KU LEUVEN

Organising & standardising research data that underpin your publication

Willeke de Haan
KU Leuven Research Coordination Office

VIB RDM training
23 May 2025

Agenda

- Organise and structure data and documentation files
  - Logical, structured and descriptive file names
- Open and standard file formats
- File versioning
- Data standards
  - Make data Interoperable, understandable and Reusable (I and R of FAIR)

Organise / structure files and folders

- Find what works for you, in line with your data collection methods, active data storage system and technical knowledge/support available
- In this presentation I show some examples of good practices that researchers at KU Leuven use

Folder structure

- 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 or write access to folders for colleagues and collaborators
- When paper is published (or end of project): Check folders and content, package structure and files into zip bundle and move to archival storage

Folder structure

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
```

- 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 or write access to folders for colleagues and collaborators
- When paper is published (or end of project): Check folders and content, package structure and files into zip bundle and move to archival storage

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**: sandbow 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

Example

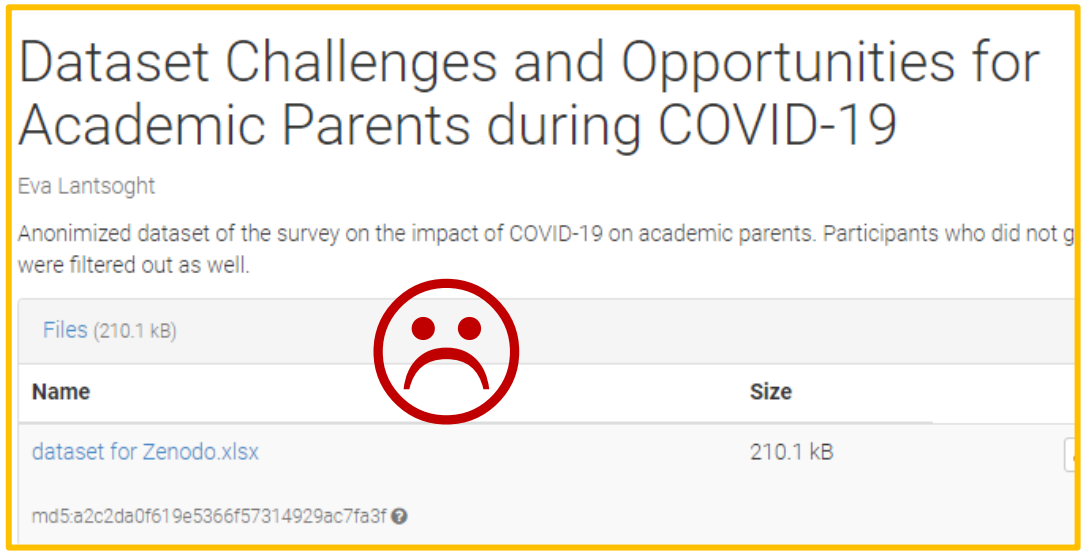
ProjectName

- **General_Information**
  - Grant_Information
  - Reports_Deliverables
  - Manuscripts
- **Shared_Protocols**
  - Protocols
  - Code
  - Sample_Inventory
  - Shared_Datasets
- **Researcher_Name**
  - Project plan
  - Documents
  - Data
    - Work_package#
    - Experiment#
    - Temp
  - Thesis
  - Presentations

File naming

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

