Decentralised Semantics: A Semantic Engine user perspective

SciDataCon 2023

organized by

WORLD DATA SYSTEM

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

as part of

IDW A FESTIVAL OF DATA 23-26 OCT 2023 SALZBURG

Carly Huitema, Ph.D.

Manager – Agri-food Data Strategy

ADC Ecosystem

Goal

- Agri-food Data Canada's vision is a research data ecosystem
  - Improving the connections between existing resources
  - Supporting data FAIRness with machine-readable context
- Agri-food Data Canada's current approach is **researcher centric**, helping them create FAIR data

Findable, Accessible, Interoperable, Reusable

AGRI-FOOD DATA CANADA
AT THE UNIVERSITY of GUELPH

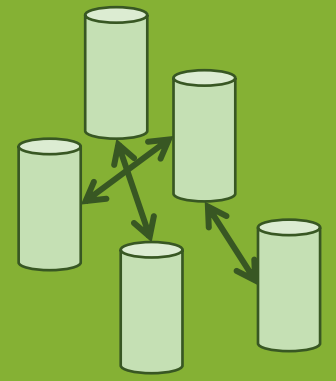
Governance

Working Groups

Ecosystem Principles and Governance

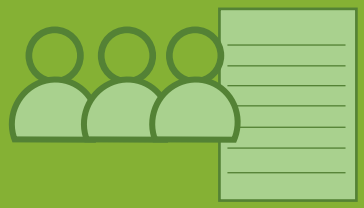
Ecosystem

Semantic Engine



Federate existing nodes

Ecosystem Scale Catalogues



Knowledge and training

The Semantic Engine for context

ADC is developing a suite of tools called the Semantic Engine

The Semantic Engine helps researchers write rich contextual and machine-readable data documentation

Data Schemas

Data schemas – the **first** contextual documentation target of ADC

Semantic Engine for better schemas

AGRI-FOOD DATA CANADA
AT THE UNIVERSITY of GUELPH

Partnership

HUMAN COLOSSUS FOUNDATION

Adopting and adapting overlays capture architecture (OCA)

Developed overlays capture architecture (OCA)

OCA is an international and open standard for data schemas

OCA Documentation:
DOI: 10.5281/zenodo.7707466

Overlays capture architecture for rich data schemas

ADC

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

Attributes

Schema

Attribute	Label (en)	Description (en)
animal_id	Animal ID	Farm-level unique animal ID
duration	Duration	Milking event Duration in minutes
session_n	Session Number	Count of the milking event for the cow. Counting resets at midnight
total_yield	Total Yield	Yield of milking event in litres
milking_location	Milking Location	Location of where the specific milking event took place

Layers of a schema

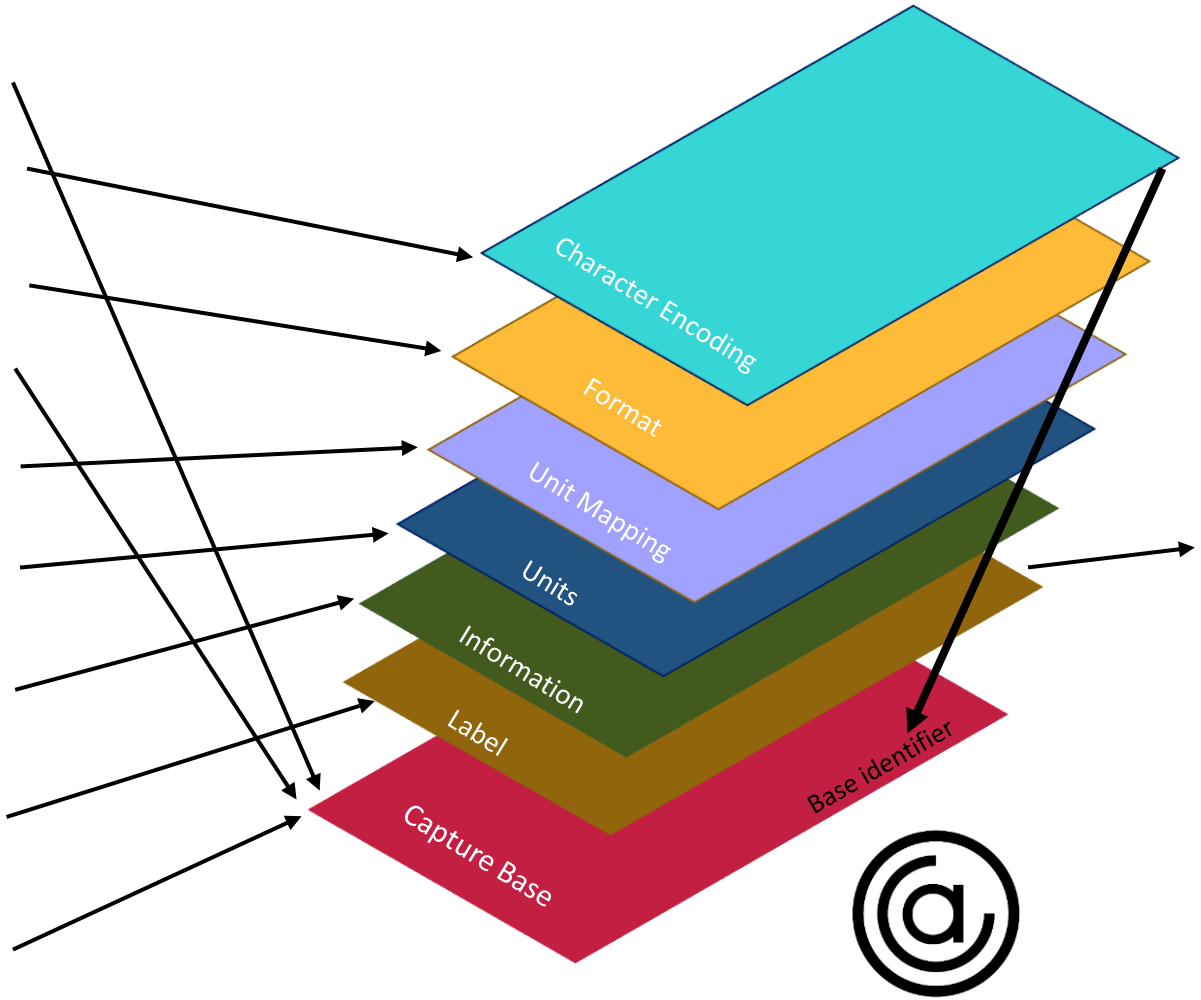
- Extend the schema with more features (details)
- Each feature is a layer of the schema in OCA
  - Advantage – can split schema layers into tasks with independent contributors

