

# A data ecosystem serving agri-food sustainability

Lucas Alcantara, Ph.D.  
Manager, Research Centre Data



**AGRI-FOOD DATA  
CANADA**

AT THE UNIVERSITY *of* GUELPH

# The Problem

## Agri-food research data ecosystem challenges

- Data sharing and discovery is hard, data reuse is minimal
- **Integration and reuse** of data from different sources or research groups **can make any dataset more valuable**

# Agri-food Data Canada's Vision

To be **researcher-centered**, providing reliable **data management** and **analytics ecosystems** that fuel innovation and enable broad access to world-leading, curated research data that promotes opportunities for **innovation and partnerships**.

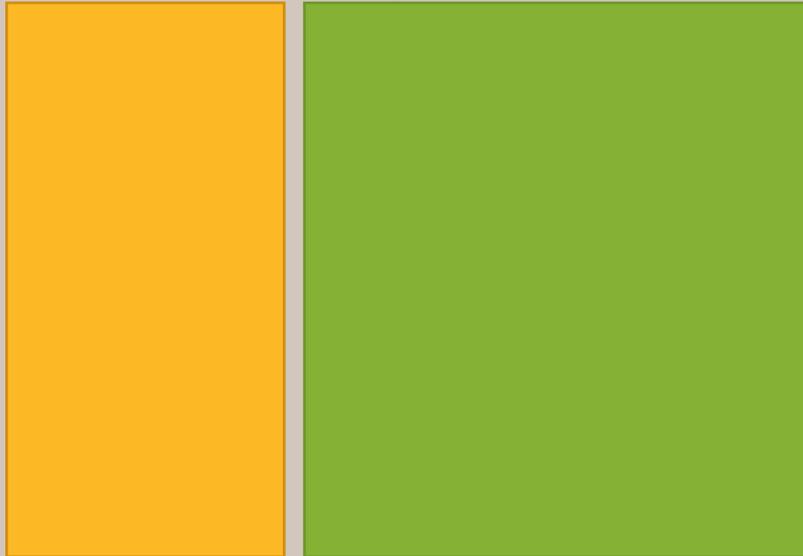
“Making agri-food data FAIR”

*Findable, Accessible, Interoperable, Reusable*



Governance

Ecosystem



# ADC Design

## Principles

- Open source, open development ideals
- Modular design
- Automate processes
- Integrate PIDs
- Researcher centric
- Connect data to metadata as priority
- Enter (meta)data once
- Knowledge accessible at time of need

# ADC Ecosystem

## Goal

- Agri-food Data Canada's vision is a research data ecosystem
- The ecosystem is not a platform to hold research data
- The ecosystem seeks to help researchers create more FAIR data