File naming



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

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_V03.csv
- Date_ProjectAbbreviation_ExperimentNumber_Location_Version

Cropped image of an ant head taken on the third of December 2020 by Mia Moore

- File name: 20201203_MM_HEAD_CROPPED_V1.jpg
- Date_CreatorData_Type_Modification_Version

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

- BDHS_SurveyProcedures_00-04.pdf
- ProjectAcronym_Type_Version

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

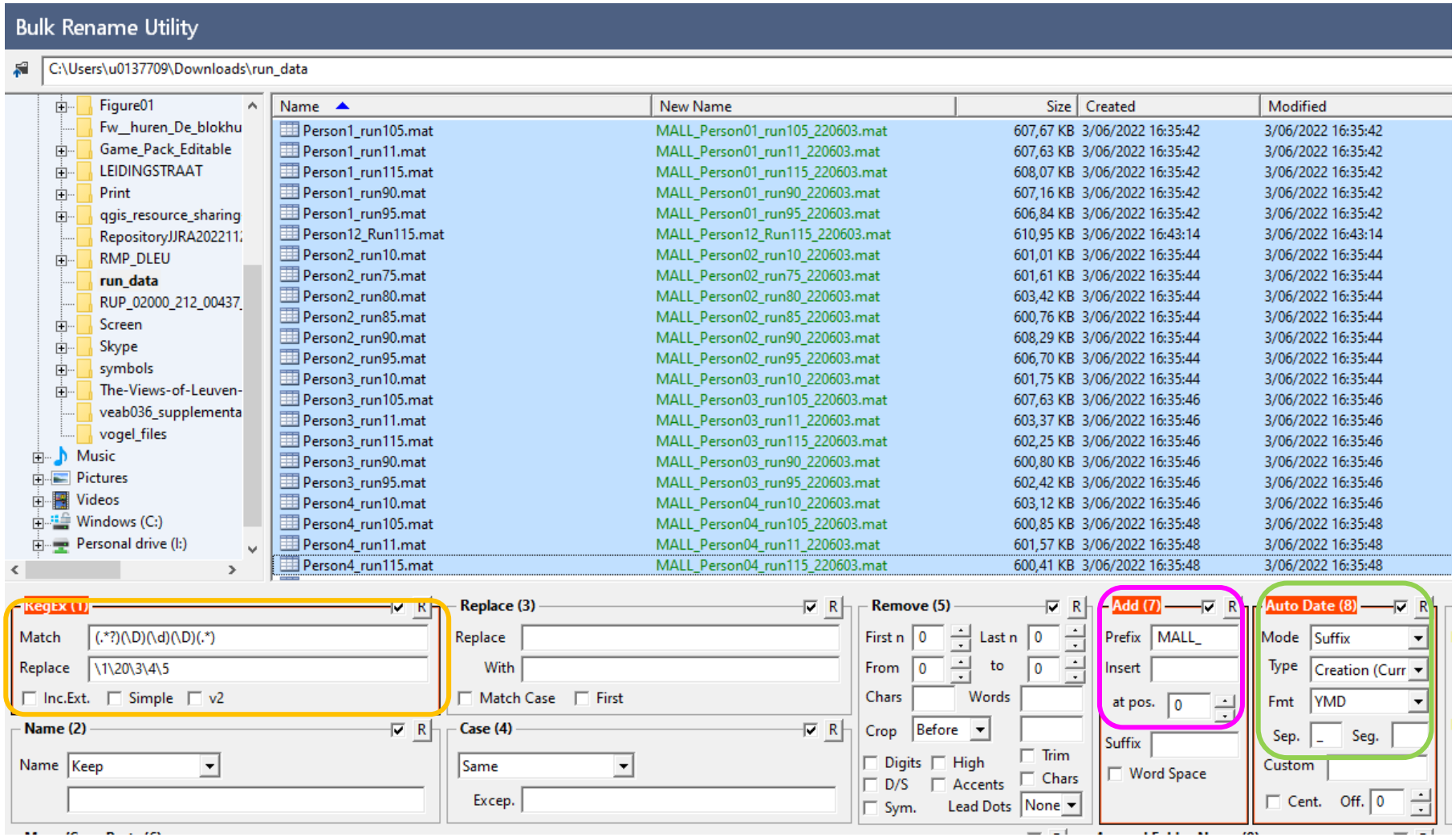
Batch file renaming

- 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



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

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 (or sample):
  - Unique ID
  - Dataset/Sample name
  - Description
  - Origin
  - Owner
  - Person responsible
  - Purpose, e.g. project name
  - Storage location
  - Contains personal data Y/N
  - Size / volume
  - Access: who has / needs access to the data

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 (or sample):
  - Unique ID
  - Dataset/Sample name
  - Description
  - Origin
  - Owner
  - Person responsible
  - Purpose, e.g. project name
  - Storage location
  - Contains personal data Y/N
  - Size / volume
  - Access: who has / needs access to the data

In our group we work with blood and tissues from genetically modified mice. For each type of assay we have an Excel template where the first sheet contains all the documentation about the experiment. For qPCR for example this sheet is based on the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines. We use prefilled fields and dropdown lists where possible and we have added some fields that are not part of the standards but are useful for us to understand each other's' data, e.g. the location of the samples

EXPERIMENTAL DESIGN			
