A data ecosystem serving agri-food sustainability

Lucas Alcantara, Ph.D.
Manager, Research Centre Data
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
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
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
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
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 helps researchers write rich contextual data documentation
Adding context through documentation

- Easy to create
- Easy to standardize
- Citable
- Easy to make machine-actionable
- Fully Portable
- Easy to share
- Easy to reuse
- Easy to extend
Context should travel with data

- Portable
- Machine-actionable
- Citable

Data’s contextual documentation

Data
The first contextual documentation tool from ADC focuses on writing rich data schemas.
Semantic Engine for better schemas

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)

<table>
<thead>
<tr>
<th>animal_id</th>
<th>duration</th>
<th>session_n</th>
<th>total_yield</th>
<th>milking_location</th>
</tr>
</thead>
<tbody>
<tr>
<td>4551</td>
<td>00:05:39</td>
<td>3</td>
<td>9.03</td>
<td>Voluntary Milking System</td>
</tr>
<tr>
<td>4551</td>
<td>00:06:23</td>
<td>4</td>
<td>10.14</td>
<td>Voluntary Milking System</td>
</tr>
<tr>
<td>4604</td>
<td>00:05:12</td>
<td>1</td>
<td>14.83</td>
<td>Rotary Parlour</td>
</tr>
<tr>
<td>4598</td>
<td>00:06:41</td>
<td>1</td>
<td>28.63</td>
<td>Rotary Parlour</td>
</tr>
</tbody>
</table>
Attributes are described in a schema

<table>
<thead>
<tr>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
</tr>
<tr>
<td><strong>Animal ID</strong></td>
</tr>
<tr>
<td>4551</td>
</tr>
<tr>
<td>4551</td>
</tr>
<tr>
<td>4604</td>
</tr>
<tr>
<td>4598</td>
</tr>
</tbody>
</table>

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

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

OCA files are machine-readable

```
{
    "capture_base": "EzB01m3oiJ7-tWwHdp-KSPdqvQpbiZTxFuu4CdcCQba",
    "digest": "ECp6yLXgqaw1A7Wx4hjNn_I-ta3eQoh6ZL5S0_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.