*Findable, Accessible, Interoperable, Reusable*



## Governance



## Ecosystem



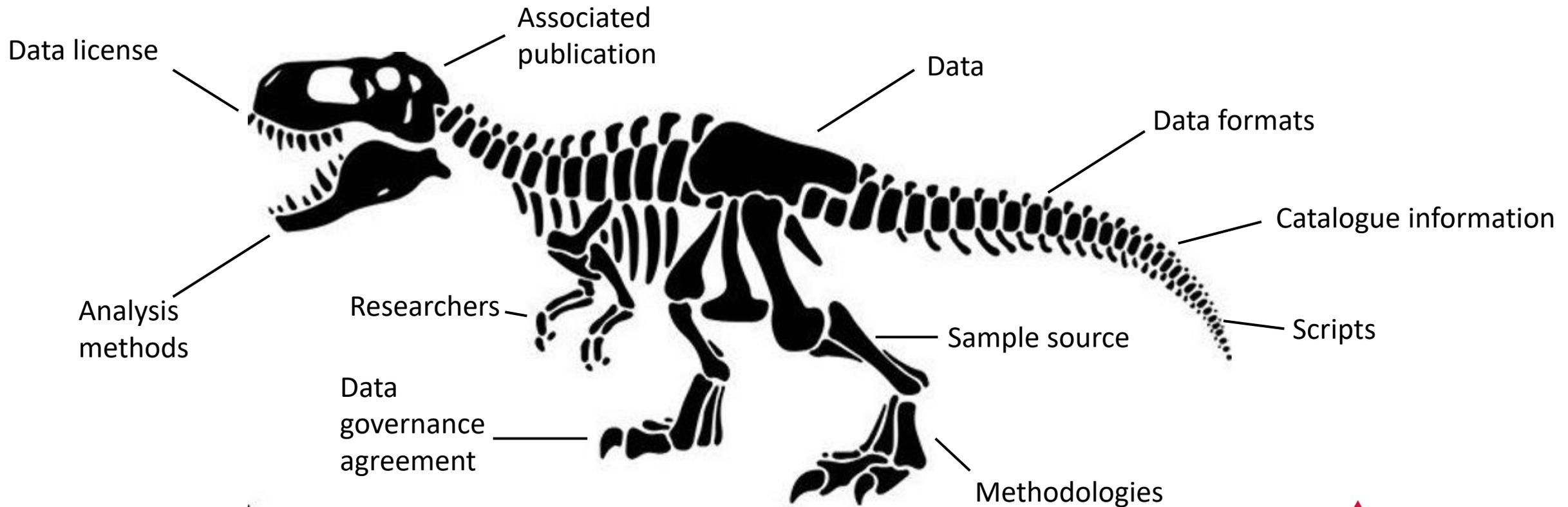
# Data and its context

Data does not speak for itself



# Data and its context

Data does not speak for itself; it is more useful when placed in context



# Increasing the value of data

Adding context increases the value of data

