

# Documenting Your Data



Dr. Carly Huitema

DOCUMENTATION IS A LOVE LETTER THAT  
YOU WRITE TO YOUR FUTURE SELF.

- DAMIAN CONWAY -

# Outline

- Introductions and rational
- FAIR data principles and the role of metadata
- Documenting files, folders and other metadata information
- Glossaries, Ontologies and more
- Persistent Identifiers in research
- Documenting Data Schemas with OCA

# Value of Documenting Data

- Ethical and legal obligations
  - Research Ethics Board.
  - Granting agency, publishing requirements
- Research data is expensive to collect
- Some research data can only be collected once!
- Every time data is reused it increases in value
- Don't make mistakes using your own data (or other's data)
- To get more out of your research data
  - Document it better to make it more understandable
  - Make it available for reuse

# Verification and Reproducibility

- Supports research verification and review
- Reproducibility of research
- What will you leave behind?
  - Can your supervisor understand
  - Can your students follow the work
  - Can you understand 6 months from now?
  - How much time will you spend when it comes time to write?

Comment on a paper by Samir Chatterjee.

[John Cornforth](#)

[Show more ▾](#)

[+ Add to Mendeley](#)  [Share](#)  [Cite](#)

[https://doi.org/10.1016/S0040-4039\(00\)71452-3 ↗](https://doi.org/10.1016/S0040-4039(00)71452-3)

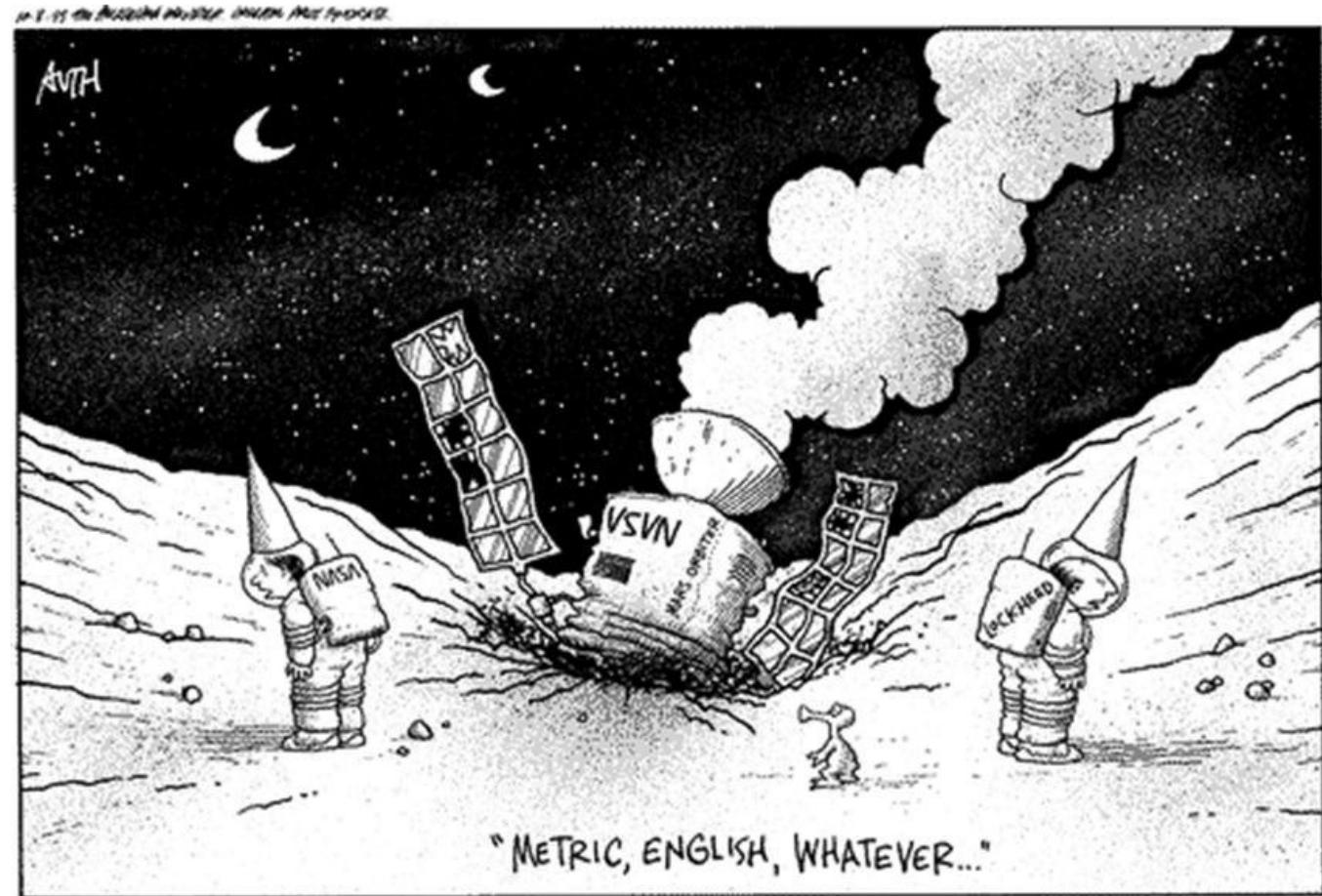
[Get rights and content ↗](#)

## Abstract

All claims in the paper cited (*Tetrahedron Lett.* 3249 (1979)) should be accepted as fact after, but not before, verification by independent experiment.

# Mars Climate Orbiter

- \$125 billion USD total loss (+ lost time and effort)
- Failure to convert from Imperial to Metric units



**Remember the Mars Climate Orbiter incident from 1999?**

Use the FAIR data principles as a  
guide for better data  
documentation

# FAIR Data Principles

- Findable, Accessible, Interoperable, Reusable (FAIR)
  - For people and machines
- Applies to digital resources (like data, software, etc.)
- Formulation of 15 principles under FAIR was in 2016
  - Endorsed by the G20
- FAIR is **not equivalent to open**
  - (and open is not equivalent to ‘free’)
- Note: metadata can be tricky because someone’s metadata can be someone else’s data.
  - That’s why FAIR principles are all written to apply to (meta)data

# FAIR Data Principles

## To be Findable:

- F<sub>1</sub>. (meta)data are assigned a globally unique and persistent identifier
- F<sub>2</sub>. data are described with rich metadata (defined by R<sub>1</sub> below)
- F<sub>3</sub>. metadata clearly and explicitly include the identifier of the data it describes
- F<sub>4</sub>. (meta)data are registered or indexed in a searchable resource

## To be Accessible:

- A<sub>1</sub>. (meta)data are retrievable by their identifier using a standardized communications protocol
- A<sub>1.1</sub> the protocol is open, free, and universally implementable
- A<sub>1.2</sub> the protocol allows for an authentication and authorization procedure, where necessary
- A<sub>2</sub>. metadata are accessible, even when the data are no longer available

## To be Interoperable:

- I<sub>1</sub>. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I<sub>2</sub>. (meta)data use vocabularies that follow FAIR principles
- I<sub>3</sub>. (meta)data include qualified references to other (meta)data

## To be Reusable:

- R<sub>1</sub>. meta(data) are richly described with a plurality of accurate and relevant attributes
- R<sub>1.1</sub>. (meta)data are released with a clear and accessible data usage license
- R<sub>1.2</sub>. (meta)data are associated with detailed provenance
- R<sub>1.3</sub>. (meta)data meet domain-relevant community standards

# Question – what does this mean to you?

## To be Findable:

- F<sub>1</sub>. (meta)data are assigned a globally unique and persistent identifier
- F<sub>2</sub>. data are described with rich metadata (defined by R<sub>1</sub> below)
- F<sub>3</sub>. metadata clearly and explicitly include the identifier of the data it describes
- F<sub>4</sub>. (meta)data are registered or indexed in a searchable resource

## To be Accessible:

- A<sub>1</sub>. (meta)data are retrievable by their identifier using a standardized communications protocol
- A<sub>1.1</sub> the protocol is open, free, and universally implementable
- A<sub>1.2</sub> the protocol allows for an authentication and authorization procedure, where necessary
- A<sub>2</sub>. metadata are accessible, even when the data are no longer available

## To be Interoperable:

- I<sub>1</sub>. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I<sub>2</sub>. (meta)data use vocabularies that follow FAIR principles
- I<sub>3</sub>. (meta)data include qualified references to other (meta)data

## To be Reusable:

- R<sub>1</sub>. meta(data) are richly described with a plurality of accurate and relevant attributes
- R<sub>1.1</sub>. (meta)data are released with a clear and accessible data usage license
- R<sub>1.2</sub>. (meta)data are associated with detailed provenance
- R<sub>1.3</sub>. (meta)data meet domain-relevant community standards

# What is Metadata

- Metadata is data about data
- Line is blurred between what is metadata and what is data
- One person's metadata is another person's data!
- Metadata examples are:
  1. Title and description
  2. Tags and categories
  3. Who created and when
  4. Who last modified and when
  5. What is the data license
  6. How was the data collected and analyzed
  7. Data column descriptions (schema or data dictionary)

You can't identify metadata just by looking at it

## **Professor James Frew's two laws of metadata:**

1. Scientists don't write metadata
2. Any scientist can be forced to write bad metadata



apiary\_monitoring\_2016\_open\_data.xlsx

## Honey bee pests and pathogens in Ontario apiaries

Get data on pests and pathogens measured in honey bee apiaries in Ontario.