Schema

Attribute	Label (en)	Information (en)	Units	Unit mapping (UCUM)	Type	Format	Character encoding	Sensitive
animal_id	Animal ID	Farm-level unique animal ID			Numeric	^[1-9]\d*$	UTF-8	Y
duration	Duration	Milking event Duration in minutes	min	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	l	Numeric	^\d*\.?\d+$	UTF-8	
milking_location	Milking Location	Location of where the specific milking event took place			Text	^.{0,250}$	UTF-8	

OCA layer structure

Attribute	Label (en)	Information (en)	Units	Unit mapping (UCUM)	Type	Format	Character encoding	Sensitive
animal_id	Animal ID	Farm-level unique animal ID			Numeric	^[1-9]\d*$	UTF-8	Y
duration	Duration	Milking event Duration in minutes	min	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	l	Numeric	^\d*\.?\d+$	UTF-8	
milking_location	Milking Location	Location of where the specific milking event took place			Text	^.{0,250}$	UTF-8	

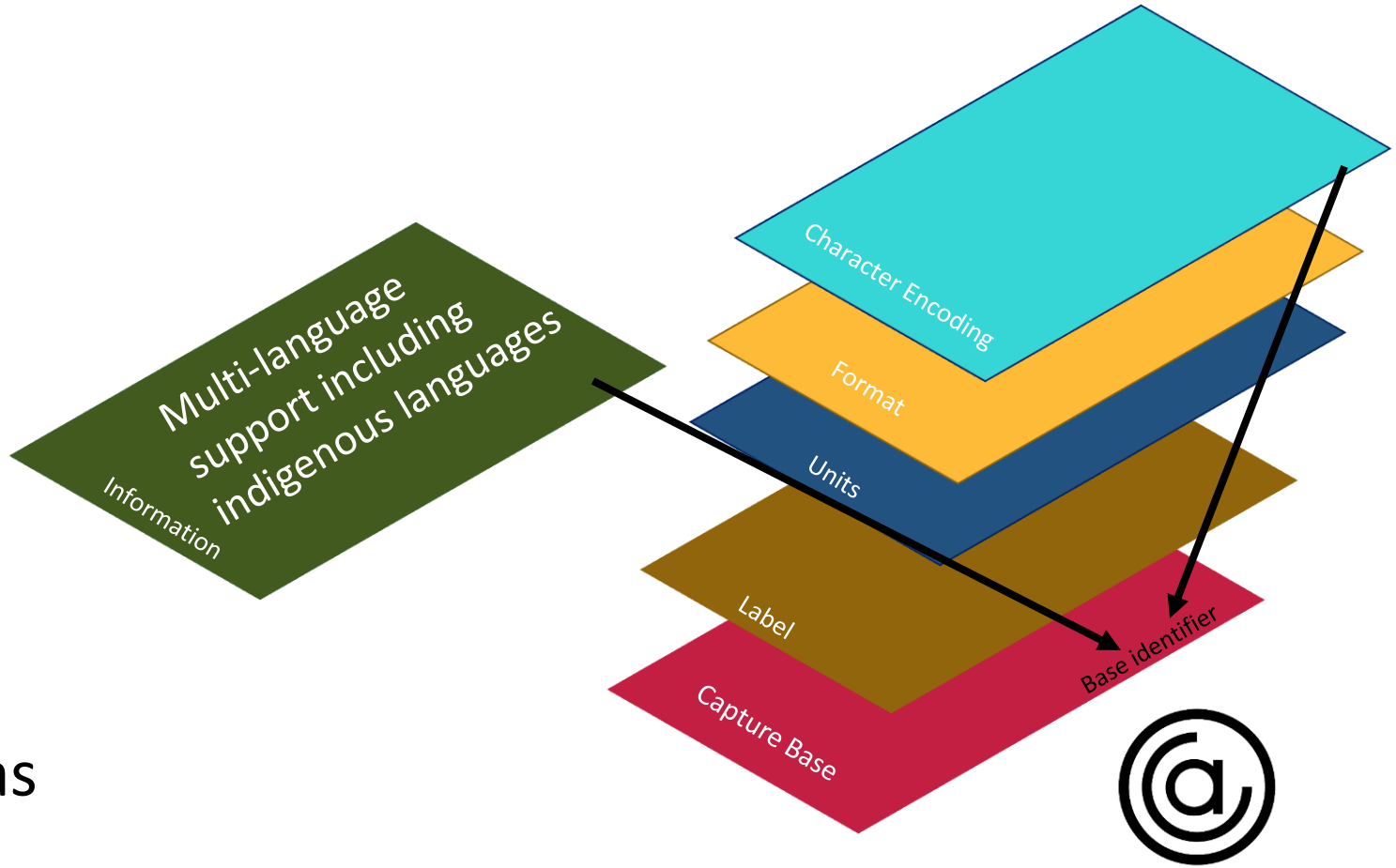


OCA is a machine-readable format

```
{
  "capture_base": "EzBO1m3oiJ7-tNwHipu-KSPdqvQpbiZTXFuu4CdcCQbA",
  "digest": "EF4w6K-SMVZ4VFxkc-jqNNGUXKbO587o3MtcpBU_zeko",
  "type": "spec/overlays/label/1.0",
  "language": "po",
  "attribute_labels": {
    "animal_id": "ID do animal",