Definition of experimental and control groups		E	
Number within each group		E	
Assay carried out by core lab or investigator's lab?		D	
Acknowledgement of authors' contributions		D	
SAMPLE			
Description		E	
Volume/mass of sample processed		D	
Microdissection or macrodissection		E	
Processing procedure		E	
If frozen - how and how quickly?		E	
If fixed - with what, how quickly?		E	
Sample storage conditions and duration (especially for FFPE samples)		E	
NUCLEIC ACID EXTRACTION			
Procedure and/or instrumentation		E	
Name of kit and details of any modifications		E	
Source of additional reagents used		D	
Details of DNase or RNAse treatment		E	
Contamination assessment (DNA or RNA)		E	
Nucleic acid quantification		E	
Instrument and method		E	
Purity (A260/A280)		D	
Yield		D	
RNA integrity method/instrument		E	
RIN/RQI or Cq of 3' and 5' transcripts		E	
Electrophoresis traces		D	
Inhibition testing (Cq dilutions, spike or other)		E	
REVERSE TRANSCRIPTION			
Complete reaction conditions		E	
Amount of RNA and reaction volume		E	
Priming oligonucleotide (if using GSP) and concentration		E	
Reverse transcriptase and concentration		E	
Temperature and time		E	

eLab Notebook

- eLab notebook systems (e.g. eLabJournal, 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.

Exercise: folder structure & file naming

- The role of basial 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
- Documents
  - SOP
  - …
  - Papers
- Code
- Results

Elements

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

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. SharePoint

- For automatic management of versioning, conflict resolution and back-tracing capabilities, use a proper version control software, e.g. GitHub, GitLab, BitBucket

Open / standard file formats

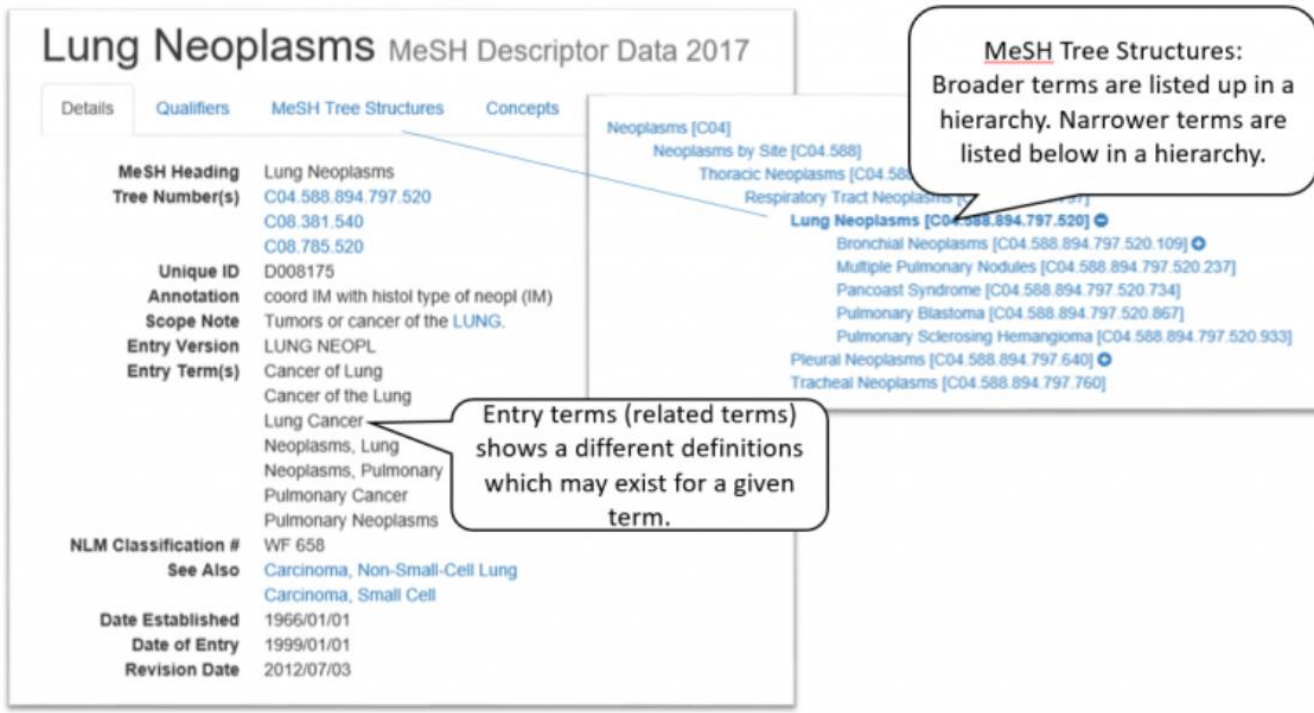
- Use or convert to open / standard file formats when you can