Monitoring honey bee pests and pathogens in Ontario apiaries is a key objective of Ontario's Pollinator Health Action Plan. To achieve this, the Ontario government began a 6-year monitoring project in 2015 to create an inventory of honey bee pests and pathogens found in Ontario apiaries and assess the prevalence and load of these pathogens. Apiary monitoring will continue until 2020.

### For more information

Contact Agriculture, Food and Rural Affairs

Metadata

Data

Site #	Inspection #	Year	Inspection Date	Region	County	Number of Colonies Inspected (#)	American Foulbrood Positive Colonies (#)	European Foulbrood Positive Colonies (#)	Sacbrood Positive Colonies (#)
1	1	2016	5/7/2016	East	STORMONT, DUNDAS AND GLENGARRY	10	0	0	0
1	2	2016	5/25/2016	East	STORMONT, DUNDAS AND GLENGARRY	10	0	0	0
1	3	2016	8/15/2016	East	STORMONT, DUNDAS AND GLENGARRY	10	0	0	0
1	4	2016	9/22/2016	East	STORMONT, DUNDAS AND GLENGARRY	10	0	0	0
2	1	2016	5/5/2016	South	WELLINGTON	10	0	0	0
2	2	2016	6/9/2016	South	WELLINGTON	10	0	0	0
2	3	2016	8/13/2016	South	WELLINGTON	10	0	0	0
2	4	2016	9/27/2016	South	WELLINGTON	10	0	0	0
3	1	2016	5/7/2016	South	BRANTFORD	10	0	0	0
3	2	2016	5/23/2016	South	BRANTFORD	10	1	0	0
3	3	2016	8/6/2016	South	BRANTFORD	6	0	0	0
3	4	2016	10/4/2016	South	BRANTFORD	10	0	0	0
4	1	2016	5/10/2016	Central	GREY	10	0	0	0

### Additional Information

Field	Value
Last updated	September 10, 2021
Created	September 10, 2021
Format	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
File size	77.5 KiB
License	Open Government Licence – Ontario
Name	Apiary Monitoring 2016
Type	Data
Language	English and French
Contains geographic markers	
Data made public date	2021-09-10
Data range start	2016-04-18
Data range end	2016-10-05
Data birth date	
Version	

# Catalogue metadata

- Let's you find, use, and cite a dataset

Who is the dataset owner?

What year was it published?

What is a title and description?

What is the data use license?

e.g. library catalogue entry, dataverse entry

Hyperspectral time series datasets of maize during the grain filling period

Sep 24, 2021 - Department of Plant Agriculture

 Craig, Valerie; Earl, Hugh; Sulik, John; Lee, Elizabeth A., 2021, "Hyperspectral time series datasets of maize during the grain filling period", <https://doi.org/10.5683/SP2/1ZVWFV>, Scholars Portal Dataverse, V1, UNF:6:aHOe23WjEa9jEdMRtXkI1Q== [fileUNF]

Remotely sensed hyperspectral data are increasingly being used to assess crop development and growth throughout the growing season. Large datasets capturing key growth stages can be useful to researchers studying many physiological plant responses. This dataset represents a time...

# Schema Metadata/Data Dictionary

The diagram illustrates the relationship between a data table and its corresponding schema metadata. A blue arrow points from the 'DATA' table to the 'DATA DICTIONARY (METADATA)' table, indicating that the dictionary provides the definition for the data in the table.

**DATA**

employee_id	first_name	last_name	nin	dept_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Bamy	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1

**DATA DICTIONARY (METADATA)**

Column	Data Type	Description
employee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current position title, e.g. Secretary
dept_id	int	Employee department. Ref: Departments
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

 Dataedo

- What format is the data in?
- Necessary to know how to use the data
- Related to interoperability
- Data in a proprietary schema (like an instrument specific result file) is not very useful.

# Additional Descriptive Metadata

Example: Protein Data Bank Entries

- Additional information about the data that helps a researcher understand it
- Specific to the data type

[Display](#)

**6M2N**

SARS-CoV-2 3CL protease (3CL pro) in complex with a novel inhibitor

DOI: [10.22110/pdb6M2N/pdb](https://doi.org/10.22110/pdb6M2N/pdb)

Classification: **VIRAL PROTEIN**

Organism(s): Severe acute respiratory syndrome coronavirus 2

Expression System: Escherichia coli BL21(DE3)

Mutation(s): No [i](#)

Deposited: 2020-02-28 Released: 2020-04-15

Deposition Author(s): Su, H.X., Zhao, W.F., Li, M.J., Xie, H., Xu, Y.C.

---

**Experimental Data Snapshot**

Method: X-RAY DIFFRACTION

Resolution: 2.20 Å

R-Value Free: 0.254

R-Value Work: 0.225

R-Value Observed: 0.227

---

**wwPDB Validation** [i](#)

Metric	Percent
Rfree	100
Clashscore	100
Ramachandran outliers	100
Sidechain outliers	100

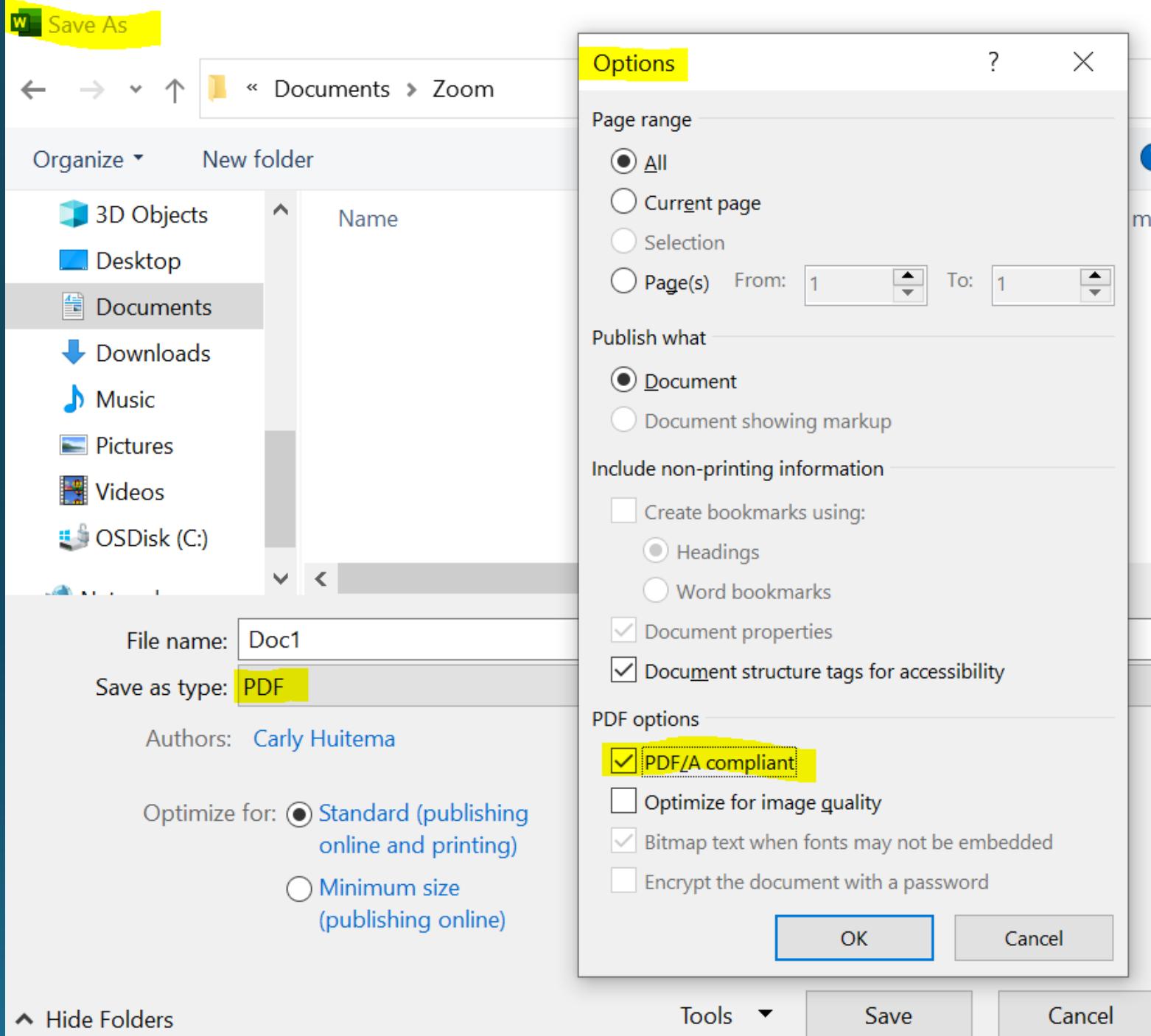
# Metadata Adds Value to Data

- Catalogue metadata helps researchers find and cite datasets
- Schema/data dictionary metadata helps researchers use data
- Additional metadata helps researchers understand the context of the data
  - Am I allowed to use this data?
  - Where and when was this data collected?
  - What units are the measurements in?
  - How was this data collected?
  - Is the dataset appropriate to answer my question?