    "begin_time": "Hora de início",
    "date": "Data de início",
    "dim": "Dias em leite",
    "duration": "Duração",
    "end_date": "Data de término",
    "end_time": "Hora de término",
    "lact_n": "Número da lactação",
    "milking_location": "Local de ordenha",
    "session_n": "Número da sessão",
    "total_yield": "Produção total"
  },
  "attribute_categories": [

  ],
  "category_labels": {

  }
}
```

Identifiers glue the schema together

- Identifiers enable a decentralized schema architecture
- OCA supports multiple contributors for a schema
  - OCA uses self-addressing identifiers to link and track content
- Enables community improvement of OCA schemas

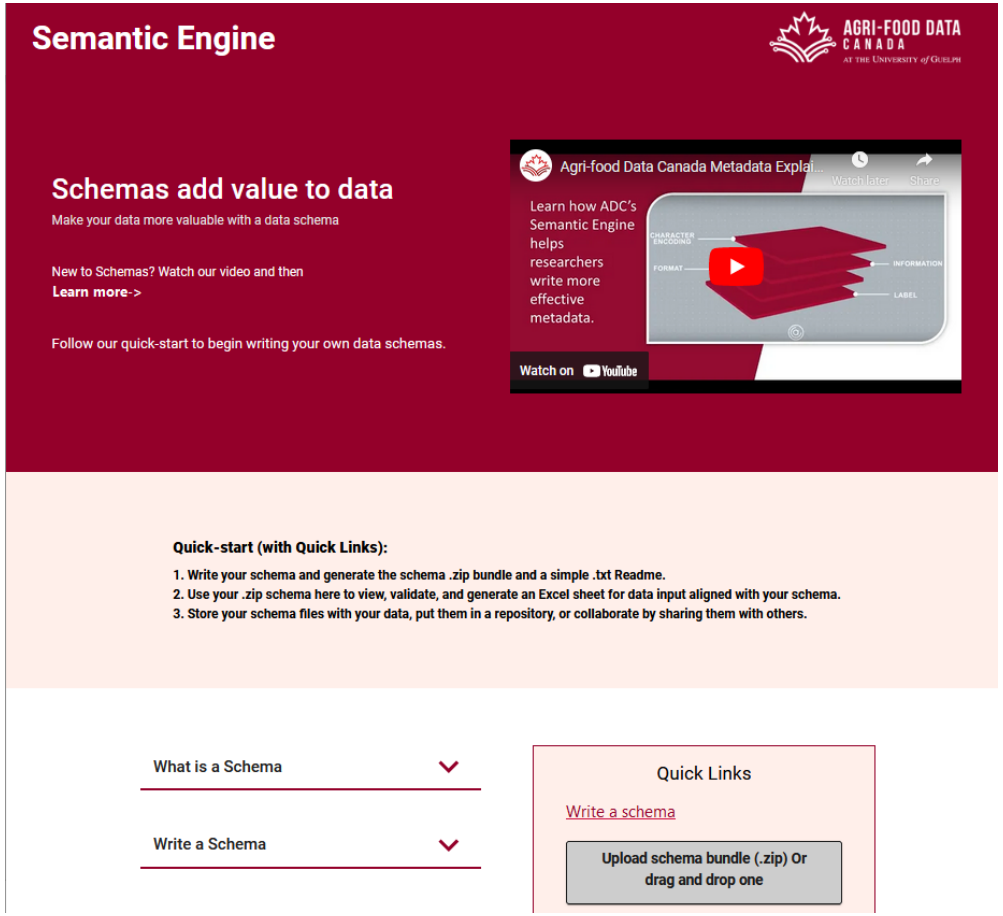


Who is using OCA

- Research projects at the University of Guelph
- University of Guelph research barns and stations
- Government of British Columbia using OCA for multi-language credential presentation
  - Gaining interest in other provinces, early pilots with Quebec
- EU project – cardiovascular consortium of 21 organizations (including universities and institutes) across Europe, CH, UK, and USA.
- IOT platform in supply chain - monitoring and telemetry data (multinational organization)