Data with context is more reusable

Data that is reused has greater value

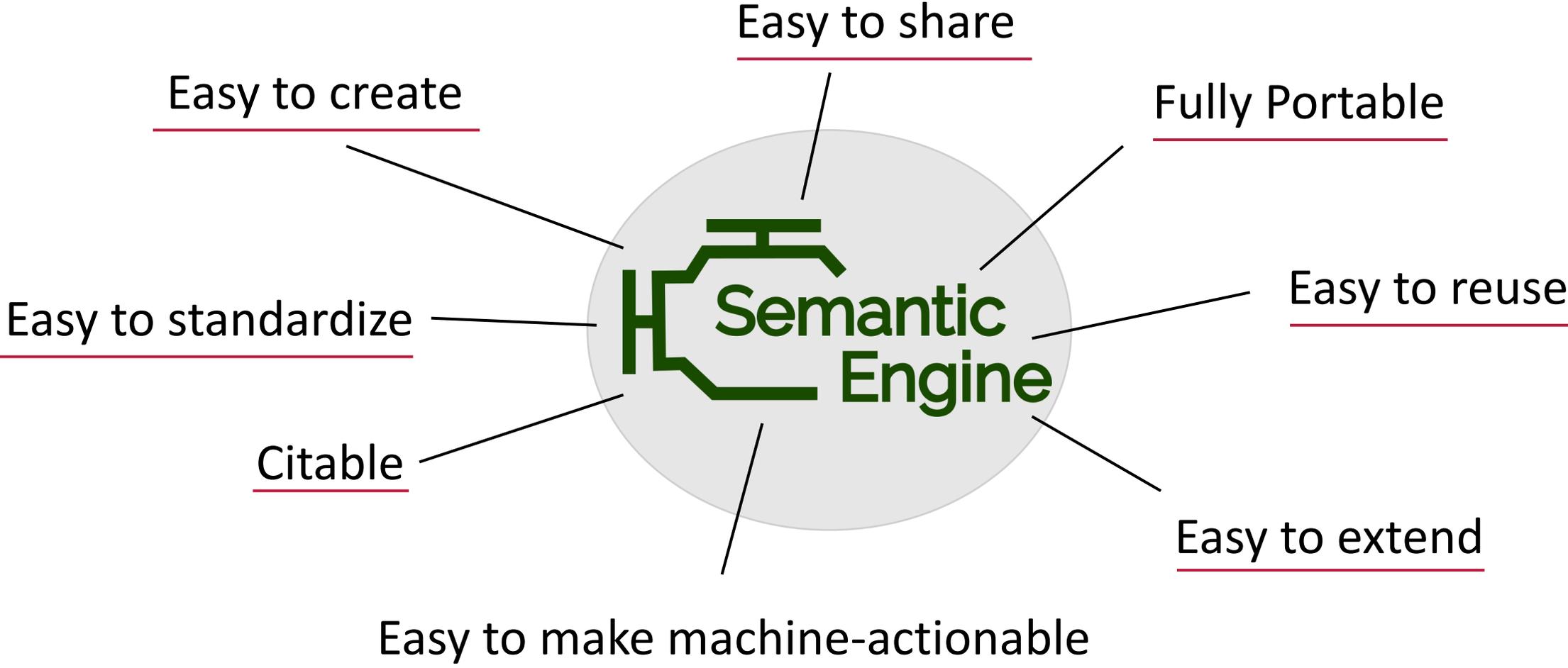
# The Semantic Engine for context

ADC is developing a suite of tools called the Semantic Engine



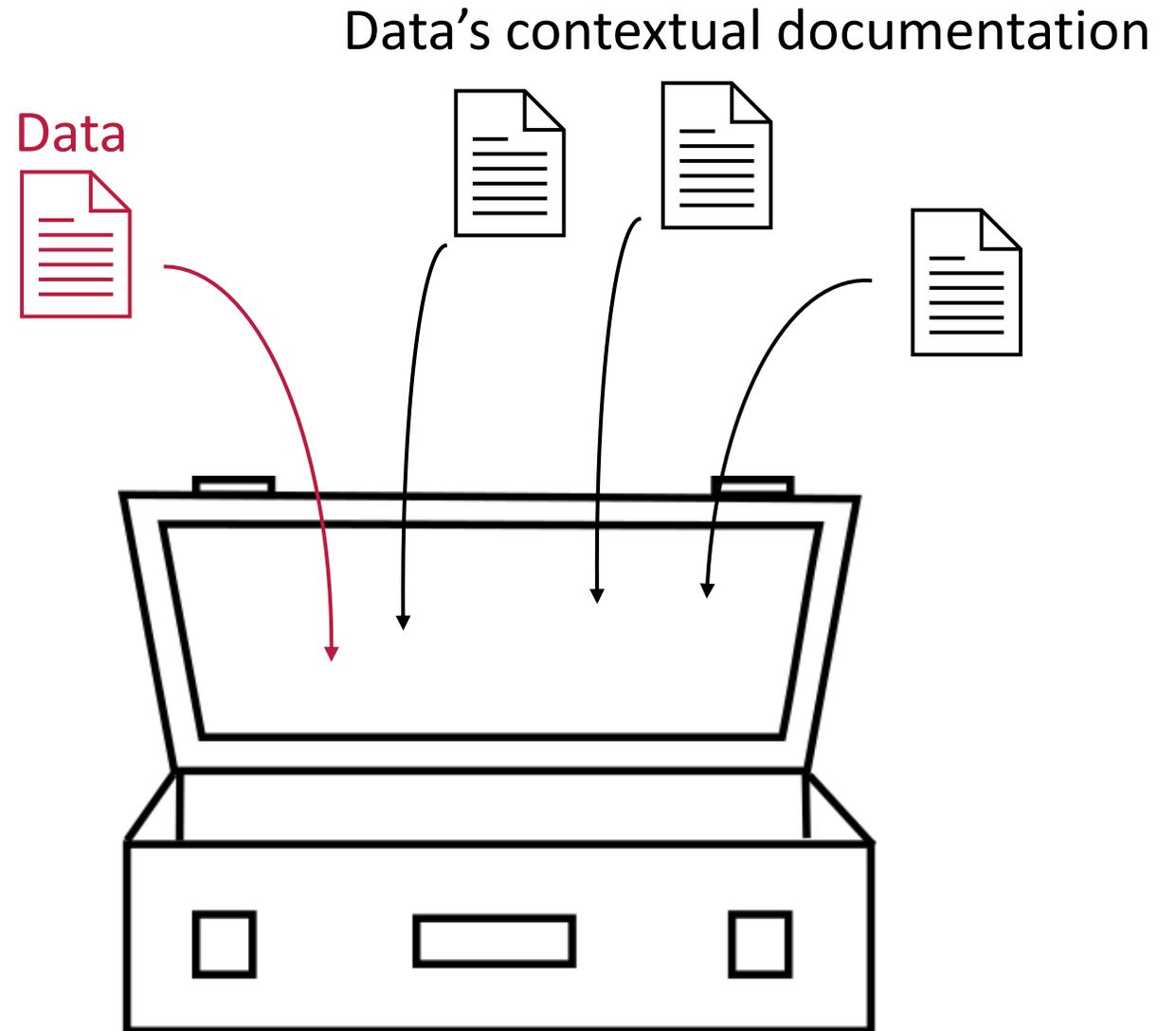
The Semantic Engine helps researchers write rich contextual data documentation

