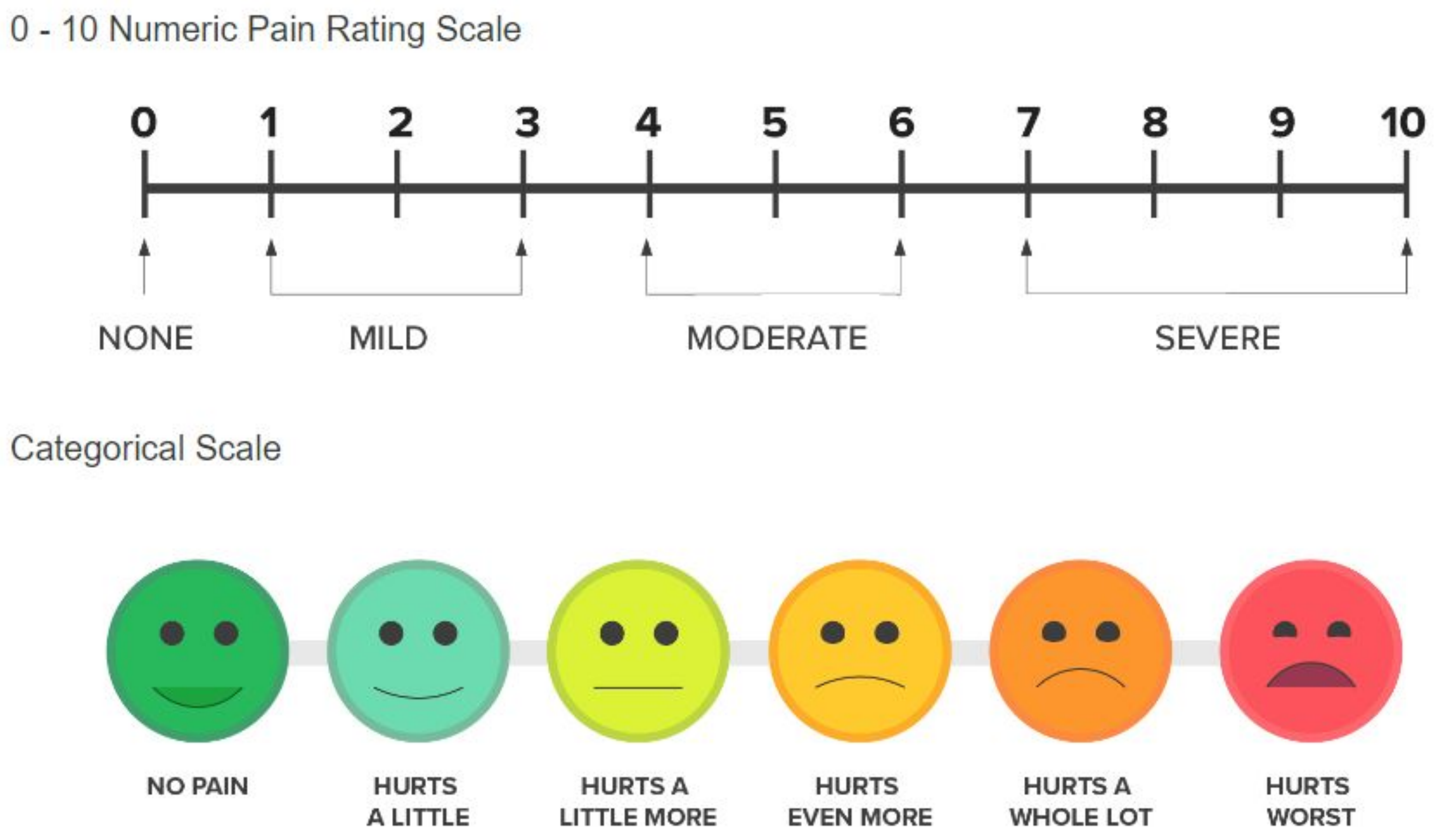


	A	B	C	D
1	TimeStamp	Sunrise	Sunset	DayLength
2	UTC	hhmm	hhmm	hrs
3	20180301	0816	1858	10.70
4	20180302	0814	1900	10.77
5	20180303	0811	1902	10.83
6	20180304	0809	1904	10.92
7	20180305	0806	1906	10.98
8	20180306	0804	1908	11.07
9	20180307	0801	1910	11.13
10	20180308	0759	1912	11.22
11	20180309	0756	1914	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