How anyone can use OCA schemas

semanticengine.org



General flow

1. Write your schema

2. Download the .zip (for machines) and the .txt (for humans)

Save and use schema directly

3. Publish schema with DOI

Deposit (.zip and .txt) schema in a repository for citation and reference

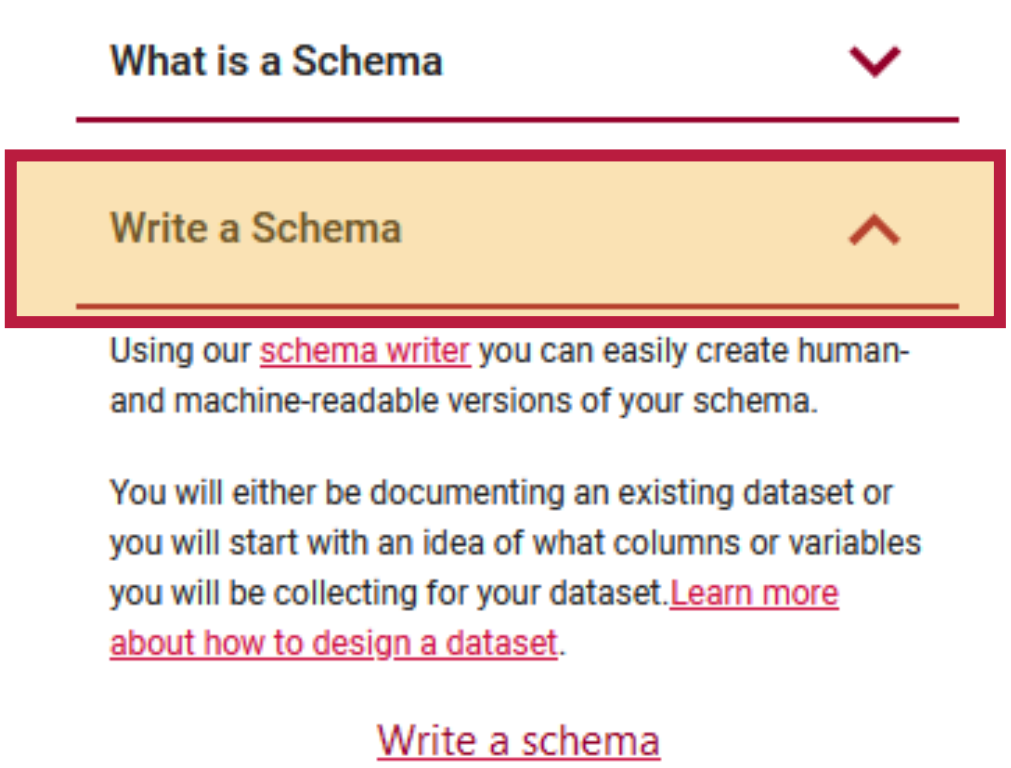
Example usage

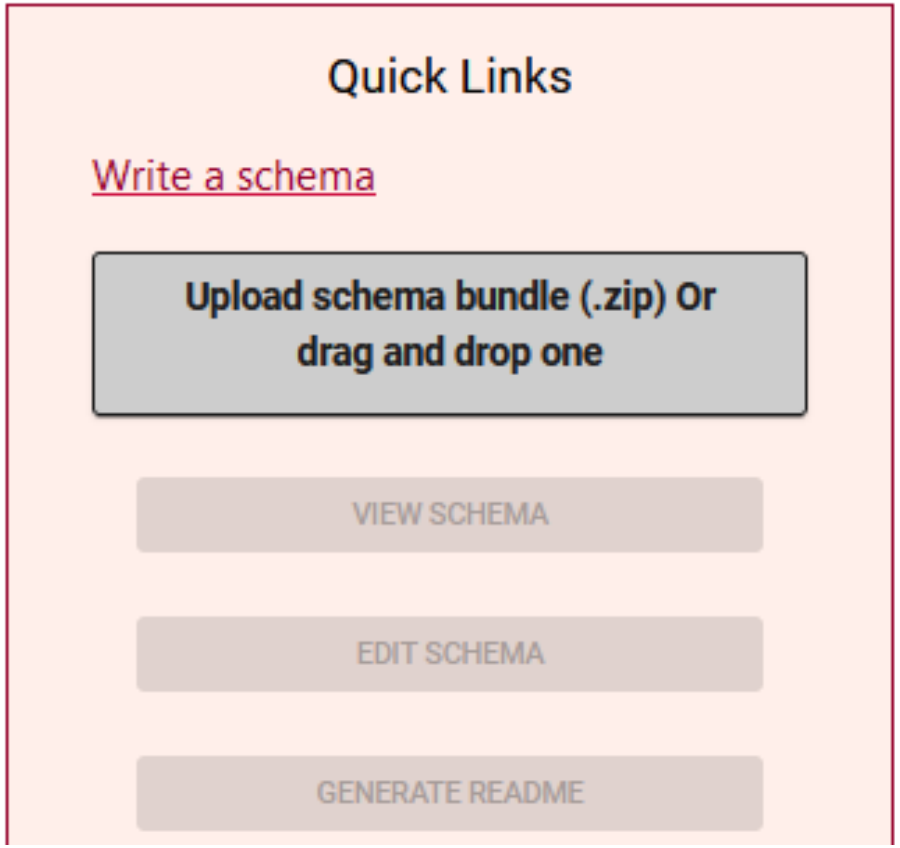
- Beef carcass data from the Meat Lab at the University of Guelph
  - All research farm products are sent to the campus meat lab
  - Meat lab is a small-scale meat packing plant operating under federal inspection regulations
    - Meat lab grades animals and collects data
    - Information on beef carcasses (multidimensional)
  - Schema documentation to help researchers use the data collected