# Levels of Metadata for your Research

- Study level
  - What is the research problem answering
  - What is the design of the experiments
  - Who is funding, who is supervising, who is involved
  - Copies of relevant documents e.g. survey text, consent request text, consent documentation, approved protocols etc.
  - Data management plan
  - Instruments used to collect data
- How to store it
  - Readme text files
  - Data Management Plan (living document)
  - PDF/A

- PDF/A
- Archive version of PDF



# Levels of Metadata for your Research

- File, dataset or database level
  - What are in each file
  - How are the files related
  - How is the database structured
  - Code used
- How to store it
  - Readme text files
  - A Readme sheet in all Excel files – what the experiment is, reference specific lab book pages, hypothesis being tested, dates of data collection or sources of data etc.
  - Code documentation and scripts
    - You can store scripts for manipulating data as a separate tab in Excel

Example of good dataset level documentation (on GitHub) – but not machine readable

# Software in research survey - 2014

---

## Introduction

---

In 2012 the Software Sustainability Institute ran a survey of researchers at 15 research-intensive universities in the UK to uncover their attitudes to software. For reasons that will be explained in more detail in a forthcoming blog post, the analysis of these results was conducted in Excel. To improve the transparency and reproducibility of these results, this analysis has now been repeated in Python.

## Important points

---

- Licence for the code and data can be found in the the LICENCE and LICENCE\_DATA files respectively.
- The code runs on Python 3.
- The data derives from the [2014 software in research survey](#).

## Summary of process

---

1. Get raw survey results from survey software ([iSurvey](#))
2. Anonymise data by manually deleting "Email" and "Further comments" fields.
3. Make Question 11 parsable in Python

# Levels of Metadata for your Research

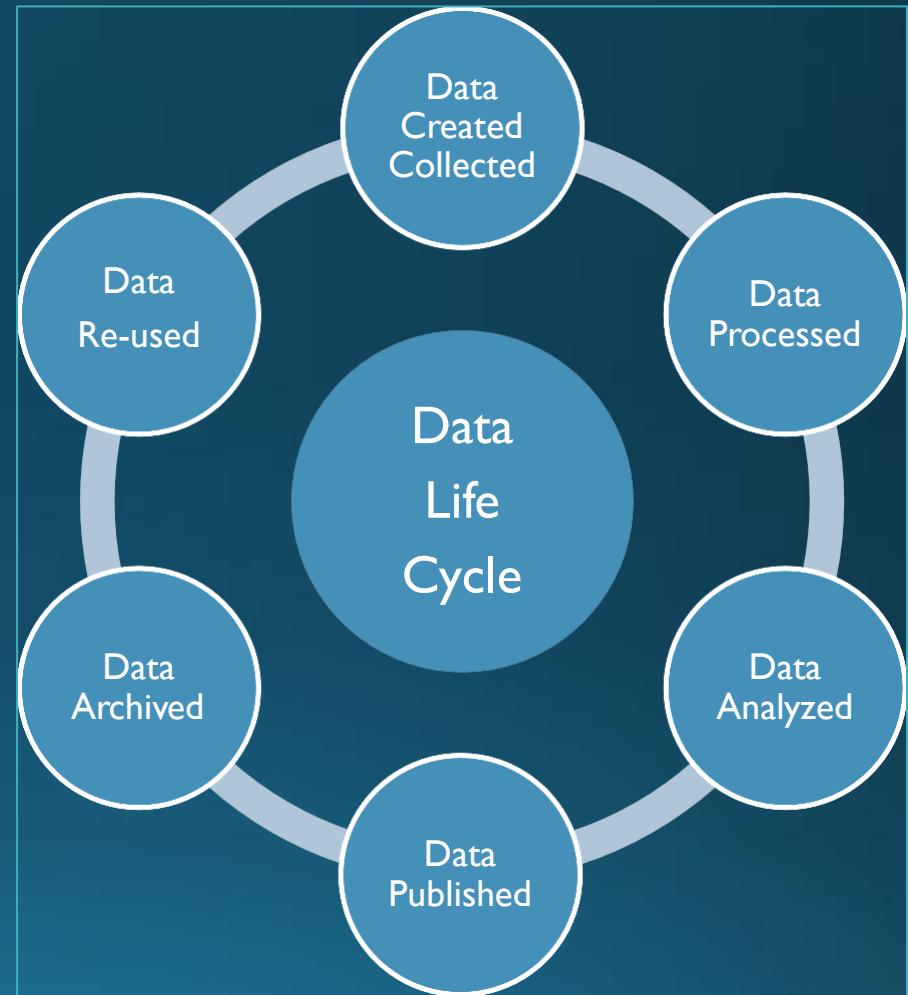
- Variable or item level
  - What are columns of data in a spreadsheet
  - What are questions in a questionnaire
- How to store it
  - Readme file or Excel tab giving details
  - Codebook/data dictionary
  - OCA data schema (Discussed later) – machine readable

# When to document data – Data Lifecycle

- At the very end when it is time to submit the paper/thesis

OR

- At every step along the way when it is easy to remember what was done?



# Tools for Data Documentation

- **Keep the raw data** – write scripts (best) or notes about how you manipulated the data for derived datasets
- Scripts instead of point-and-click or remember processes
  - Document point-and-click or “I’m sure I’ll remember what I did” things
- SOPs (Standard Operating Procedures)
- Readme files
- Readme tabs to Excel sheets
- Version control systems like GitHub (especially for code but also for documentation)

# File Organization

- Saves time and storage space
  - Finding things now and in the future
- Necessary for collaboration
  - Your supervisor can find your files
  - Your collaborators can all work together on a project and share data
- Prevent data loss
  - Find the data when you need to write up the paper/thesis
- Find the data for publication
- Let others reuse your data

# File and folder naming

- Helps identify content of the file (or folder)
- Think of how you want files sorted in folder
- Less than 25 characters – preferred
  - Avoid unusual characters !@#\$%^&\*()+
  - Use underscores between words or capitalize first letter of each word
- Names independent of location (create project id or acronym)
- Version identification e.g. v01
- Example:
  - afs\_codebook\_2018-02-13.pdf
  - afsCodebook20180213.pdf



*Interpretation* = project id\_description of file\_ISO date format.file format

# Dates

- Year than month than day
- Keeps files in chronological order in the folder

## PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27.  $\frac{2}{2}$ -13 2013.158904109

MMXIII-II-XXVII MMXIII  $\frac{LVII}{CCCLXV}$  1330300800

$((3+3)\times(111+1)-1)\times3/3-1/3^3$  2013 Mississ 2-27-13

10/11011/1101 02/27/20/13  $\frac{2}{5} \frac{3}{6} \frac{1}{7} \frac{4}{8}$  01237

# Folder Organization

- Set up Project Folder structure
  - Follow the structure of your project
  - Avoid overlapping categories or similar folders
  - Restructure when needed
  - Archive instead of delete (archive folders)
- Assign an acronym to your project:
  - E.g. Alpaca Fibre Study = AFS
- All folders will start with this acronym – e.g. AFS\_Budget
  - Keep your folder names short and clear to understand
  - NO spaces!!!! Use an underscore \_

# Organizing your project folders

- AFS ← Top folder for project
  - AFS\_Data ← Where all data will be saved for this project
    - AFS\_Data\_201806 ← Data collected in June 2018 is saved here
      - AFS\_Data\_201806\_Suri.xlsx ← Data collected in June 2018 from Suri breeders
      - AFS\_Data\_201806\_Huacaya.xlsx ← Data collected in June 2018 from Huacaya breeders

# Organizing your project folders

- Create a README file – save in your top directory or main folder A text file that:
  - Defines your acronyms
  - Describes your project and the folder structure
  - Defines what files will be in each directory or folder
  - Think of this README file as an annotated Table of Contents to your project folder structure

AFS\_README.txt

# Readme – Starting to Document

Title: Alpaca Fibre Study (AFS)

Short abstract or project statement

AFS\_Budget = Budget information for the project

AFS\_Data = Data collected

- AFS\_Data\_201806; AFS\_Data\_201807; AFS\_Data\_201808
  - AFS\_SAS = All SAS programs
  - AFS\_Output = All SAS outputs
  - Data collected from 2018-06-01 to 2019-06-01
- Price data collected in dollars per pound
- NOTE: 2018-06-15 Rain caused data collection to be delayed until 2018-06-20

# Filenames - examples

- AFS\_Budget
  - AFS\_Budget\_2018\_Expenses.xlsx
  - AFS\_Budget\_2018\_Revenues.docx
- AFS\_SAS
  - AFS\_SAS\_20180630\_DescStats.sas
  - AFS\_SAS\_20180630\_Model.sas
- Provide detailed explanation of contents in the README file!