	A	B	C	D	E	F	G	H	I	J	K	L							S
1	TimeStamp	Time	Td	Tw	RH	Sdur		Sdur_accum_der	TSoil5	TSoil10	TSoil20	TSoil30							ccum_de
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							0
4	20180308	0010	4.84	4.15	92.8	0	0.00	0.00	5.63	5.76	5.84	5.87							0
5	20180308	0015	4.83	4.08	92.0	0	0.00	0.00	5.63	5.75	5.83	5.87							0
6	20180308	0020	4.86	4.06	91.2	0	0.00	0.00	5.62	5.74	5.82	5.86							0
7	20180308	0025	4.72	3.95	90.8	0	0.00	0.00	5.62	5.73	5.81	5.87							0
8	20180308	0030	4.53	3.77	90.1	0	0.00	0.00	5.62	5.73	5.81	5.86							0
9	20180308	0035	4.52	3.74	90.1	0	0.00	0.00	5.62	5.72	5.80	5.86							0
10	20180308	0040	4.50	3.71	90.4	0	0.00	0.00	5.61	5.72	5.79	5.86							0
11	20180308	0045	4.51	3.71	90.4	0	0.00	0.00	5.61	5.71	5.79	5.86							0
12	20180308	0050	4.49	3.69	90.2	0	0.00	0.00	5.61	5.70	5.78	5.86							0
13	20180308	0055	4.49	3.67	90.0	0	0.00	0.00	5.62	5.70	5.77	5.85	5.35	5.77	-0.1		0.00	0.0	0
14	20180308	0100	4.48	3.67	90.2	0	0.00	0.00	5.61	5.70	5.77	5.85	5.35	5.77	-0.2		0.00	0.0	0

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.



Quiz data standards

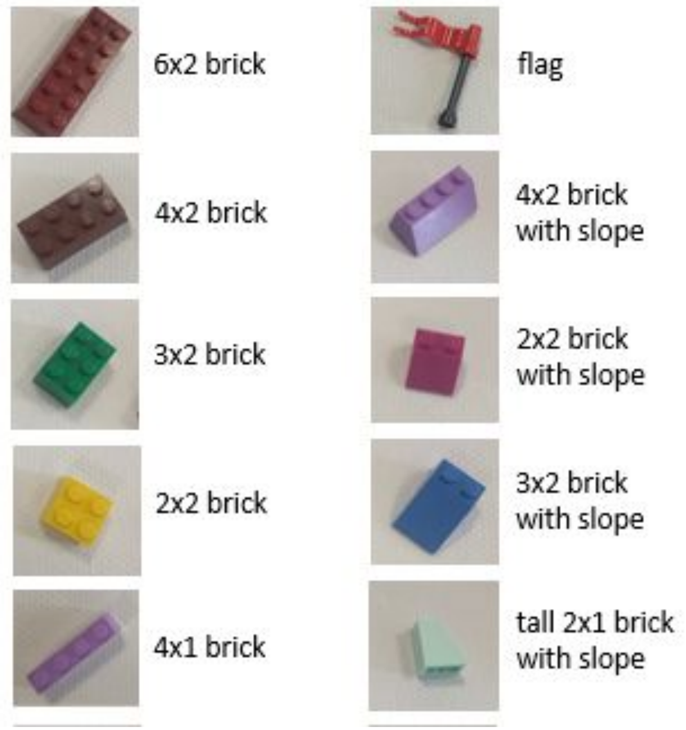
https://www.mentimeter.com/app/presentation/n/alz4ewsydnoq2jeb14afvgro7c4nrk4b/present?question=gd2bx18qsv7t