Raw data collection

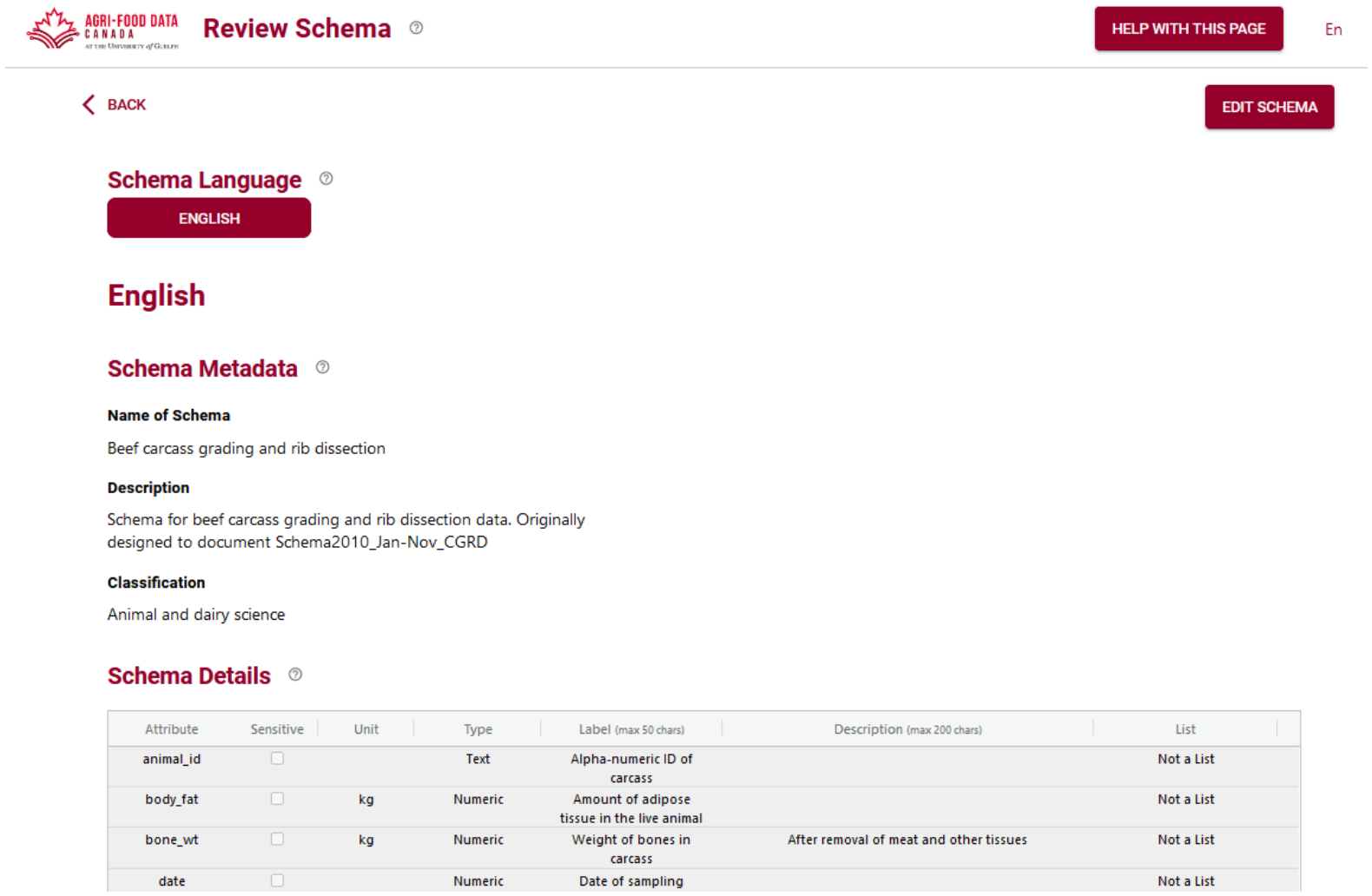
	A	B	C	D	E	F	G	H	I	J
1	date	animal_id	hotcarcass_wt	rib_fat1	rib_fat2	rib_fat3	grade_fat	ribeyearea_sqin	ribeyearea_sqmm	marblin
2	19970401	46F	244	9	12	10	8	9.2	.	4.5
3	19970421	82F	323	17	7	9	9	12.1	.	5
4	19970429	35F	310	17	14	10	10	10.5	.	6
5	19970429	44F	278	11	18	8	7	9.1	.	5.5
6	19970429	35F	291	17	15	8	8	11.9	.	3
7	19970506	OCA 54F	324	22	10	10	10	12.5	.	5.5
8	19970506	7258	318	28	12	12	11	10.7	.	4.5
9	19970513	7240	316	16	16	11	11	11.6	.	4.5
10	19970513	OCA 21F	291	27	10	8	8	11.5	.	5.5
11	19970513	86F	347	11	12	8	8	12.1	.	3.5
12	19970521	OCA 3F	381	21	10	9	9	14.3	.	4.5
13	19970521	11F	373	8	17	8	8	13.6	.	4.5
14	19970521	F105	326	11	20	10	10	10.6	.	4.5
15	19970521	7259	313	25	8	9	9	9.3	.	4.5
16	19970527	119F	364	7	19	9	9	12	.	4.5
17	19970527	F78	280	10	18	11	9	10.5	.	4
18	19970527	23F	353	10	24	12	12	11.6	.	5.5
19	19970527	9F	373	10	25	9	9	12.2	.	5
20	19970603	OCA 18F	356	8	16	9	9	13	.	6
21	19970603	F54	339	9	21	11	10	11.2	.	5.5
22	19970603	28F	372	8	15	8	8	11.8	.	5
23	19970603	OCA 51F	294	19	15	11	11	10.9	.	5
24	19970610	93F	309	25	13	12	10	11.8	.	5