# Plan now and Save time later!

- Sounds like a lot of work to plan out your directories, file names, and document it!
- It will save you a lot of time later!
  - Especially when you go back after being away for a bit.
- Be consistent



# Discussion - Organization

- How are you currently organizing your files?
- Let's try to have a conversation

# Best Practices for variable names

1. Set Maximum length to 32 characters
2. ALWAYS start variable names with a letter
3. Numbers can be used anywhere in the variable name AFTER the first character
4. ONLY use underscores "\_" in a variable name
5. Do NOT use blanks or spaces
6. Use lowercase

# Variable names inside my files

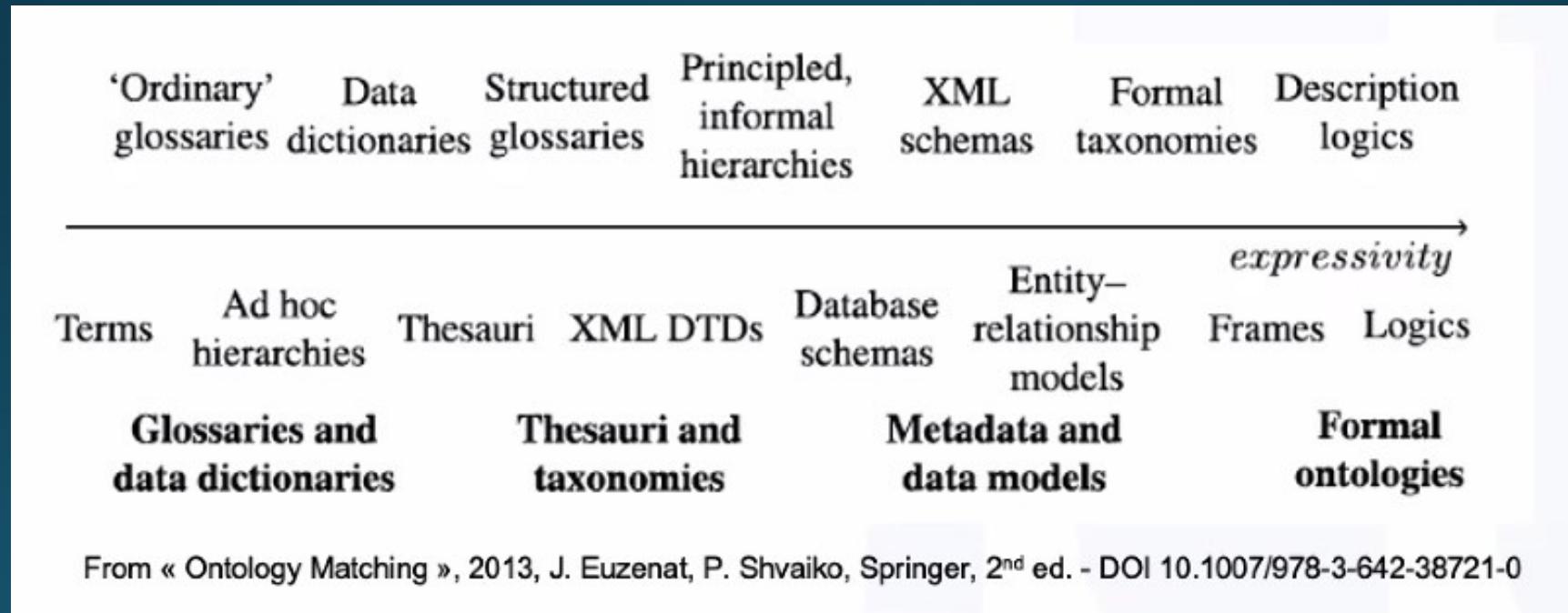
- Information or data that we are collecting:
  - Diet A → diet\_a
  - Fibre length in centimetres → fibre\_cm
  - Location of farm → location
  - Price paid for fleece → price

Coffee Break  
10 minutes



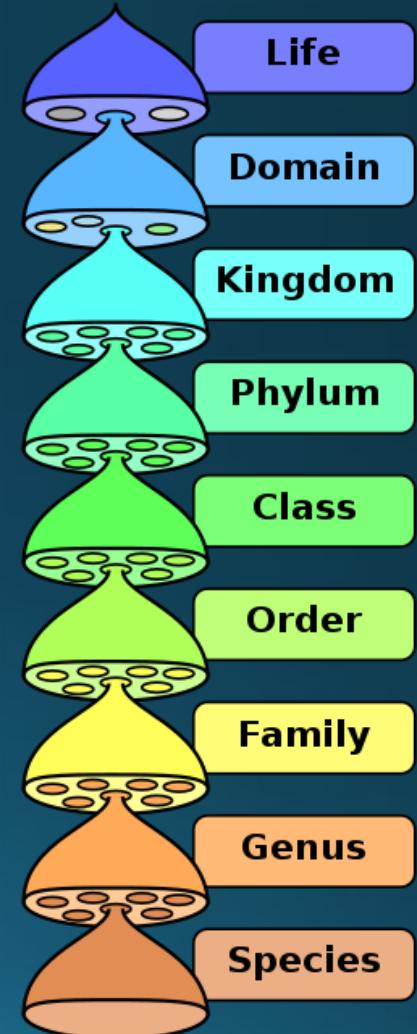
# Semantic Artefacts

- Glossaries, ontologies and more
- To unambiguously use defined terms (and their relationships)



# Glossaries, ontologies and more

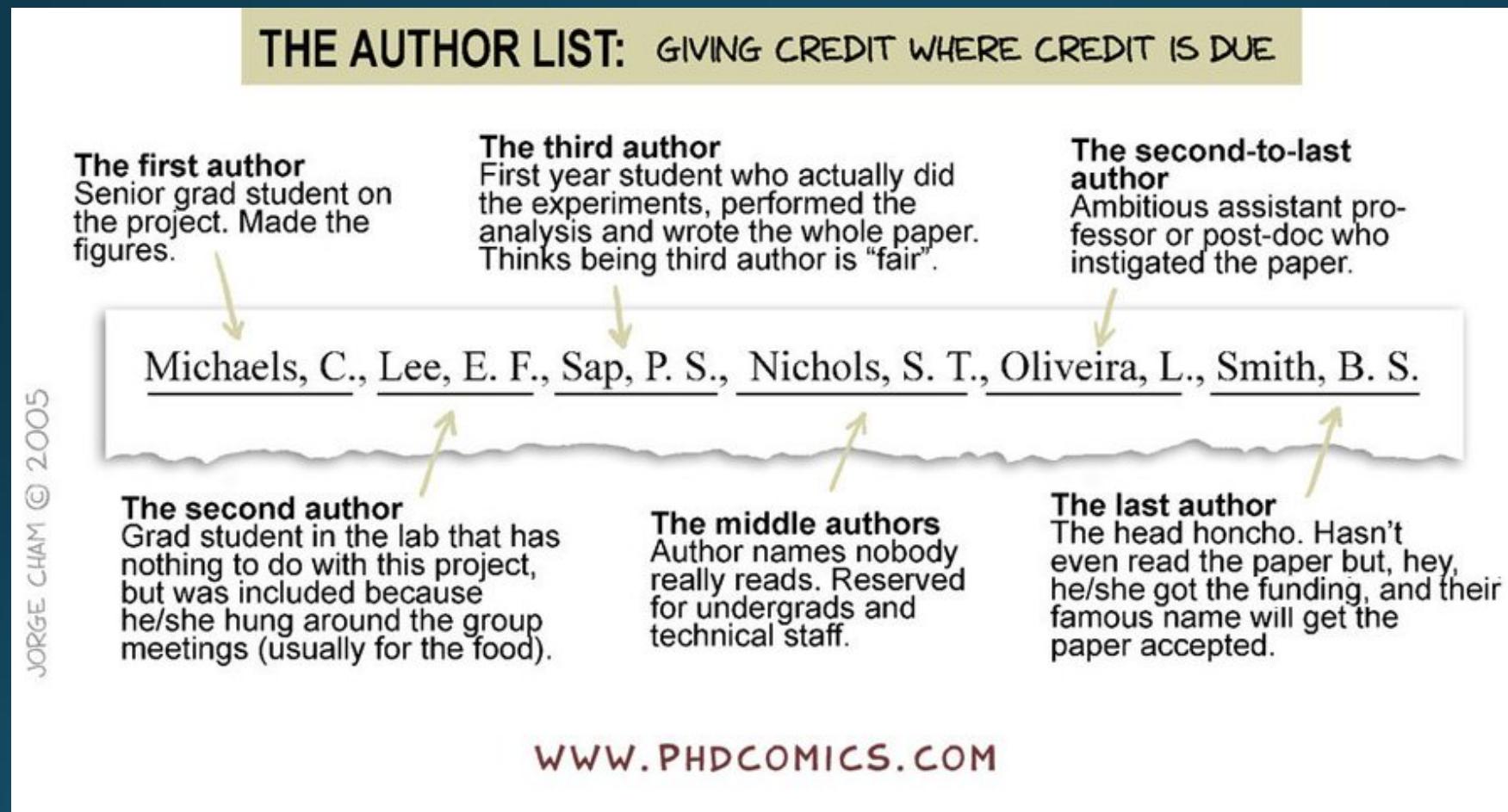
- Examples of semantic artefacts:
  - Linnaean taxonomy, IUPAC chemical nomenclature, SI units of measurement, MeSH terms for PubMed etc.
- Usage improves data FAIRness
  - Especially interoperability



# Controlled Vocabulary

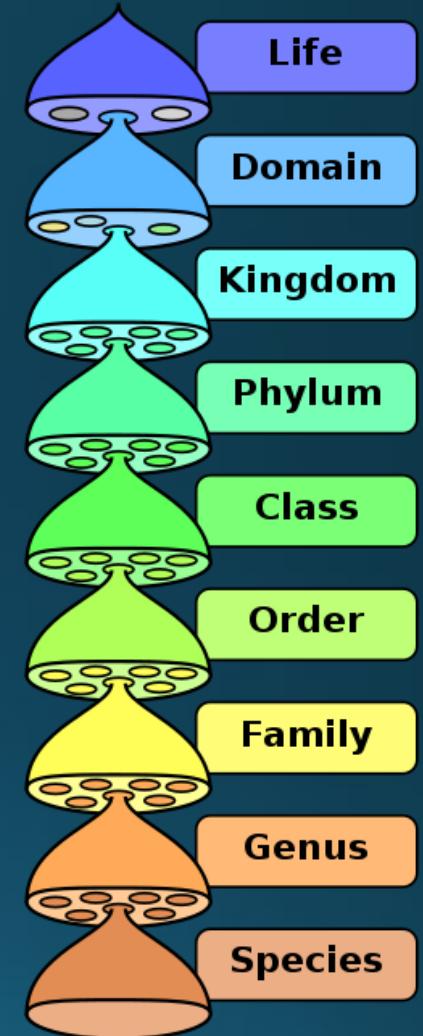
- For data/information harmonization
  - E.g. names of offices at a University
  - Office of Research Services (ORS) will be known as Research Services Office (RSO) at the University of Guelph
- To improve searching
  - Protease vs proteinase vs proteases etc.
- To reference a known term without having to define it again
- Challenge – term lists may reduce noise but decrease precision
  - E.g. classifying your research from a keyword list