- For long-term access and reuse of research data
  - Images: TIFF, JPEG 2000, PNG, GIF
  - Text: XML, PDF/A, HTML, JSON, TXT, RTF
  - Containers: TAR, ZIP
  - Databases: XML, CSV, JSON
  - Video: MPEG (mp4), AVI
  - Sounds: WAVE, AIFF, MP3,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

Data standards

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

Data standards

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



Standards

International

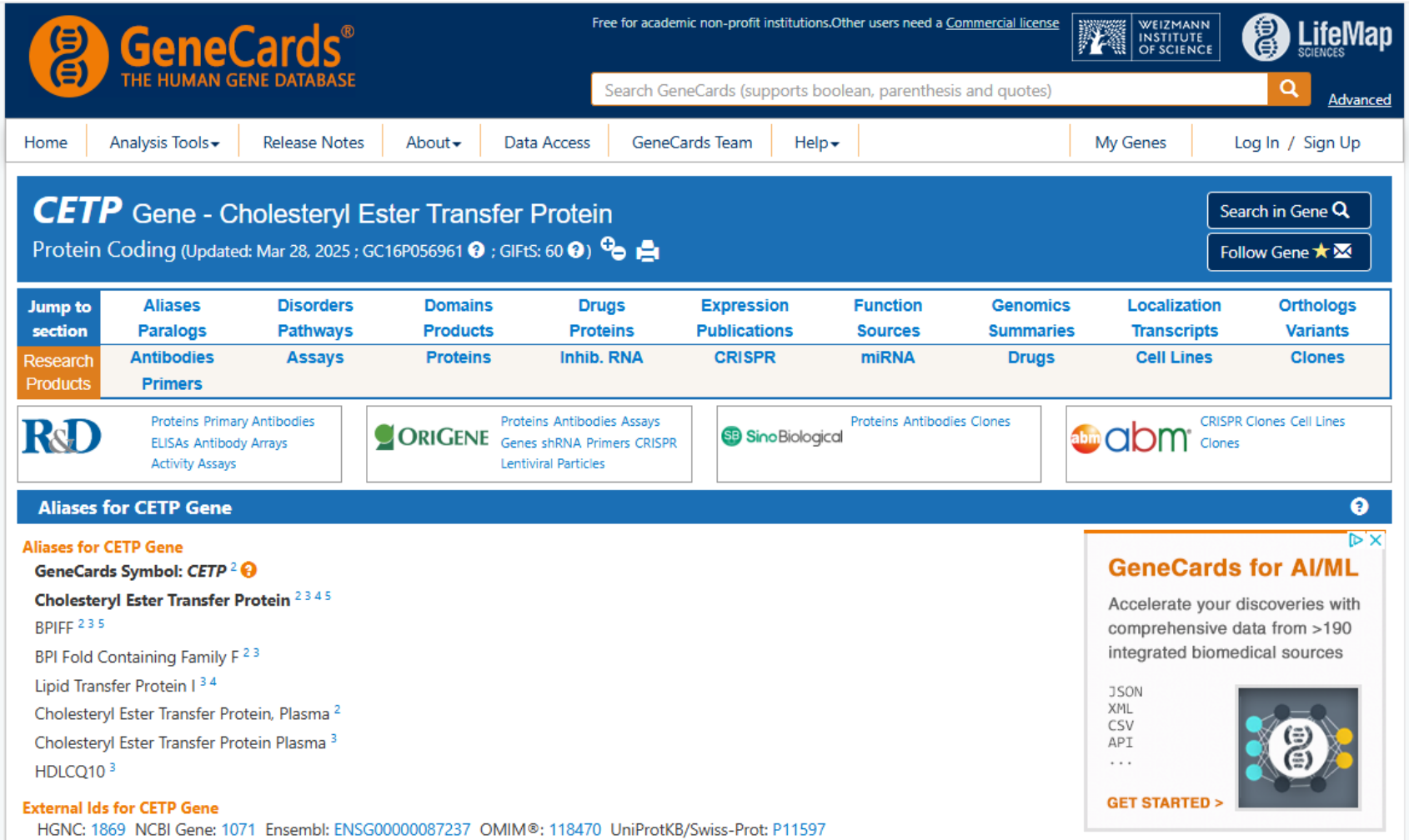
- ISO 8601 standards for date / time
- ISO 3166 standard for country codes

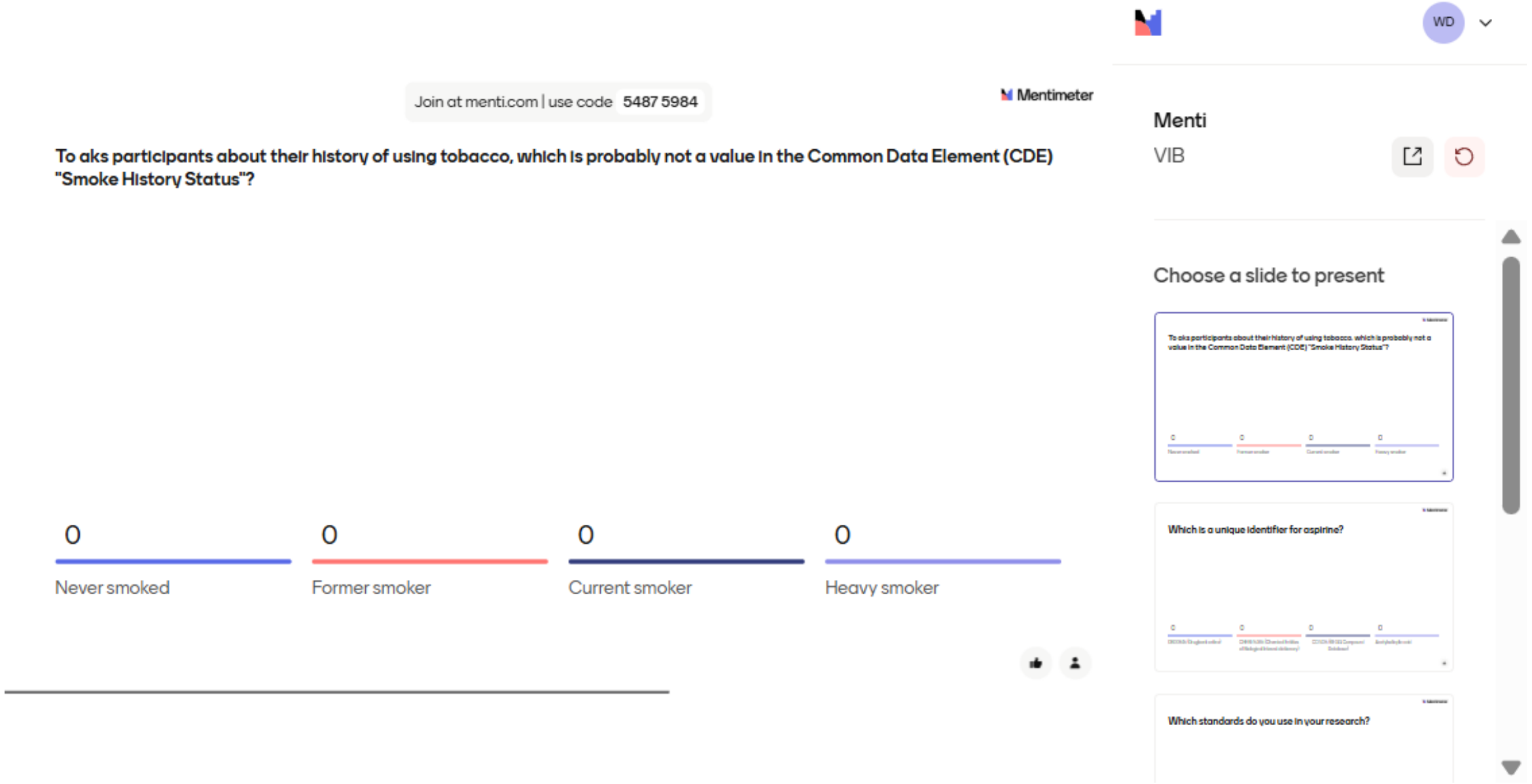
Community

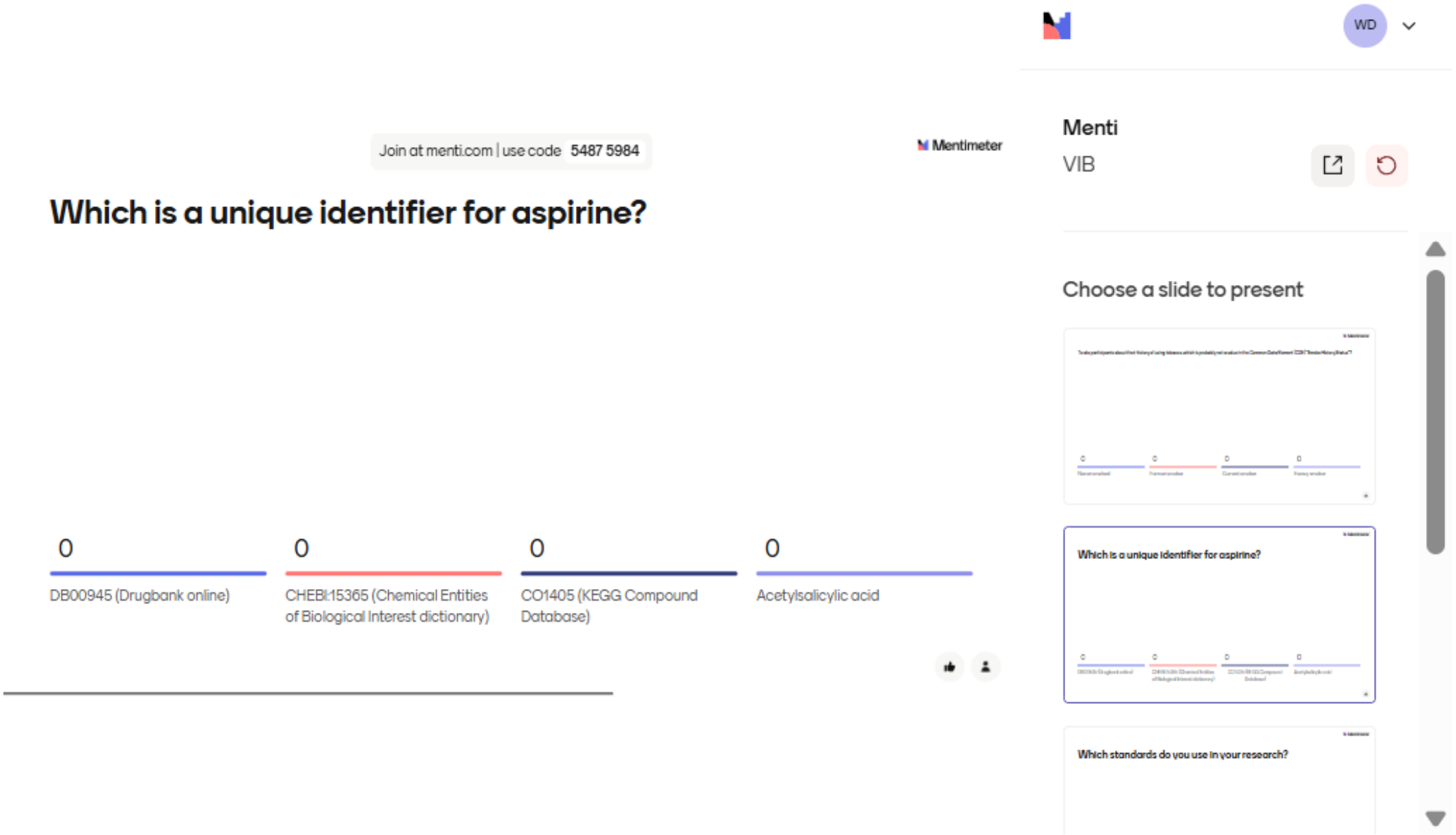
- DICOM MRI data
- Standard International Age Classification, UNStat 1982
- OME for microscopy
- MINSEQE for sequencing

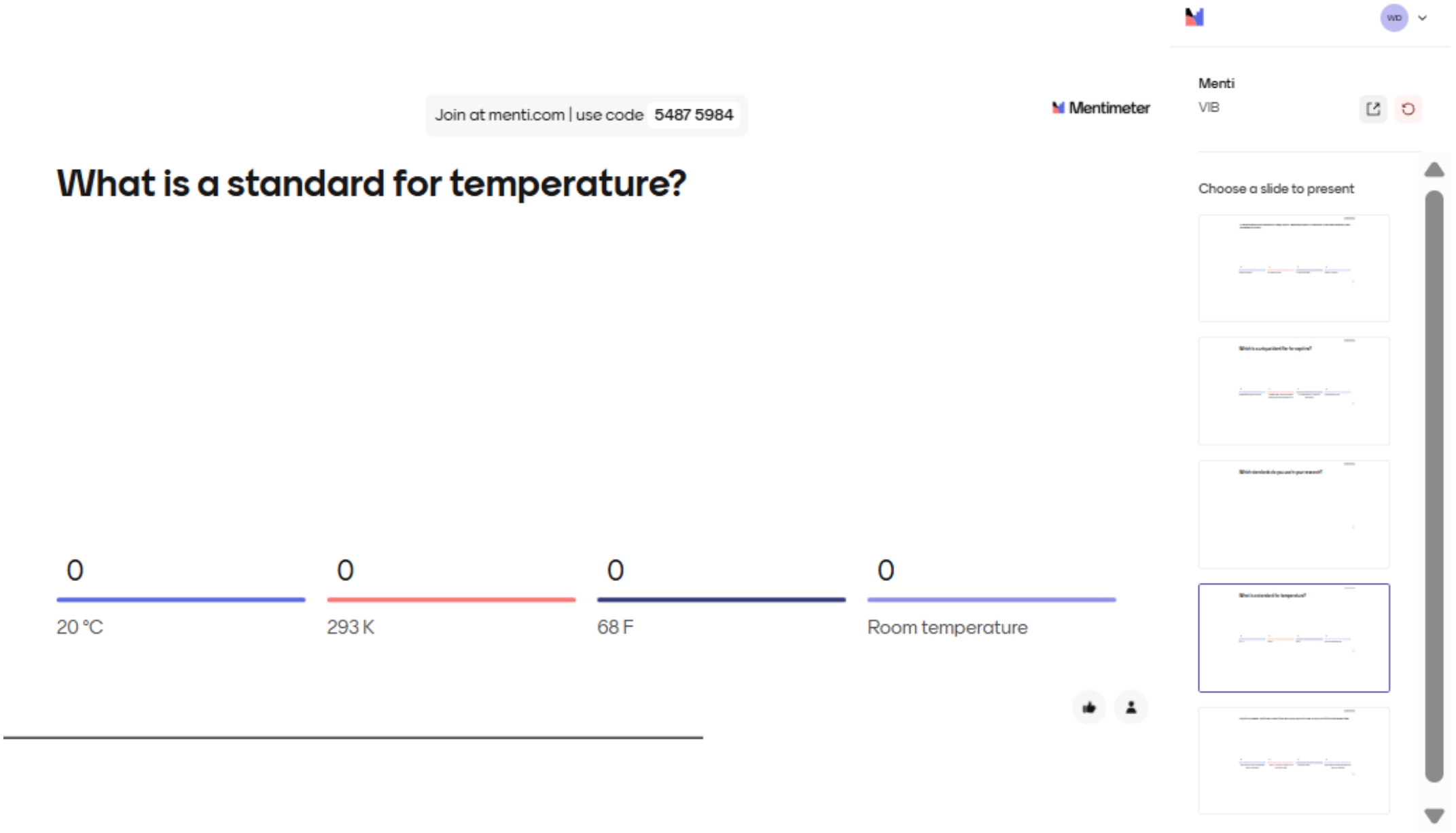
C. Learning and education services	
1. Enrolment in regular and adult education	2-4; 5 y.gr. 5-24; 10 y.gr. 25-64; 65+
2. Educational attainment	5 y.gr. 15-24; 10 y.gr. 25-64; 65+
3. Illiteracy	5 y.gr. 10-24; 10 y.gr. 25-64; 65+
G. Health, health services and nutrition	
1. Morbitiy and handicaps (for 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+

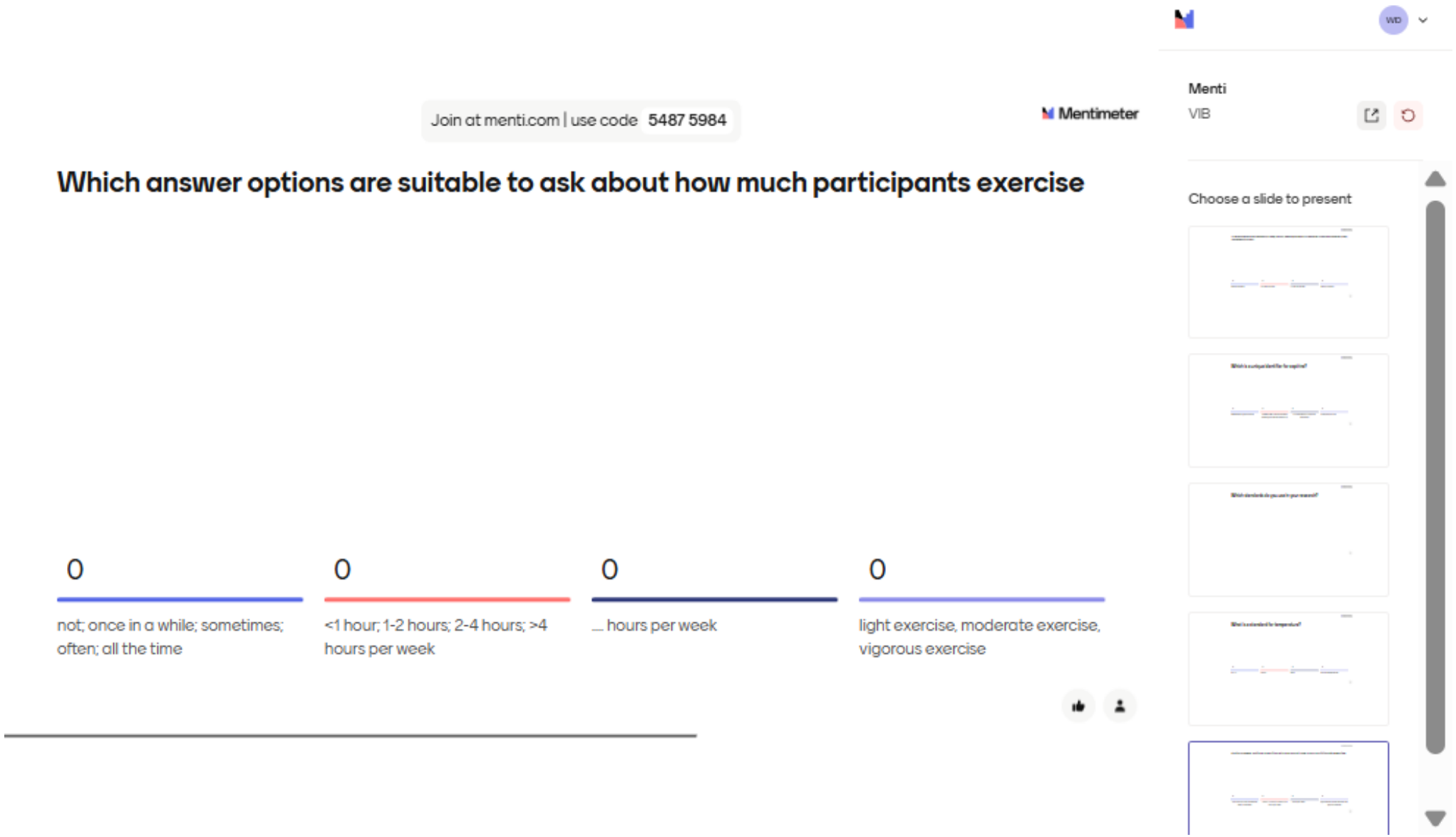
When age classification categories are applied consistently at an international level, datasets can be easily linked, combined and compared. But: different disciplines / purposes will need different categories !


GeneCards®
THE HUMAN GENE DATABASE
Free for academic non-profit institutions.Other users need a Commercial license
WEIZMANN INSTITUTE OF SCIENCE
LifeMap SCIENCES
Search GeneCards (supports boolean, parenthesis and quotes)
Advanced
Home
Analysis Tools
Release Notes
About
Data Access
GeneCards Team
Help
My Genes
Log In / Sign Up
CETP Gene - Cholesteryl Ester Transfer Protein
Protein Coding (Updated: Mar 28, 2025 ; GC16P056961 ; GIFtS: 60)
Search in Gene