Author Schema using Semantic Engine





Author Schema using Semantic Engine



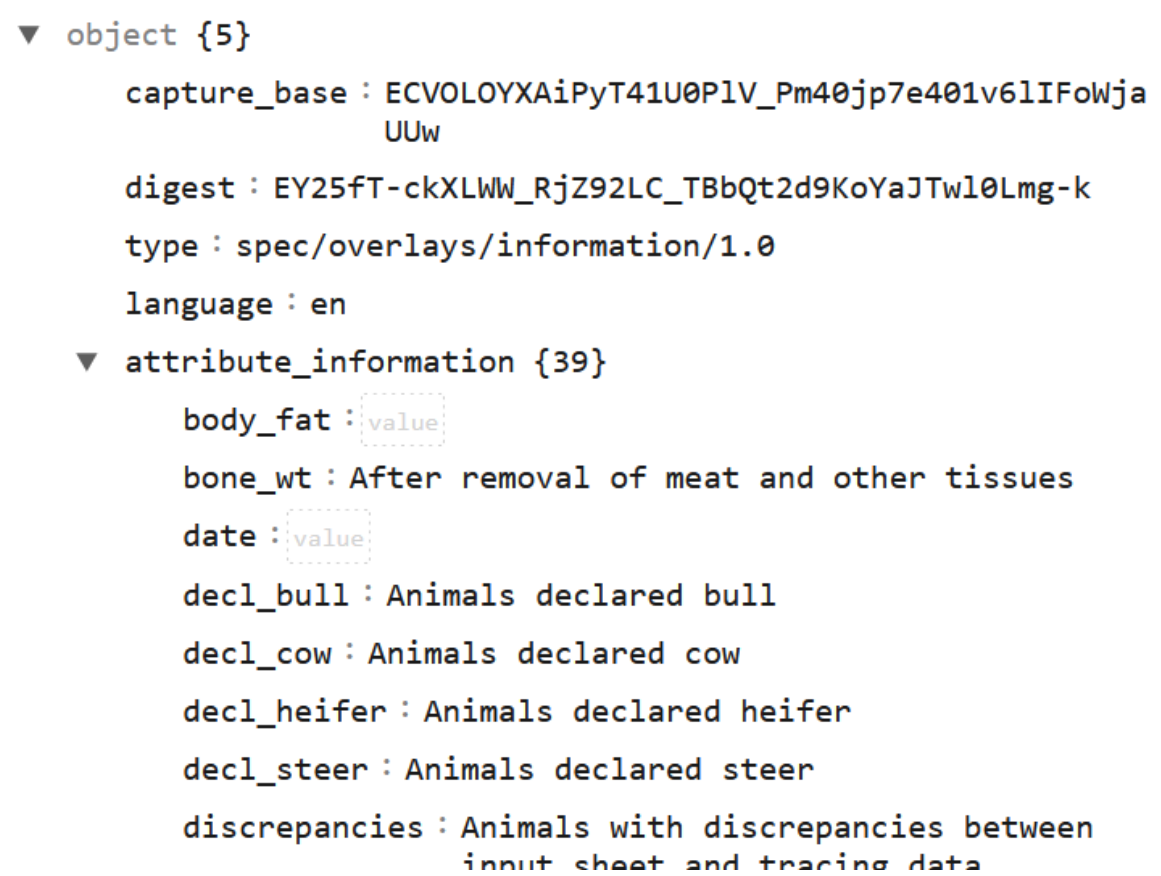
Attribute	Sensitive	Unit	Type	Label (max 50 chars)	Description (max 200 chars)	List
animal_id	☐		Text	Alpha-numeric ID of carcass		Not a List
body_fat	☐	kg	Numeric	Amount of adipose tissue in the live animal		Not a List
bone_wt	☐	kg	Numeric	Weight of bones in carcass	After removal of meat and other tissues	Not a List
date	☐		Numeric	Date of sampling		Not a List