# CRedit Taxonomy ....before



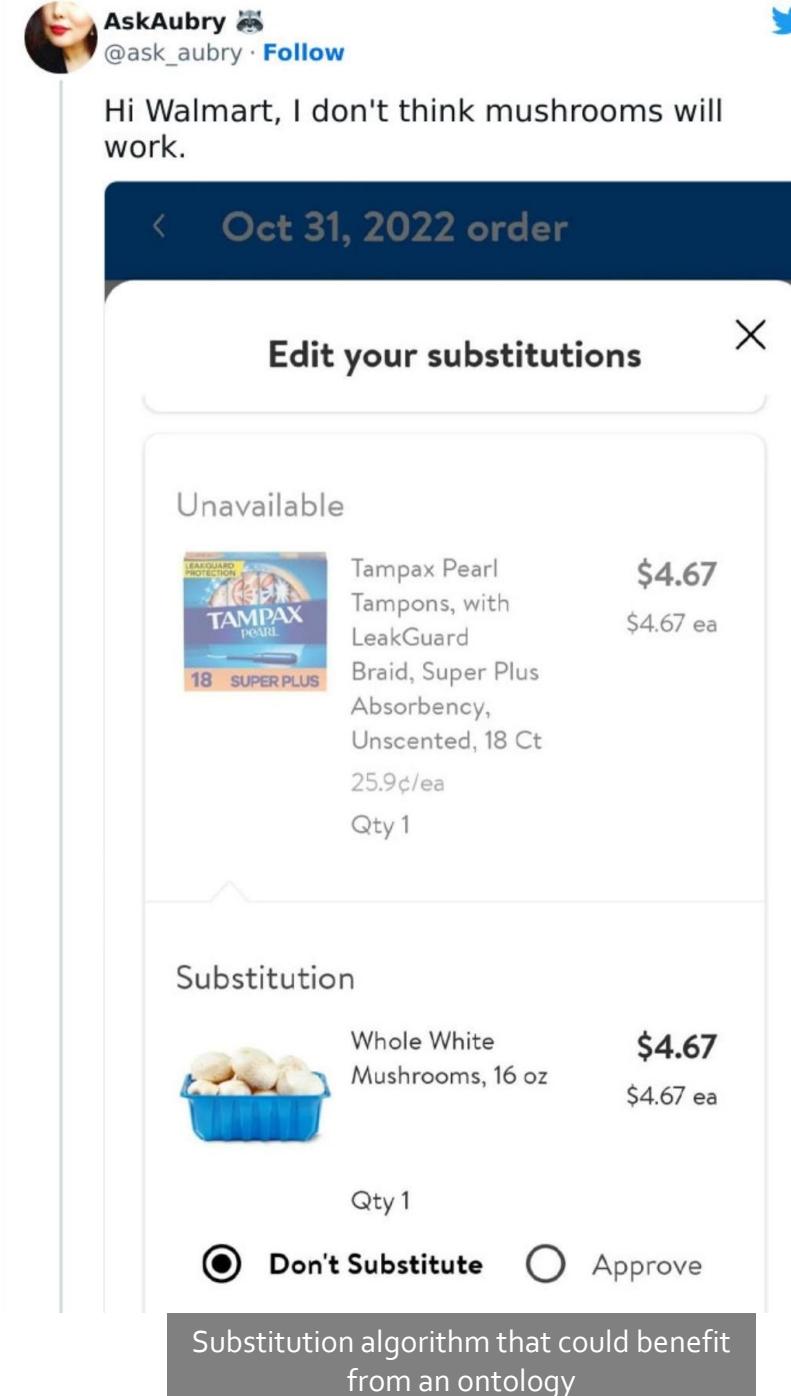
# Taxonomy

- Biological Taxonomy
- CRediT taxonomy
  - Contributor Roles Taxonomy – used for publication contributors
  - High-level taxonomy, including 14 roles.
  - The roles describe each contributor's specific contribution to the scholarly output.
  - <https://credit.niso.org/>
- is-a relationships
  - parent-child relationships between concepts
  - A cat is-a mammal



# Ontologies

- Formal representation of domain knowledge
  - Concepts and their associated relationships
  - Can be hierarchical
  - Pure vs Practical
- Machine actionable and useful for computers
  - AI and Machine Learning
- Valuable in model building, searching, linking resources
- More complex relationships can be modelled
  - Has-a, use-a



A screenshot of a mobile application interface. At the top, there's a header with a profile picture of a woman, the handle '@ask\_aubry', and a 'Follow' button. To the right of the header is a small blue Twitter icon. Below the header, a message from the user says, 'Hi Walmart, I don't think mushrooms will work.' Underneath this message is a dark blue bar with the text 'Oct 31, 2022 order' and a back arrow icon. The main content area has a light gray background and a white rounded rectangle for the 'Edit your substitutions' section. This section contains two items:

- Unavailable:** A product image of Tampax Pearl Tampons, with LeakGuard, Braid, Super Plus Absorbency, Unscented, 18 Ct. The price is \$4.67 and it costs \$4.67 ea. The quantity is Qty 1.
- Substitution:** An image of Whole White Mushrooms in a blue basket. The price is \$4.67 and it costs \$4.67 ea. The quantity is Qty 1.

At the bottom of the substitution section are two buttons: a filled circle labeled 'Don't Substitute' and an empty circle labeled 'Approve'. A note at the bottom states, 'Substitution algorithm that could benefit from an ontology'.

# Using semantic artifacts

- Identify those semantic artefacts used in your domain
- Features to evaluate:
  1. What license and terms of use does it mandate?
  2. What format does it come in?
  3. Is it well maintained
    - i.e. frequent release, term requests handling, versioning and deprecation policies clarified?
  4. Are there stable persistent resolvable identifiers for all terms?
  5. Who use it and what resources are being annotated with it?
  6. Is it well documented?
    - There should be enough metadata for each class in the artefact and enough metadata about the artefact itself.

# Examples of semantic artefacts

- Gene Ontology Resource (GO)
- Enzyme Commission Number (EC number)
- FoodON – Food Ontology
  - A harmonized food ontology to increase global food traceability, quality control and data integration
- OBO Foundry (Open Biological and Biomedical Ontology Foundry)
  - Search for ontologies
- KEGG
  - Kyoto Encyclopedia of Genes and Genomes

# Check-in: semantic artefacts

- What kinds of semantic artefacts are you using now?
- How would you want to use them more?

# Persistent Identifiers (PIPs)

- We want to consistently reference/cite specific references
- A PID is a long-lasting digital identifier to a resource
- A URL is not a PID

The screenshot shows the Okta website's 404 error page. The header includes the Okta logo, navigation links for Customers, Solutions, Products, Services, Resources, Company, Contact Sales, a Try Okta button, a Sign In button, a search icon, and a globe icon. The main content features a large blue curved graphic on the right. The title "Page Not Found" is prominently displayed, followed by the subtext "Ah, the dreaded 404. Let's get you back on track." Below this, there is a section titled "Common issues that can result in broken links:" with three items: "Truncated URLs", "Case sensitivity", and "Old news". Each item has a brief description.

Common issues that can result in broken links:
Truncated URLs
Some email applications may break a URL due to line breaks. You may need to cut and paste a URL from your email application into your browser's location field.
Case sensitivity
URLs are case sensitive. "File.html" and "file.html" are not considered the same file by the server. Please verify that the URL is entered correctly as printed.
Old news
The web page or file you were trying to reach may have been moved to a different location or removed because the information it contained is no longer valid.