Follow Gene
Jump to section
Aliases
Paralogs
Disorders
Pathways
Domains
Products
Drugs
Proteins
Expression
Publications
Function
Sources
Genomics
Summaries
Localization
Transcripts
Orthologs
Variants
Research Products
Antibodies
Primers
Assays
Proteins
Inhib. RNA
CRISPR
miRNA
Drugs
Cell Lines
Clones
R&D
Proteins Primary Antibodies
ELISAs Antibody Arrays
Activity Assays
ORIGENE
Proteins Antibodies Assays
Genes shRNA Primers CRISPR
Lentiviral Particles
SinoBiological
Proteins Antibodies Clones
abm
CRISPR Clones Cell Lines
Clones
Aliases for CETP Gene
Aliases for CETP Gene
GeneCards Symbol: CETP
Cholesteryl Ester Transfer Protein
BPIFF
BPI Fold Containing Family F
Lipid Transfer Protein I
Cholesteryl Ester Transfer Protein, Plasma
Cholesteryl Ester Transfer Protein Plasma
HDLCQ10
GeneCards for AI/ML
Accelerate your discoveries with comprehensive data from >190 integrated biomedical sources
JSON
XML
CSV
API
...
GET STARTED >
External Ids for CETP Gene
HGNC: 1869 NCBI Gene: 1071 Ensembl: ENSG00000087237 OMIM®: 118470 UniProtKB/Swiss-Prot: P11597
Previous GeneCards Identifiers for CETP Gene
GC16M047282, GC16P057046, GC16P056730, GC16P056771, GC16P055553, GC16P056996, GC16P042865
Search aliases for CETP gene in PubMed and other databases
Summaries for CETP Gene
NCBI Gene Summary for CETP Gene
The protein encoded by this gene is found in plasma, where it is involved in the transfer of cholesteryl ester from high density lipoprotein (HDL) to other lipoproteins. Defects in this gene


WD
Join at menti.com | use code 5487 5984
Mentimeter
To aks participants about their history of using tobacco, which is probably not a value in the Common Data Element (CDE) "Smoke History Status"?
0
Never smoked
0
Former smoker
0
Current smoker
0
Heavy smoker
Menti
VIB
Choose a slide to present
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?
Which standards do you use in your research?