Schema in .txt and .zip (JSON) formats

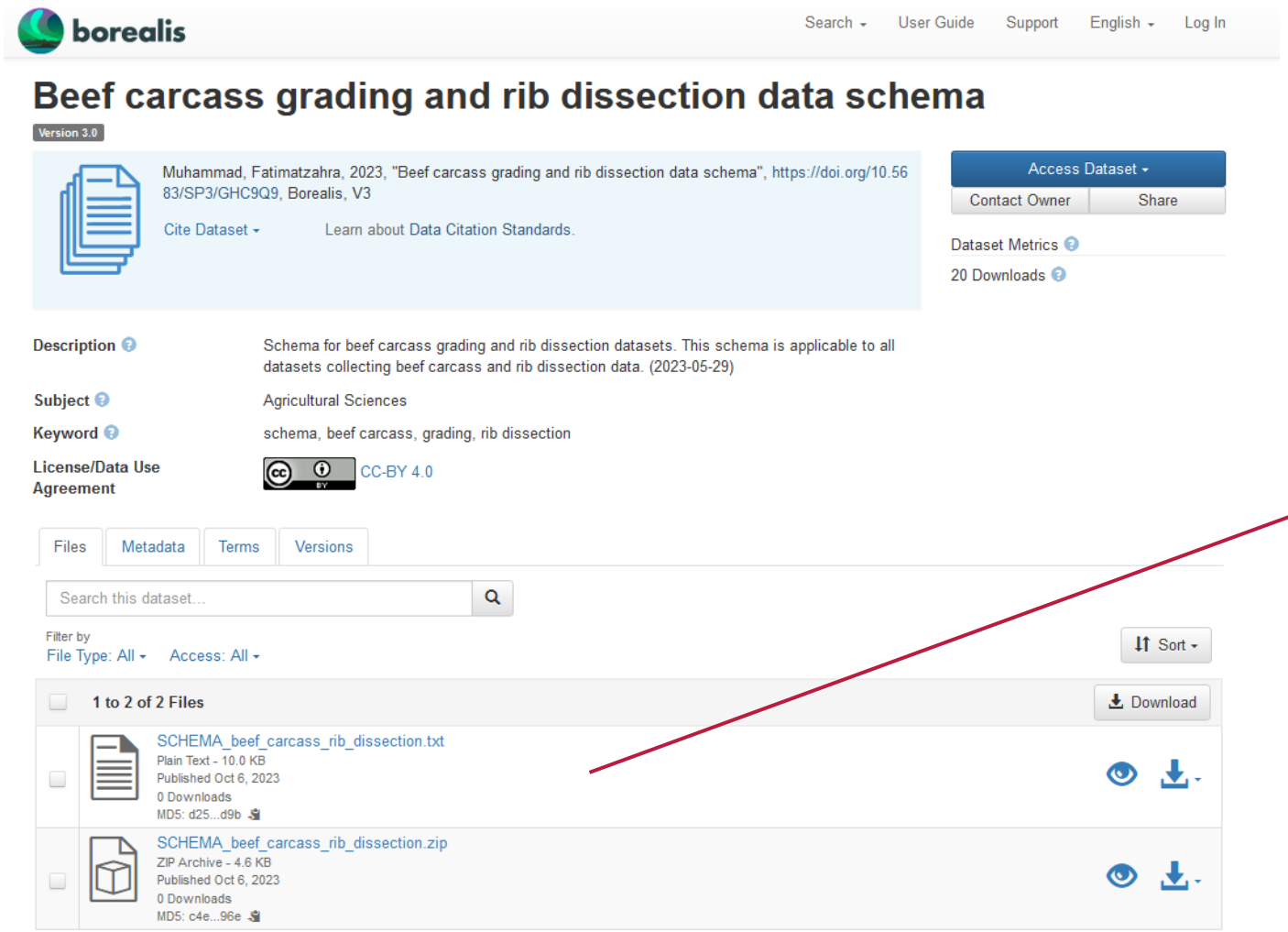
Readme.txt

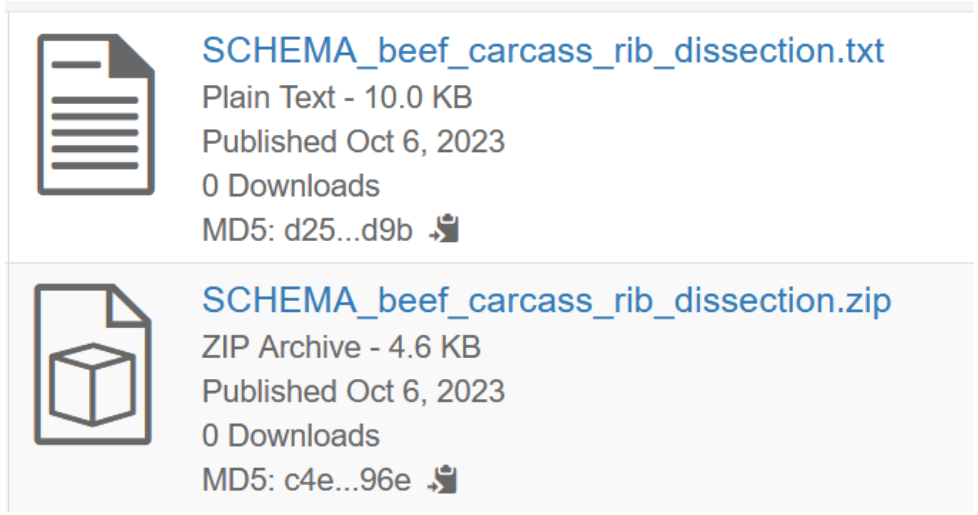
```
BEGIN_REFERENCE_MATERIAL
**********************************************************************
OCA_READ_ME/1.0
This is a human-readable schema, based on the OCA schema standard.

Reference for Overlays Capture Architecture (OCA):
https://doi.org/10.5281/zenodo.7707467

Reference for OCA_READ_ME/1.0:
https://github.com/agrifooddatacanada/OCA_README

A schema describes details about a dataset.
In OCA, a schema consists of a capture_base which documents the attributes and their most basic features.
A schema may also contain overlays which add details to the capture_base.
For each overlay and capture_base, a hash of their original contents has been calculated and is reported here as the SAID value.

This README format documents the capture_base and overlays that were associated together in a single OCA Bundle.
OCA_MANIFEST lists all components of the OCA Bundle.
For the OCA_BUNDLE, each section between rows of ****'s contains the details of one "layer type/version" of the OCA Bundle.
**********************************************************************
END_REFERENCE_MATERIAL

BEGIN_OCA_MANIFEST
**********************************************************************
Bundle SAID: unavailable