# Adding context through documentation



# Context should travel with data

- Portable
- Machine-actionable
- **Citable**



# Data Schemas

The first contextual documentation tool from ADC focuses on writing rich data schemas



# Semantic Engine for better schemas

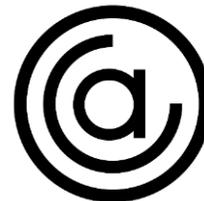


HUMAN COLOSSUS  
FOUNDATION

Adopting and adapting  
OCA for agri-food data



Developed OCA, an international and  
open standard for data schemas



Overlays Capture Architecture (OCA)  
for rich data schemas

# Data documentation: Schemas

All datasets have a schema

- Explicit or implicit
- Contains useful details or ‘user-must-guess’

A schema describes the attributes (variables)

animal_id	duration	session_n	total_yield	milking_location
4551	00:05:39	3	9.03	Voluntary Milking System
4551	00:06:23	4	10.14	Voluntary Milking System
4604	00:05:12	1	14.83	Rotary Parlour
4598	00:06:41	1	28.63	Rotary Parlour

← Attributes

# Attributes are described in a schema

Dataset

animal_id	duration	session_n	total_yield	milking_location
4551	00:05:39	3	9.03	Voluntary Milking System
4551	00:06:23	4	10.14	Voluntary Milking System
4604	00:05:12	1	14.83	Rotary Parlour
4598	00:06:41	1	28.63	Rotary Parlour

Schema

Attribute	Label (en)	Description (en)	Units	Type
animal_id	Animal ID	Farm-level unique animal ID		Numeric
duration	Duration	Milking event Duration in minutes	min	DateTime
session_n	Session Number	Unique count of the milking event per cow, per day, per milking system. Resets at midnight.		Numeric
total_yield	Total Yield	Yield of milking event in litres	L	Numeric
milking_location	Milking Location	Location of where the specific milking event took place		Text

Attributes



# Task-specific features of a schema

OCA recognizes a schema is made of different related features, which are also independent (overlays)