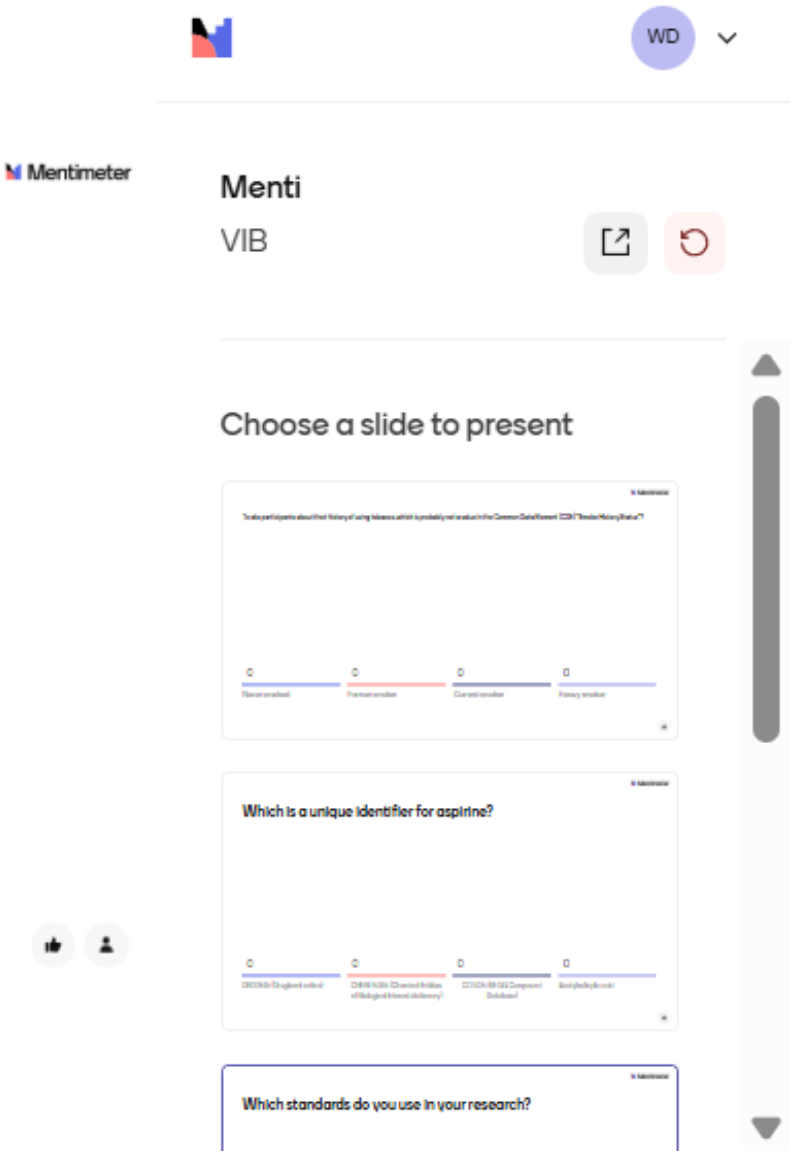

WD
Join at menti.com | use code 5487 5984
Mentimeter
Which is a unique identifier for aspirine?
0
DB00945 (Drugbank online)
0
CHEBI:15365 (Chemical Entities of Biological Interest dictionary)
0
CO1405 (KEGG Compound Database)
0
Acetylsalicylic acid
Menti
VIB
Choose a slide to present
Which is a unique identifier for aspirine?
Which standards do you use in your research?


Join at menti.com | use code 5487 5984
Mentimeter
Menti
VIB
What is a standard for temperature?
0
20 °C
0
293 K
0
68 F
0
Room temperature
Choose a slide to present


Join at menti.com | use code 5487 5984
Mentimeter
Which answer options are suitable to ask about how much participants exercise
0
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
light exercise, moderate exercise, vigorous exercise
Menti
VIB
Choose a slide to present

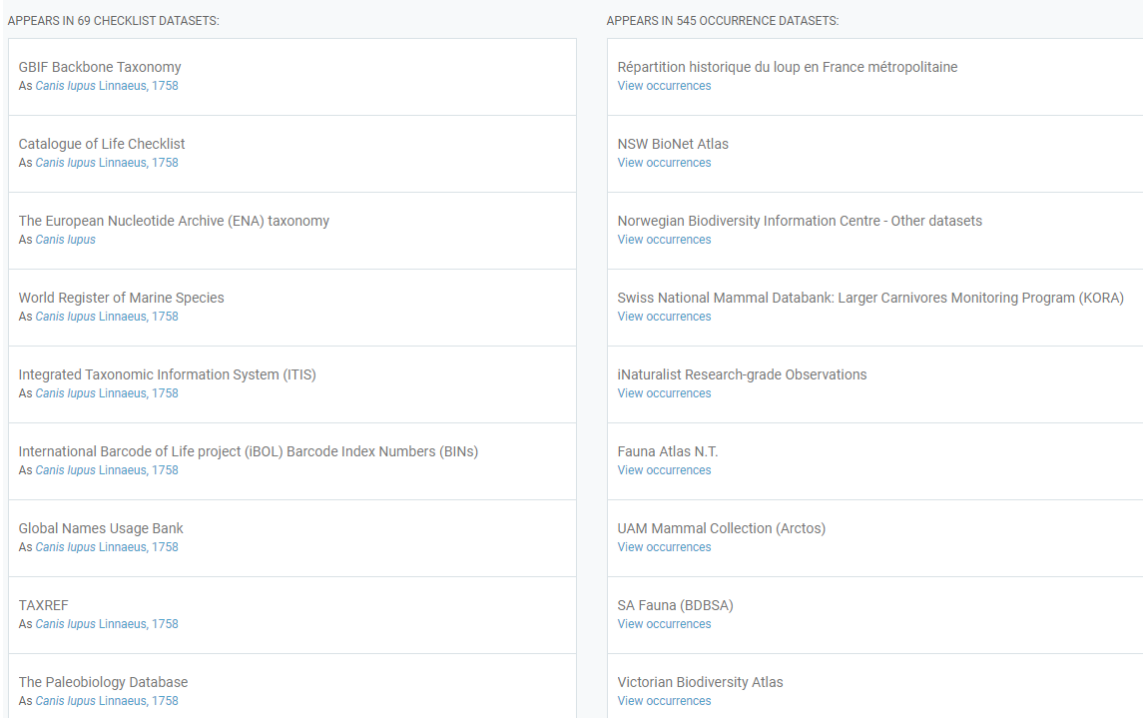
Join at menti.com | use code 5487 5984

Which standards do you use in your research?