"capture_base" SAID: "ECVOLOYXAiPyT41U0PlV_Pm40jp7e401v6lIFoWjaUUw",
"character_encoding" SAID: "EjtthI2d4NN8LU9qrly914IfuFkesoWs50uaL1fiJFnY",
"information (en)" SAID: "EY25fT-ckXLWW_RjZ92LC_TBbQt2d9KoYaJTwl0Lmg-k",
"label (en)" SAID: "EIp_dI-md43kpMTHWSMw0COOURcpqu_2yINFs-V4wTto",
"meta (en)" SAID: "EpWlGBud8OJlwoQ222tpHY_3PW5getsCTbWHM4zvgx38",
"unit" SAID: "EAUOJ5BG1zgp63nk7XnRVhVq1tUTw2Kzm4tHBgOk2dko"
**********************************************************************
END_OCA_MANIFEST

BEGIN_OCA_BUNDLE
**********************************************************************
Layer name: capture_base/1.0
SAID: ECVOLOYXAiPyT41U0PlV_Pm40jp7e401v6lIFoWjaUUw
classification: RDF402

Schema attribute: data type
   animal_id: Text
   body_fat: Numeric
   bone_wt: Numeric
   date: Numeric
   decl_bull: Text
   decl_cow: Text
   decl_heifer: Text
   decl_steer: Text
```

JSON Information overlay (viewer)



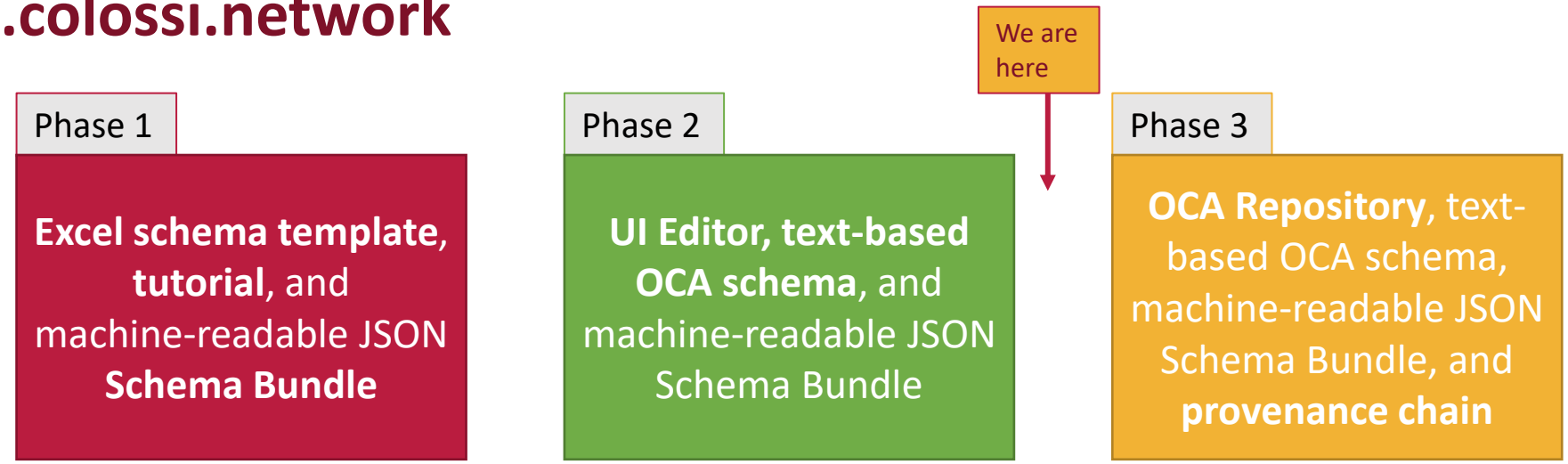
Schema now available in repository





How to follow our work

- **Agrifooddatacanada.ca** – the root of all our work
- Tools available at **semanticengine.org**
- Reach out to us at adc@uoguelph.ca with ideas, bugs, comments, collaborations
- Participate in the OCA standard and more at **humancolossus.org** and **oca.colossi.network**



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

agrifooddatacanada.ca

UNIVERSITY of GUELPH