Each column is a feature, and we can write more features as needed

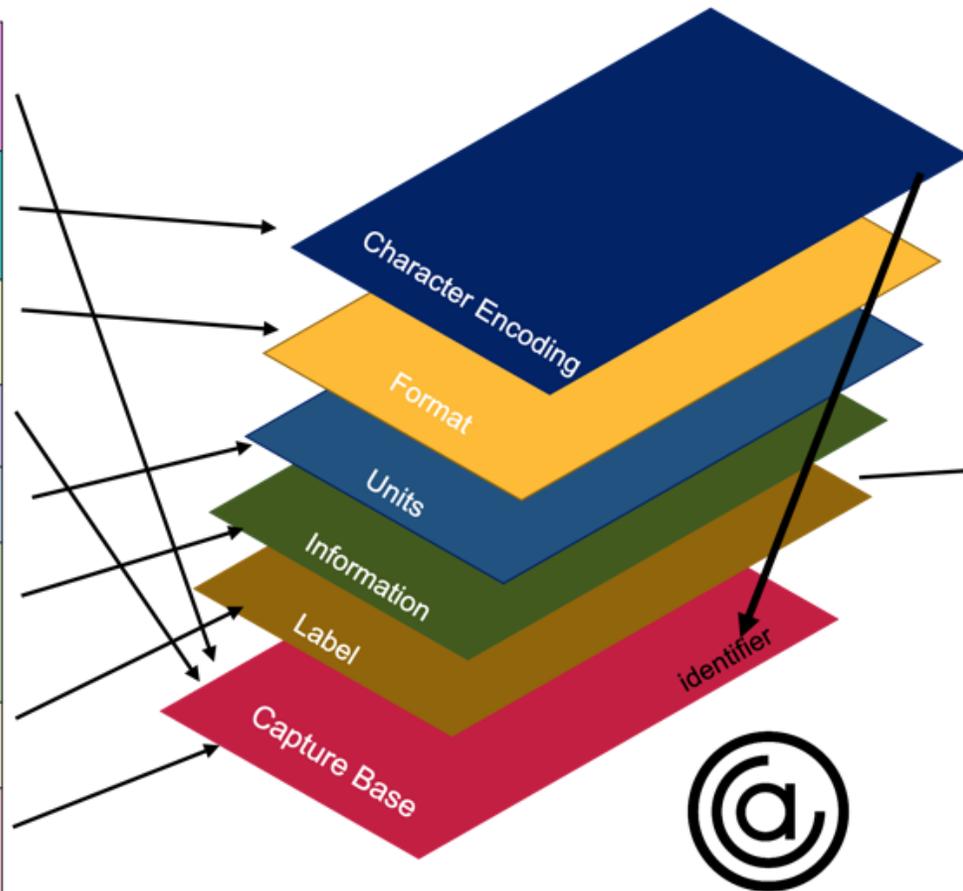
Attribute	Label (en)	Information (en)	Units	Type	Format	Character encoding	Flagged
animal_id	Animal ID	Farm-level unique animal ID		Numeric	^[1-9]\d*\$	UTF-8	Y
duration	Duration	Milking event Duration in minutes	min	DateTime	hh:mm:ss	UTF-8	
session_n	Session Number	Unique count of the milking event per cow, per day, per milking system. Resets at midnight.		Numeric	^[1-9]\d*\$	UTF-8	
total_yield	Total Yield	Yield of milking event in litres	L	Numeric	^\d*\.\?\d+\$	UTF-8	
milking_location	Milking Location	Location of where the specific milking event took place		Text	^\.{0,250}\$	UTF-8	

Etc...



# OCA layered structure

Attribute	Label (en)	Information (en)	Units	Type	Format	Character encoding	Flagged
animal_id	Animal ID	Farm-level unique animal ID		Numeric	^[1-9]d+\$	UTF-8	Y
duration	Duration	Milking event Duration in minutes	min	Date/Time	hh:mm:ss	UTF-8	
session_n	Session Number	Unique count of the milking event per cow, per day, per milking system. Resets at midnight.		Numeric	^[1-9]d+\$	UTF-8	
total_yield	Total Yield	Yield of milking event in litres	L	Numeric	^\d*\.\d+\$	UTF-8	
milking_location	Milking Location	Location of where the specific milking event took place		Text	^[0,250]s\$	UTF-8	



OCA files are machine-readable

```
{
  "capture_base": "EzB01m3oiJ7-tNwHipu-KSPdqvQpbiZTXFuu4CdcCQbA",
  "digest": "ECp6yLXgqvawIA7Wx4hJNn_I-taY3teQoh6ZL5So_h5A",
  "type": "spec/overlays/label/1.0",
  "language": "en",
  "attribute_labels": {
    "animal_id": "Animal ID",
    "duration": "Duration",
    "session_n": "Session Number",
    "total_yield": "Total Yield",
    "milking_location": "Milking Location"
  }
}
```

## Benefits of creating better schemas

- **Helps your present self, your future self, and your collaborators**
  - Avoid 'mystery' data with better descriptions
  - Deposit better quality data with less work
- **Helps others use your data**
  - Spend less time supporting other people who are using your data
  - 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**

# Benefits for ICAR members

- Facilitates data interoperability and harmonization by:
  - Enabling the incorporation of ontological terms and data standards endorsed by the committee
- Reduces barriers for the uptake of new sources and uses of data recording by:
  - Simplifying the process of documentation of data from new sources
  - Supporting automated pipelines for data validation

# THANK YOU

Agri-food Data Canada at the University of Guelph is an innovation platform for Canada's agriculture and food sectors.



**AGRI-FOOD DATA  
CANADA**

AT THE UNIVERSITY *of* GUELPH

50 Stone Road East, Guelph, ON N1G 2W  
adc@uoguelph.ca

519-824-4120 Ext. 53547

[agrifooddatacanada.ca](http://agrifooddatacanada.ca)