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:

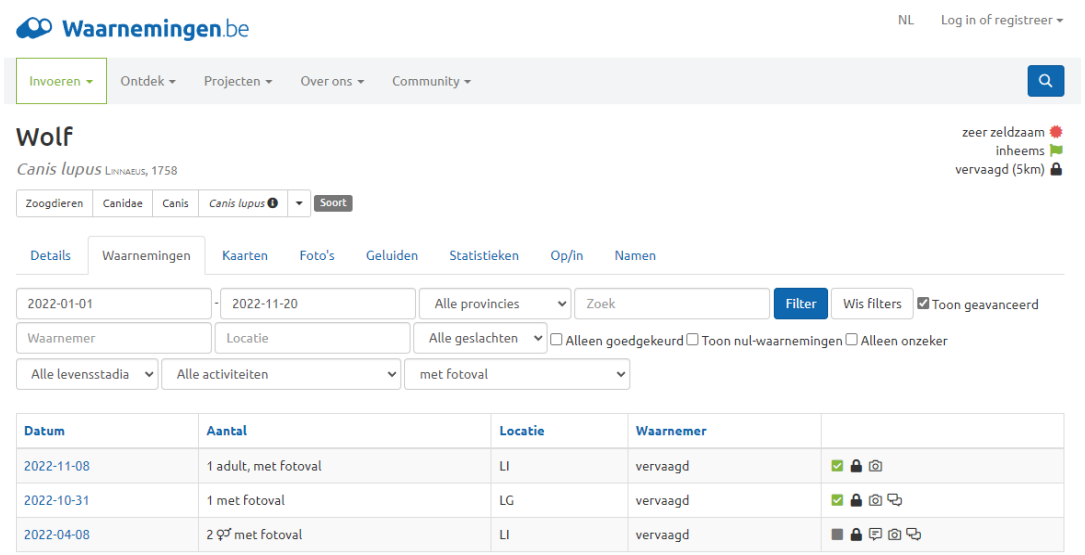
- Répartition historique du loup en France métropolitaine
  View occurrences
- NSW BioNet Atlas
  View occurrences
- Norwegian Biodiversity Information Centre - Other datasets
  View occurrences
- Swiss National Mammal Databank: Larger Carnivores Monitoring Program (KORA)
  View occurrences
- iNaturalist 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	

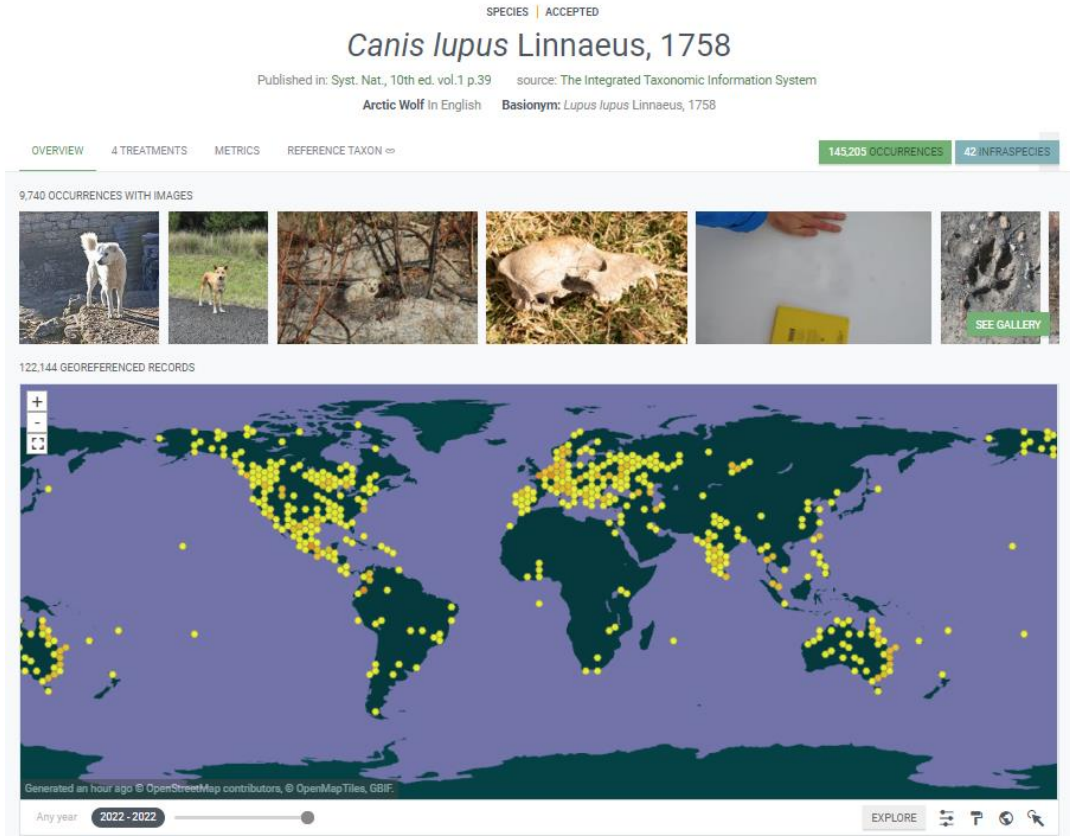
Community standard: biodiversity data

Wolf observations Flanders 2022



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



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

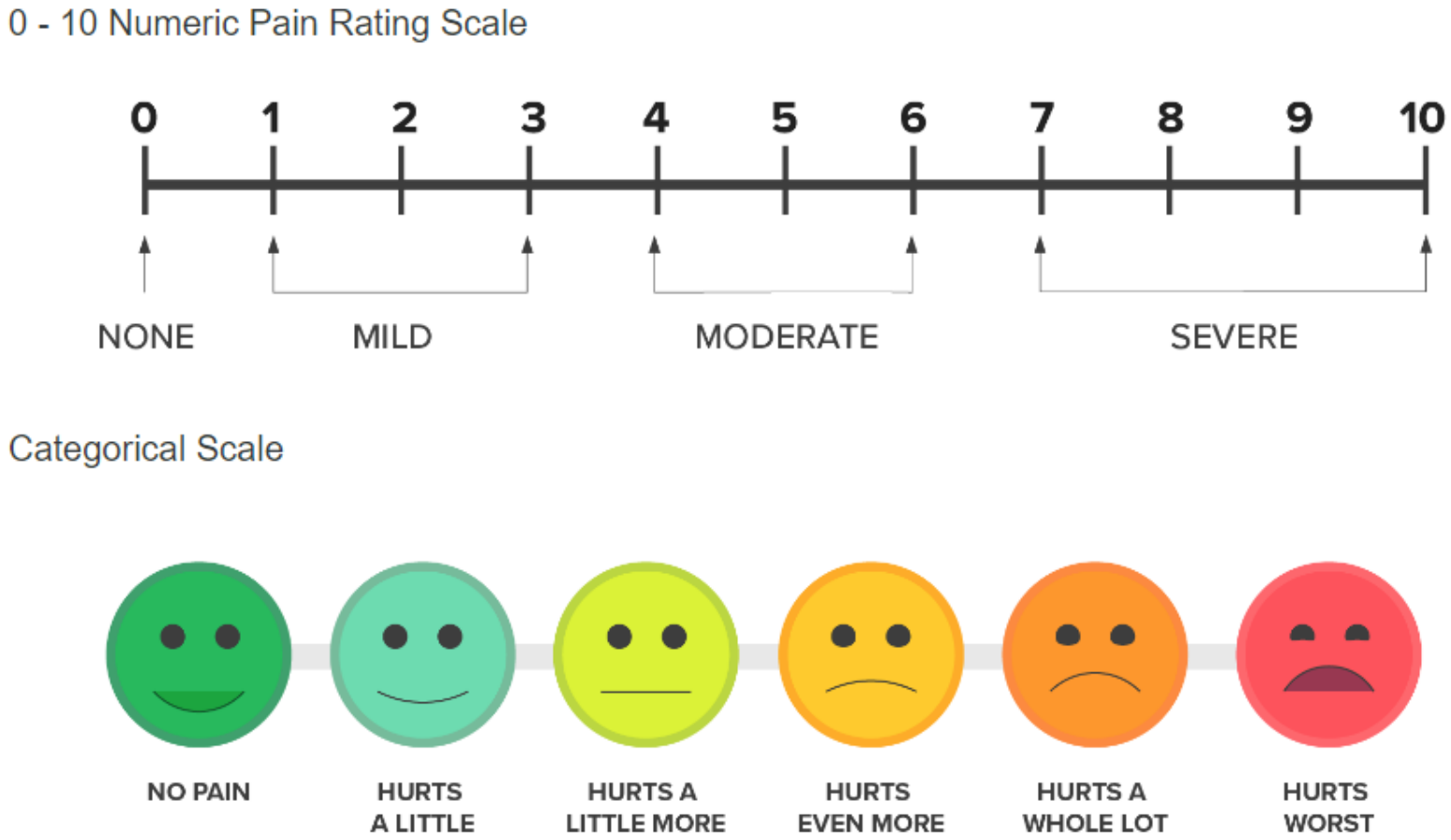
TimeStamp in both datasets facilitates interoperability

	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
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

NIH Common Data Elements

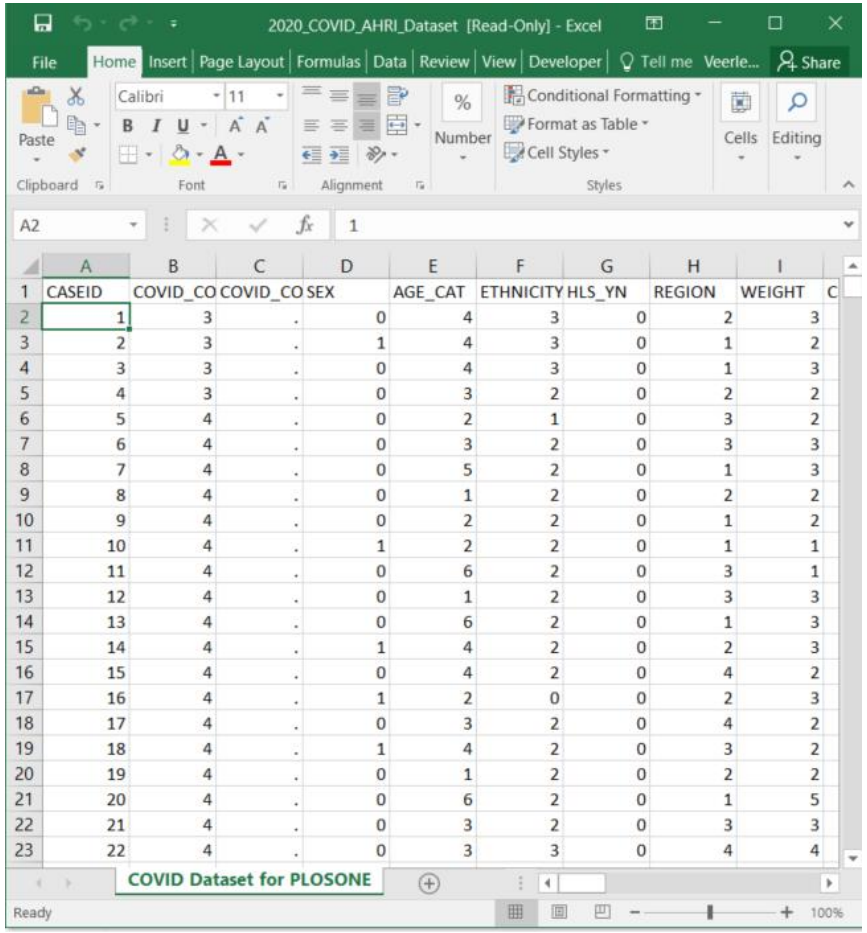
A **Common Data Element** (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection.



[NIH Common Data Elements (CDE) Repository]

Codebook to understand data

Dataset



	A	B	C	D	E	F	G	H	I
1	CASEID	COVID_CO	COVID_CO	SEX	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