# Basic structure of PIDs

- A PID resolver that takes a given identifier and returns the associated metadata record from a lookup table
  - The metadata record is controlled/updated by ‘someone’
  - The metadata record typically contains a URL to digital resources
    - This URL should be updated to reflect when the location of the digital resource changes by the ‘someone’ who controls the metadata record
- You enter the PID into a resolver service, which retrieves the metadata record and possibly redirects you to the URL
- PIDs don’t have to point to just *digital* resources
  - E.g. physical samples (like ARK identifiers)
- An organization **pledges to maintain the PID lookup table**

# DOI

- Digital identifier of an object
- 12 DOI Registration Agencies
- Registration agencies delegate minting of DOIs to content owners (like publishers)
  - Each agency has a specific catalogue metadata schema they use
    - Crossref (specialized for articles etc.)
    - DataCite (initially for data, but many other types as well)
  - Content owners keep DOI metadata updated
    - E.g. when the journal webpage is reorganized all URLs need to be updated

# DOI as a URL

- A DOI is not represented as a URL, it is an identifier with form:

10.3352/jeehp.2013.10.3

- But you can put <https://doi.org> in front of it to have it resolve automatically via a URL

<https://doi.org/10.3352/jeehp.2013.10.3>

- Citation rules change for how DOI is represented

# Places to get DOIs

- Journal submissions
  - Part of the publication process
- Borealis (dataverse) submissions at the University of Guelph
  - Submissions curated by the library
- Zenodo – CERN backed repository
  - Open submissions.  
E.g. Archive GitHub repository on Zenodo and get a DOI
- FigShare
  - Open Submissions
- Open Science Framework (OSF)
  - Project data, project registrations and more

# ORCID Identifiers

- Researcher identifier
  - Unique, open, digital
  - Differentiates you from similar named individuals
  - Use the same identifier throughout your career
  - Your ORCID stays the same even as you change names etc.
- Only the ORCID registry can assign identifiers
  - Unlike DOIs which can delegate
- Users can update their own metadata records
  - Unlike DOIs where metadata is supplied by content owners to registration agencies
- Format: <https://orcid.org/0000-0001-2345-6789>

# ROR Identifiers

- Research Organization Registry (ROR)
- PIDs for research organizations
- Preferred format is the URL
  - <https://ror.org/o1r7awg59>
- RORs are community curated and maintained

 <https://ror.org/01r7awg59>

## University of Guelph

### ORGANIZATION TYPE

Education

### LOCATION

Guelph (GeoNames ID [5967629](#))

Canada

### OTHER NAMES

University of Guelph

### OTHER IDENTIFIERS

 <https://ror.org/03srgbm67>

## University of Guelph-Humber

### ORGANIZATION TYPE

Education

### LOCATION

Etobicoke (GeoNames ID [5950267](#))

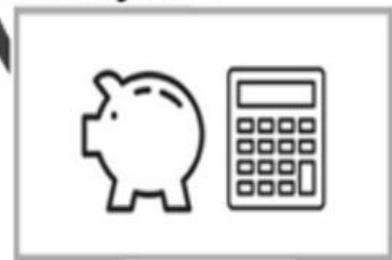
Canada

### OTHER NAMES

### OTHER IDENTIFIERS



Grant IDs



Funder IDs



# Checkin

- What kinds of identifiers do you use?
- Do you have an ORCID?
- Have you created identifiers for research outputs such as blogs, posters, presentations, etc.?

# Overlays Capture Architecture Documenting your Data Schemas

# What are schemas?

The data

110521	12.2	56.8
120521	10.8	12.6
110521	12.6	NA
110621	9.8	55
110521	8.4	32.4
120621	6.3	52.0

# Schemas describe the ‘columns’ (attributes) of your dataset

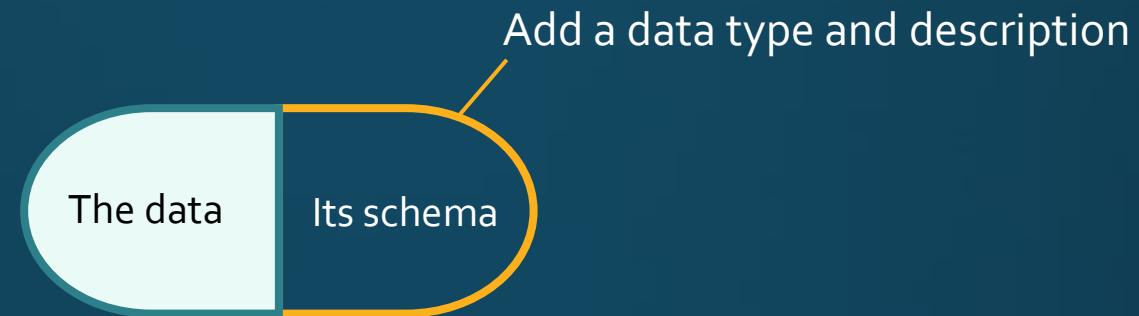


Date	Depth	conc
110521	12.2	56.8
120521	10.8	12.6
110521	12.6	NA
110621	9.8	55
110521	8.4	32.4
120621	6.3	52.0

**This data is described weakly**

What format are the dates in?  
What units are depth and conc?  
Can a machine use this?  
Can I combine with similar datasets?

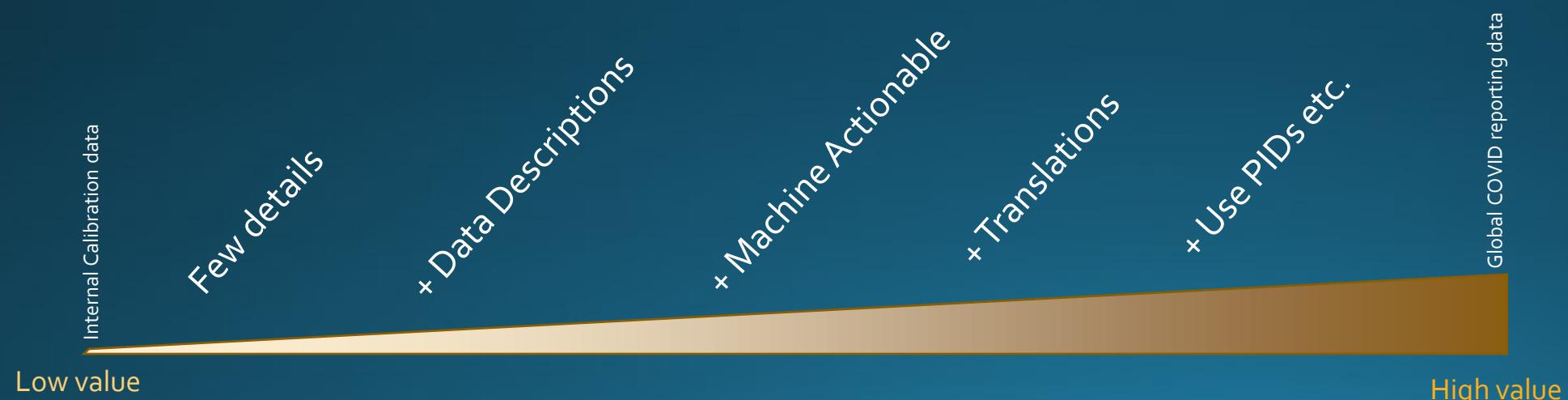
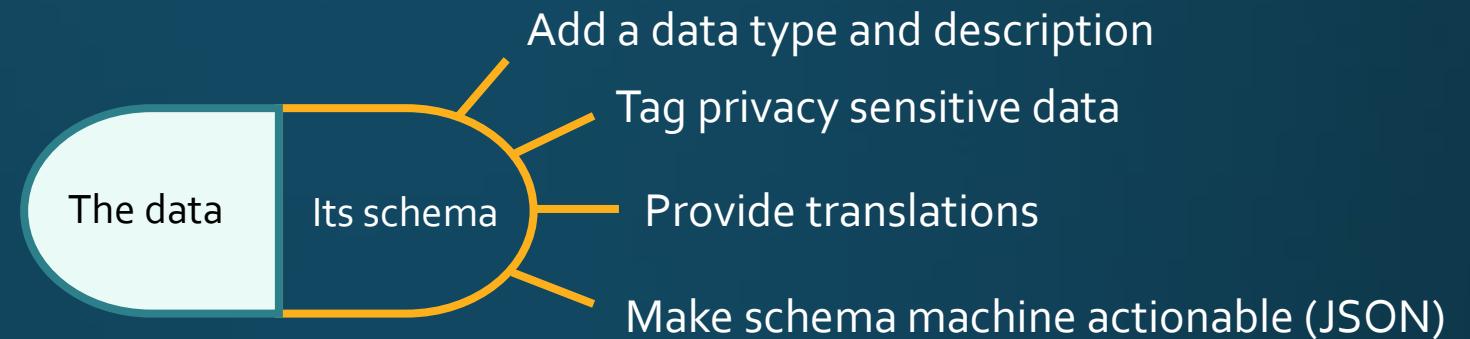
# You can improve data quality by better describing its organization



## A Schema

Column	Data Type	Description
Date	DateTime	Collection date in format DDMMYY
Depth	Numeric	Sample depth in cm
Conc	Numeric	Concentration of Carbon in uM

# Improve data quality by better describing its organization



The value and usability of the data depends on how well it is described



# Universal Semantic Engine – creating a schema

The data

Its schema

- Bob has created a table of data. He wants to describe his data better with a schema.
- He goes to the Semantic Engine webpage and selects create a new schema.
- He adds entries for each column of his data.
- Then he adds information describing the columns. He describes the data type, he adds a description. He has the option to add more information but decides that this is all his data needs.
- He exports his schema as an Excel sheet and a machine readable JSON and he stores it together with his data.
- His research is more interoperable



Column	Data Type	Description
Date	DateTime	Collection date in format DDMMYY
Depth	Numeric	Sample depth in cm
Conc	Numeric	Concentration of Carbon in uM

# The semantic engine and schemas

The data

Date	Depth	conc
110521	12.2	56.8
120521	10.8	12.6
110521	12.6	NA
110621	9.8	55
110521	8.4	32.4
120621	6.3	52.0

Data needs to be structured to be understood and used.

**A schema describes the structure of data.**

Its schema

For example, a schema describes what information is contained in the columns of a dataset.

More complete and well documented schemas increase the usability and value of the data.