Lego replication game

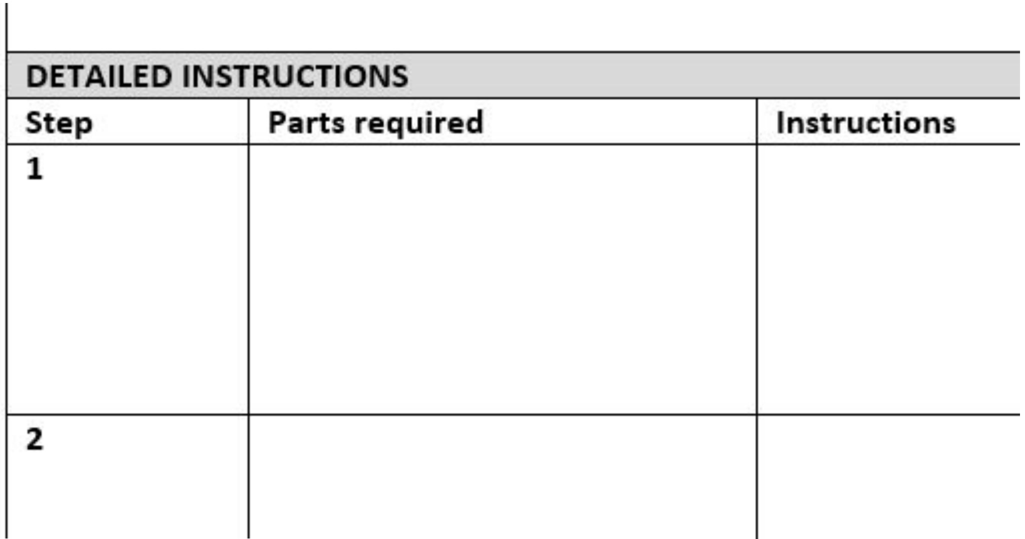
Lego replication game: discussion

- 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 vehicel

Standardisation …

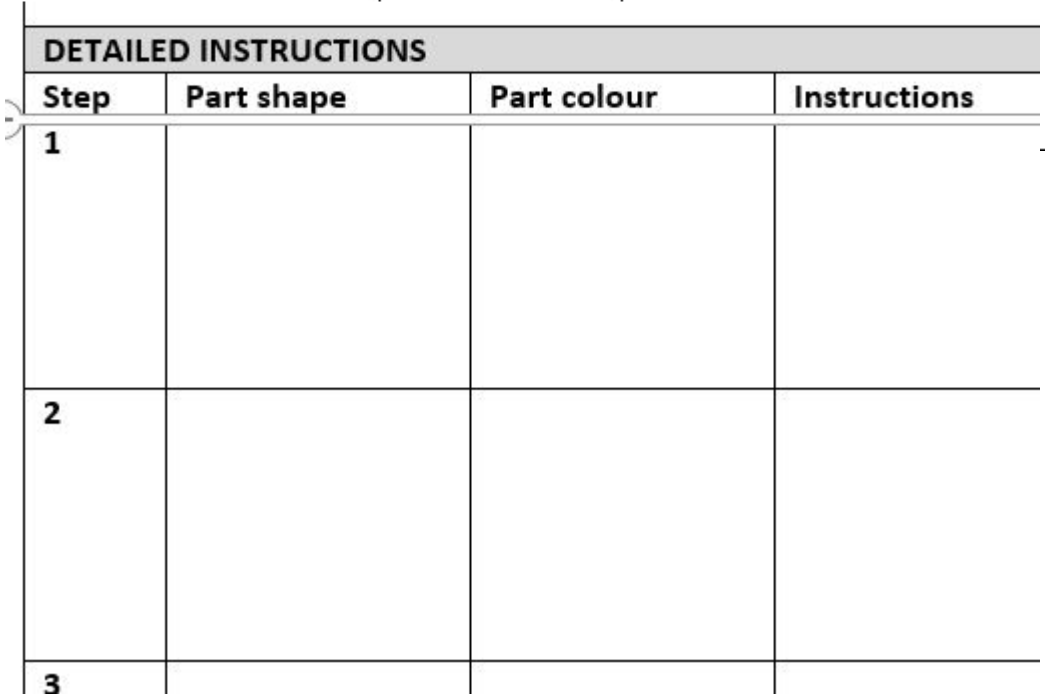


	6x2 brick		flag
	4x2 brick		4x2 brick with slope
	3x2 brick		2x2 brick with slope
	2x2 brick		3x2 brick with slope
	4x1 brick		tall 2x1 brick with slope



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