	A	B	C
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	Cohorst (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; 5:5
7	ETHNICITY	Ethnicity	0:Other or more than 1; 1:Black or African American; 2:White;
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:Mo
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 complicatio	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 compli	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, cirr	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

- Descriptive: often hard to reproduce
- Visuals, drawing or pictures: helps to reproduce
- Structure: helps to write out and follow 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

Standardisation …


6x2 brick


flag


4x2 brick


4x2 brick with slope


3x2 brick


2x2 brick with slope


2x2 brick

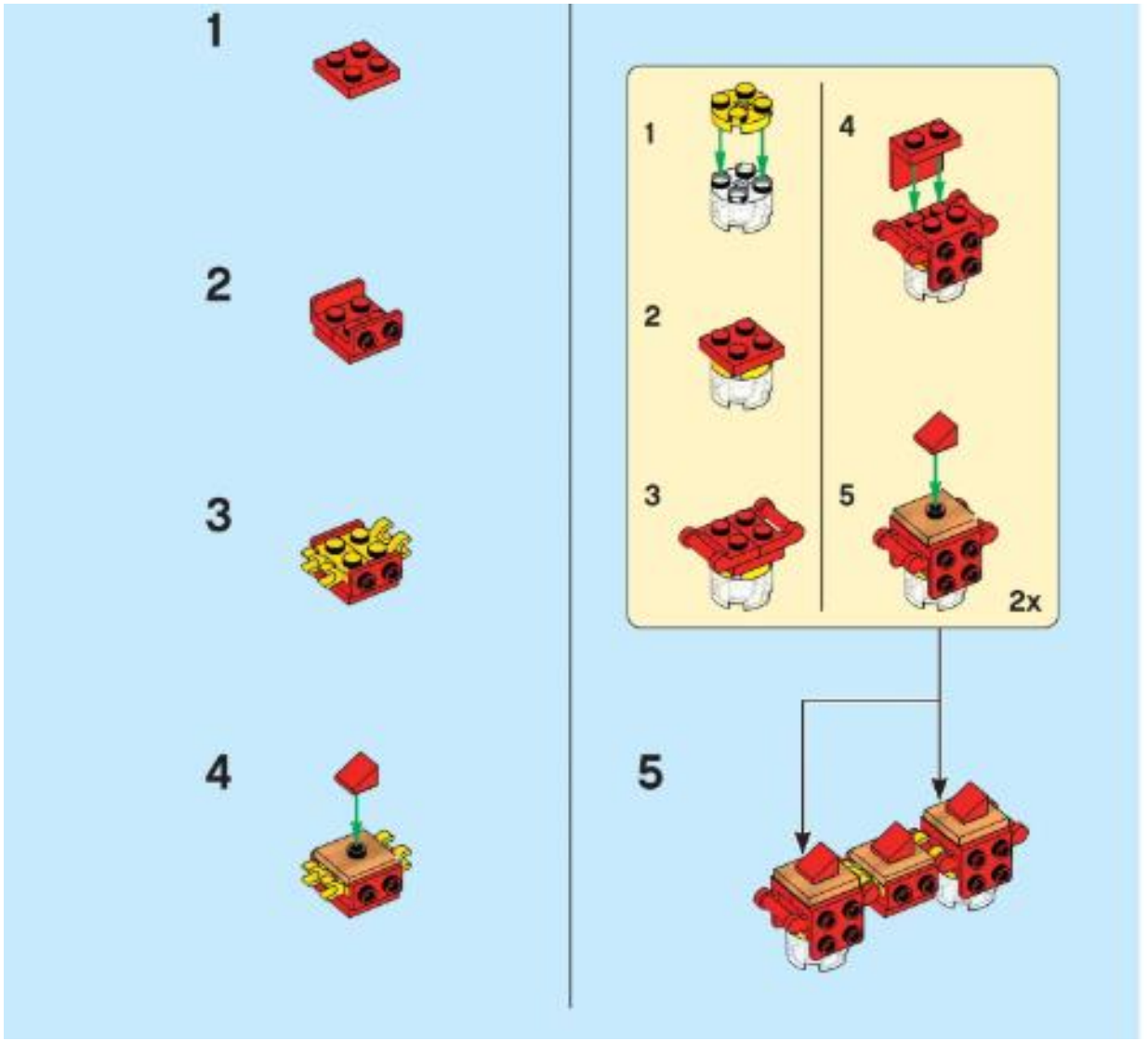

3x2 brick with slope


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			



Brick lists have unique numbers / codes for each brick
= controlled vocabulary / community standard

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, correct where needed, zip and archive
- 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