Date	Depth	conc
110521	12.2	56.8
120521	10.8	12.6
110521	12.6	NA
110621	9.8	55
110521	8.4	32.4
120621	6.3	52.0

The data

Its schema

# Benefits of creating better schemas

- Helping your present self, your future self, and your collaborators
  - Avoid ‘mystery’ data with better descriptions
  - Tune the detail depending on the need
  - Deposit better quality data with less work
- Help others use your data
  - Spend less time supporting other people who are using your data
  - RTFS (read the fine schema)
  - Communicate data context better
  - Especially valuable in cross-disciplinary research
- Help machines find and use your data
  - Schemas can be machine readable
- Publish schemas for better collaboration and interoperability
  - Publish the schema with a separate DOI = others can cite and use
- Better science from better data

# Semantic Engine for better schemas



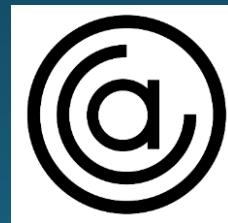
Partnership



Adopting and adapting  
overlays capture  
architecture (OCA)

Developed overlays capture  
architecture (OCA)

OCA is an international and open  
standard for data schemas

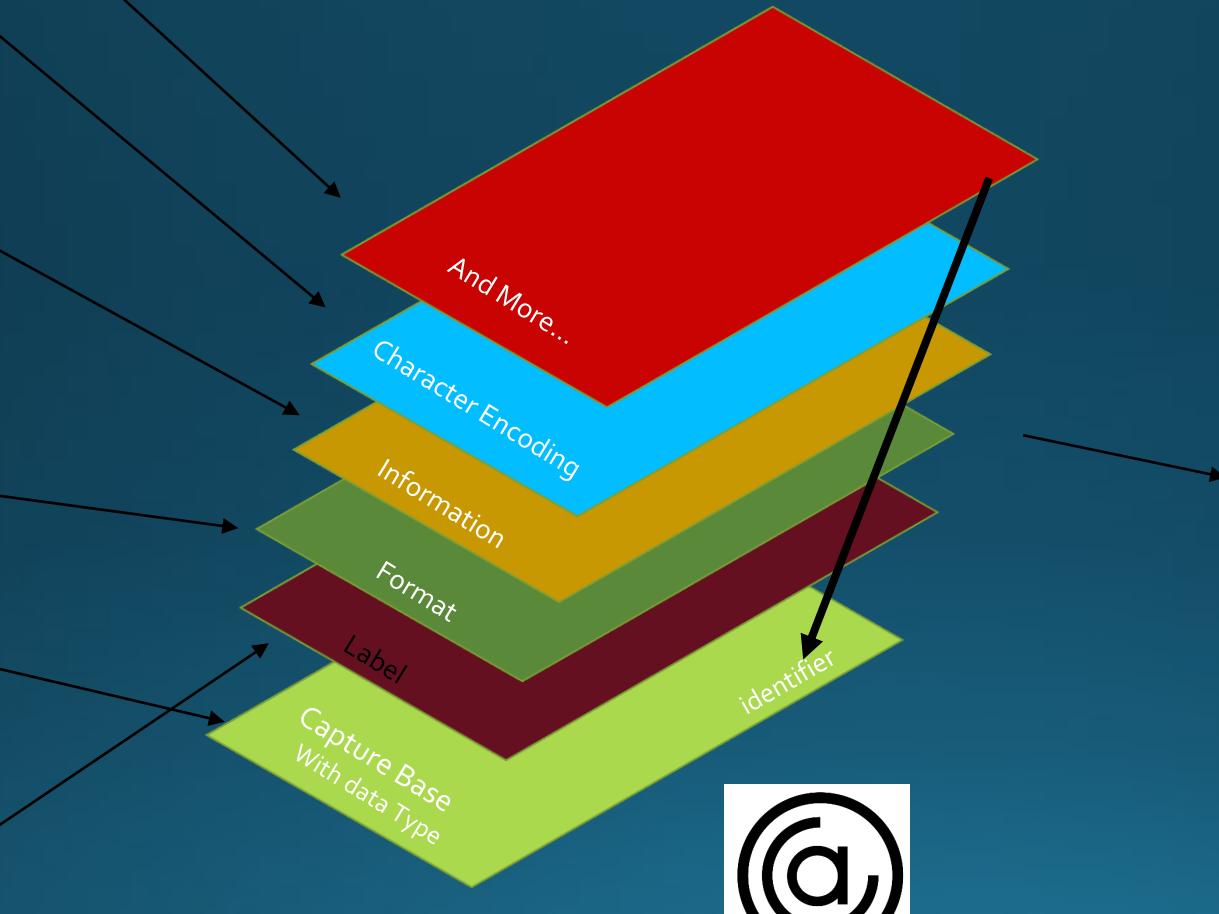


Overlays capture architecture  
for rich data schemas

# OCA recognizes a schema is made of different related but independent features

Label	Data Type	Format	Information	Character encoding	And more...
Date	Datetime	DDMMYY	This represents the date of sample collection in the field. The year codes are all in the 21 <sup>st</sup> century.	UTF-8	
Depth	Numeric	Double	The sample depth is the depth (in cm) of the sample hole (top of sample collection) before an additional ~5 cm of dirt was excavated and mixed and a sub-sample was removed.	UTF-8	
Conc	Numeric	Double	Carbon concentration in uM was calculated according to methods in DOI:example.	UTF-8	

Label	Data Type	Format	Information	Character encoding	And more...
Date	Datetime	DD/MM/YY	This represents the date of sample collection in the field. The year codes are all in the 21 <sup>st</sup> century.	UTF-8	
Depth	Numeric	Double	The sample depth is the depth (in cm) of the sample hole (top of sample collection) before an additional ~5 cm of dirt was excavated and mixed and a sub-sample was removed.	UTF-8	
Conc	Numeric	Double	Carbon concentration in µM was calculated according to methods in DOI:example.	UTF-8	



# OCA expresses the schema in a series of distinct overlays

OCA is a machine-readable format

```
{
  "@context": "https://oca.tech/overlays/v1",
  "type": "spec/overlay/label/1.0",
  "issued_by": "",
  "role": "",
  "purpose": "",
  "schema_base": "hl:fqSVkf4H6JKuEgu6AUUkD3iYzcBofhKPiuNBTxoL9Dw1",
  "language": "en_US",
  "attr_labels": {
    "BIRTHDT": "Date/Time of Birth",
    "AGE": "Age",
    "AGEU": "Age Units",
    "SEX": "Sex",
    "RACE": "Race",
    "ETHNIC": "Ethnicity"
  },
  "attr_categories": [
    "_cat-1_"
  ],
  "cat_labels": {
    "_cat-1_": ""
  },
  "cat_attributes": {
    "_cat-1_": [
      "BIRTHDT",
      "AGE",
      "AGEU",
      "SEX",
      "RACE",
      "ETHNIC"
    ]
  }
}
```

# Example overlays

*Label*

A human-readable label for the data field of the data table.

*Unit*

What specific units are used for each specific data field in your data table, e.g. uM.



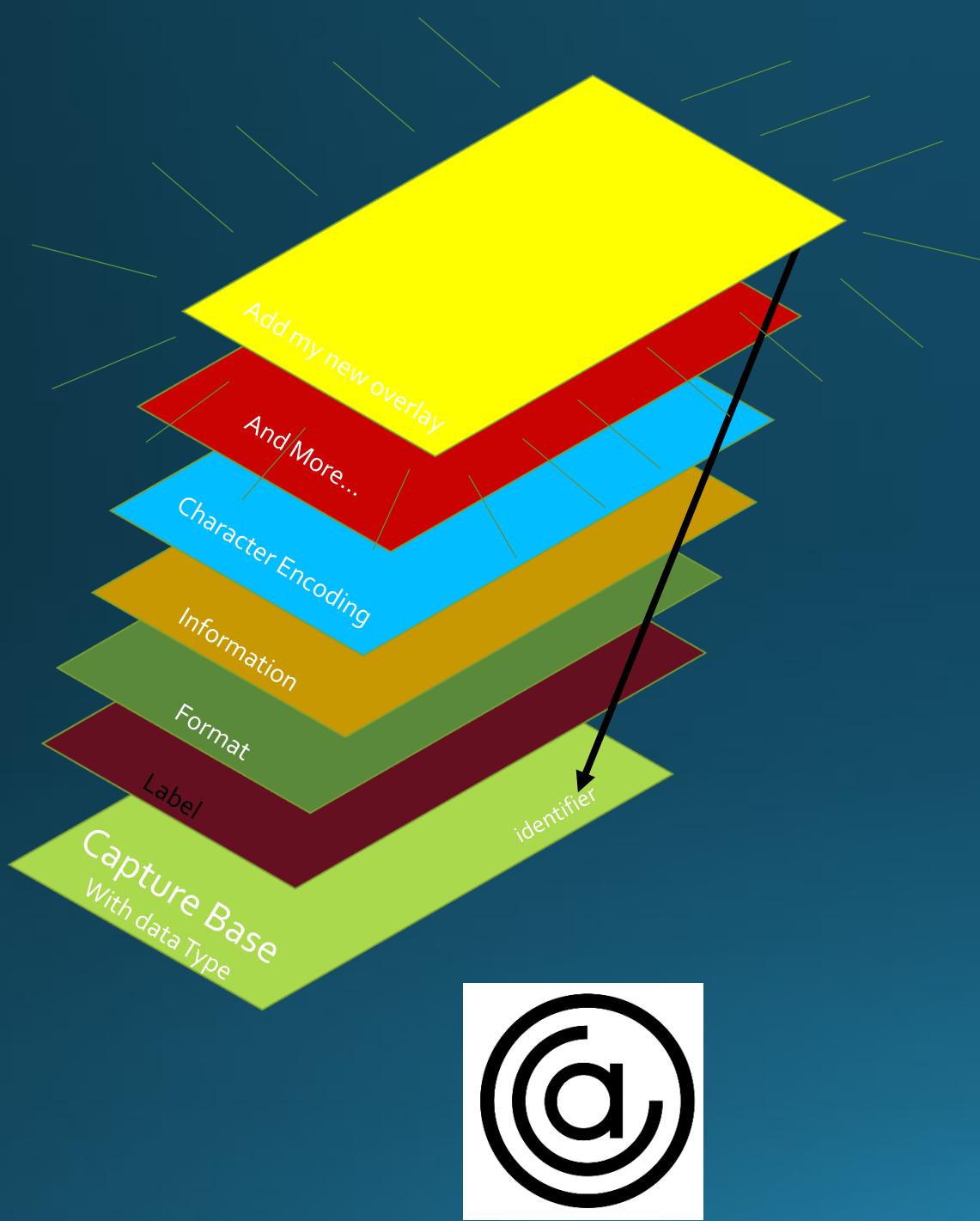
*Information*

Text that describes the data in a specific data field, e.g. what protocol was used to measure

*Entry*

Restrict entry of a specific data field to a list you provide, e.g. select from a list of three specific field locations to avoid confusing names.

# Benefits of a layered schema architecture



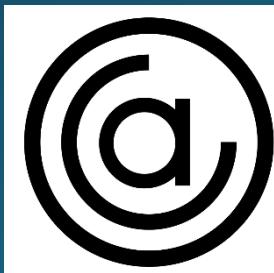
## Interoperability, extensibility

Each overlay is independent and references the schema base unique identifier.

You can extend and improve other people's schemas to fit your needs.

Add layers – e.g. add a new language or a data transformation when merging datasets using two different schemas.

Extend the schema while keeping the schema base which keeps your data **interoperable**.



# JSON Output

Name	Type
E_-EHd3f_BYYZMcJdn5lg1opqi8DNcYVsPAF9twK3kAo.json	JSON File
E3YrRA50HeBj34V_VblXJnHBU1vAxYCxIAxZWW9wfbkl.json	JSON File
EA0-kx2Mmc6jBooCmZFfJpwVgXPEUnVsK4FUVo-H0vLY.json	JSON File
Eae4FCcmF820ZNxwdAPexp7U26NlnmJj3bD08zkvP8HE.json	JSON File
EaO6295i-N3FI3kuYtMChk4EkK3nb792sLB0DlcWYoo.json	JSON File
Edlzt94ovFazdzX4vKxkRyNSSnAUTy5-7ZfknuVwLU.json	JSON File
EeFv1ODqi3mvo_jAseg0PbMpvJe6j30CUAed_-R0eUsq.json	JSON File
EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8.json	JSON File
EFU5APo-Ply2FR3nsu7AxbRyNhhpytJTX3N7_SlpKssM.json	JSON File
EIGY0etUWBBjbQhVlwzs1vkWIQ3yoL6egYmBTEiFsE44.json	JSON File
Ej2tha7fyvhB9r4h3LMlwHyjEFmCYpgkwU0aBm8pH-9A.json	JSON File
ELH2JnH_joaJqrCq9_m6g5iY29Vzkpw81usdCMITzrGI.json	JSON File
EQCWrzQmX1LDBW90VhUWx1icluAHgbtT6AAKxsy6F3HA.json	JSON File
ESJige6PuKfdK0YA8zupbTqP5PZM0emcpMvJP-vs6-EI.json	JSON File
EWPLQqdl3hYKoqLedO63gg94yKO1U1PGIFZIHnqumeMo.json	JSON File
EyoW90rtJwdwX7_j3NGHwkM3zL4LqVRRGZ0Nbpbkb_uc.json	JSON File
meta.json	JSON File

The JSON uses lots of SAIDs...

```
{  
  "capture_base": "EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8",  
  "type": "spec/overlays/entry_code/1.0",  
  "attribute_entry_codes": {  
    "insectType": [  
      "501",  
      "527"  
    ],  
    "location": [  
      "BAFF",  
      "TH"  
    ]  
  },  
  "files": {  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] character_encoding": "EyoW90rtJwdwX7_j3NGHwkM3zL4LqVRRGZ0Nbpbkb_uc",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] conditional": "EA0-kx2Mmc6jBooCmZFfJpwVgXPEUnVsK4FUVo-H0vLY",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] conformance": "E3YrRA50HeBj34V_VblXJnHBU1vAxYCxIAxZWW9wfbkl",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] credential_layout": "Eae4FCcmF820ZNxwdAPexp7U26NlnmJj3bD08zkvP8HE",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] entry (en-CA)": "E_-EHd3f_BYYZMcJdn5lg1opqi8DNcYVsPAF9twK3kAo",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] entry (fr-CA)": "EFU5APo-Ply2FR3nsu7AxbRyNhhpytJTX3N7_SlpKssM",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] entry_code": "EeFv1ODqi3mvo_jAseg0PbMpvJe6j30CUAed_-R0eUsq",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] form_layout": "EWPLQqdI3hYKoqLedO63gg94yKO1U1PGIFZIHnqumeMo",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] format": "Ej2tha7fyvhB9r4h3LMlwHyjEFmCYpgkwU0aBm8pH-9A",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] information (en-CA)": "Edlzt94ovFazdzX4vKxkRyNSSnAUTy5-7ZfknuVwLU",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] information (fr-CA)": "ESJige6PuKfdK0YA8zupbTqP5PZM0emcpMvJP-vs6-EI",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] label (en-CA)": "ELH2JnH_joaJqrCq9_m6g5iY29Vzkpw81usdCMITzrGI",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] label (fr-CA)": "EQCWrzQmX1LDBW90VhUWx1icluAHgbtT6AAKxsy6F3HA",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] meta (en-CA)": "EaO6295i-N3FI3kuYtMChk4EkK3nb792sLB0DlcWYoo",  
    "[EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8] meta (fr-CA)": "EIGY0etUWBBjbQhVlwzs1vkWIQ3yoL6egYmBTEiFsE44",  
    "capture_base-0": "EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8"  
  "root": "EFdwDXYtpNhTHRkmPZECuNMG9WqZL3wJJyJ01okrXpc8"  
}
```

# SAIDs for Identifiers

- If you find a resource (like a dataset) - could you find its identifier?
  - Can you tell if any data has been changed?
- Self-Addressing Identifiers (SAID)
  - Digital fingerprints of the file
  - If you change the file *in any way*, the fingerprint changes.
- If you find an object, you can calculate its identifier
  - Don't need to trust that you received it from the authorized source

# Hashing— creating digital fingerprints

- Hash functions
  - One-way formula
  - You can't take a hash and calculate the original data
    - Good for privacy of the data
- The SAID identifier is calculated with a short algorithm
  - Calculate the hash of the data/document
  - Insert the hash into the original data/document

# Try out schema writing yourself

- [https://agrifooddatacanada.github.io/OCA\\_training\\_pathway/](https://agrifooddatacanada.github.io/OCA_training_pathway/)
- Preview to our June 2<sup>nd</sup> launch announcement
- Workflow:
  - Fill out the Excel Template File
  - Create your OCA Schema Bundle
  - Save the Excel Template and Schema Bundle with your dataset
  - Save your schema in the UoG Dataverse repository to give it a DOI

## OCA\_training\_pathway

---

### A pathway for creating data schemas

---

Schemas are a way to document your data and help make it more FAIR (Findable, Accessible, Interoperable, Reusable). Creating a schema is a process of continuous improvement. You don't need to create the most perfect and complete schema at the beginning. Instead, follow this pathway to gradual improvement where each step produces something usable for researchers.

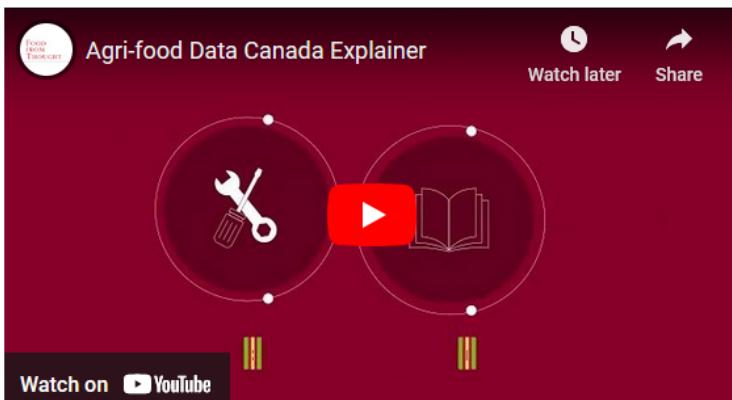
What are some benefits:

- Increase value and usability of data by adding context
- Enable early metadata creation when it is easier
- Easier to use standards with metadata
- Machine readable to reduce the number of times the same information must be manually inputted
- Easy to reuse existing metadata to generate new metadata
- Easier to deposit data into archive repositories

[Feedback for creating a schema can be done in this Form.](#)

#### 1. Introduction to the rationale

---



# Next up

- Introduction to R – Wednesday May 10
- Documenting your data and processes with R Markdown – May 17
- R Shiny – May 31
- Introduction to GitHub – June 14
- Introduction to Linux – June 21

<https://agrifooddatacanada.github.io/Research-Data-Workshop-Series/>