Agri-food Data Canada: A data ecosystem serving agri-food sustainability

L.M. Alcantara, C. Huitema and A.M. Edwards

Agri-food Data Canada, University of Guelph, Office of Research, 50 Stone Road East, N1G 2W1, Guelph, ON, Canada
Corresponding Author: alcantal@uoguelph.ca

Abstract

Agri-food Data Canada (ADC) is creating a data ecosystem serving agri-food sustainability. Through investments in technology, infrastructure, and culture, we are helping researchers and the research community get more value from the data researchers are already collecting. Agri-food Data Canada's approach is guided by the FAIR data principles (that data should be Findable, Accessible, Interoperable and Reusable). To improve data FAIRness ADC is

1. Creating a semantic engine that will help researchers create and use better machine-actionable, reusable, and accessible descriptions and governance for their data, projects, algorithms, tools, workflows, and other digital research outputs;
2. Collaborating on projects supporting the federation of data silos, to ensure that data, metadata, and access rights can travel with the data from source to destination within the ADC federation;
3. Developing tools to help researchers with data provenance and traceability; and
4. Creating a culture of FAIR data by developing knowledge-sharing resources such as webinars, training, and teaching materials.

ADC works with partners to align our approaches and contribute to the global research community, with the goal to ensure research data is FAIR. One collection of tools that are under development at ADC is the Semantic Engine. While there are many approaches to harmonizing data through the creation of data platforms, ADC sees the value in adding value to heterogeneous data through the creation of tools that improve data without the necessity of data platform infrastructure. Researchers can improve their data documentation workflows by adding context to their data through the creation of machine-actionable data schemas. At the heart of the Semantic Engine is the Overlays Capture Architecture (OCA), an international open standard created by the non-profit organization Human Colossus Foundation. OCA's layered architecture is machine-actionable and easy to generate. OCA schemas allow multiple contributors to improve a schema independently and permits the bundling of schemas with appropriate task-specific schema overlays. Schemas can be internationalized through the creation of language-independent overlays, and their additions do not change the underlying structure of the schema which ensures interoperability and allows schemas to be continually improved throughout the dataset's lifecycle. OCA also permits the use of downstream data validation rules carried by schemas and enables the incorporation of ontological terms. For example, ontologies, terms, and data standards endorsed by ICAR can be added to schemas to improve data interoperability and harmonization, which are essential for advancing the international agri-food sector. Agri-food Data Canada is developing a powerful collection of tools and creating a data ecosystem that will reduce barriers to data documentation, ease data sharing, and support the international agri-food sector's data needs.

Keywords: FAIR data, metadata, data governance

Introduction

Agri-food Data Canada (ADC) is creating a data ecosystem serving agri-food sustainability. Through investments in technology, infrastructure, and culture, we are helping researchers and the research community get more value from the data researchers are already collecting. Agri-food Data Canada's approach is directed by the FAIR Guiding Principles. According to Wilkinson *et al.* (2016), the FAIR principles provide a framework for making data Findable, Accessible, Interoperable, and Reusable.

Findability emphasizes the need for data to be assigned globally unique and persistent identifiers, described with rich metadata that includes identifiers, and registered or indexed in searchable resources. Accessibility focuses on ensuring data can be easily retrieved using standardized protocols that are open, free, and universally implementable, while also maintaining access to metadata even when the data itself is no longer available. Interoperability emphasizes the use of formal and broadly applicable languages for knowledge representation, along with vocabularies that align with FAIR principles, and the inclusion of qualified references to other data. Reusability stresses the importance of richly describing metadata with accurate attributes, associating data with clear and accessible usage licenses, and meeting domain-relevant community standards. By adhering to the FAIR principles, data becomes discoverable, accessible, compatible, and usable, enabling broader and more effective data sharing and integration across disciplines and communities.

To enhance the FAIRness of data, ADC is implementing several initiatives. These include:

- Creating the Semantic Engine: A tool designed to assist researchers in generating and utilizing machine-actionable, reusable, and accessible descriptions and governance for their data, projects, algorithms, tools, workflows, and other digital research outputs.
- Collaborating on federated data projects: ADC actively engages in collaborative efforts aimed at supporting the federation of data silos. This ensures that data, metadata, and access rights can seamlessly travel with the data from its source to its destination within the ADC federation.
- Developing data provenance and traceability tools: ADC is committed to building tools that aid researchers in capturing and tracking data provenance, promoting transparency and reproducibility in research.
- Cultivating a culture of FAIR data: ADC fosters a culture that promotes FAIR data principles by creating and sharing knowledge-sharing resources, such as webinars, training sessions, and teaching materials. These resources aim to educate and empower researchers to adopt and implement FAIR data practices.

Agri-food Data Canada is actively contributing to improve the FAIRness of data and facilitate the adoption of FAIR data principles within the agri-food research community. Unlike other methods of harmonizing data through data platforms, ADC recognizes the significance of enhancing heterogeneous data directly without relying on complex data infrastructure. One promising suite of tools that are under development at ADC is the Semantic Engine, which will empower researchers by enabling them to enhance their data documentation workflows effortlessly for effective uptake of data share and reuse.

Semantic Engine

The Semantic Engine is a suite of tools being developed by ADC to help researchers write rich contextual data documentation based on machine-actionable data schemas, thereby improving its overall quality, portability, standardization, and reuse. These tools

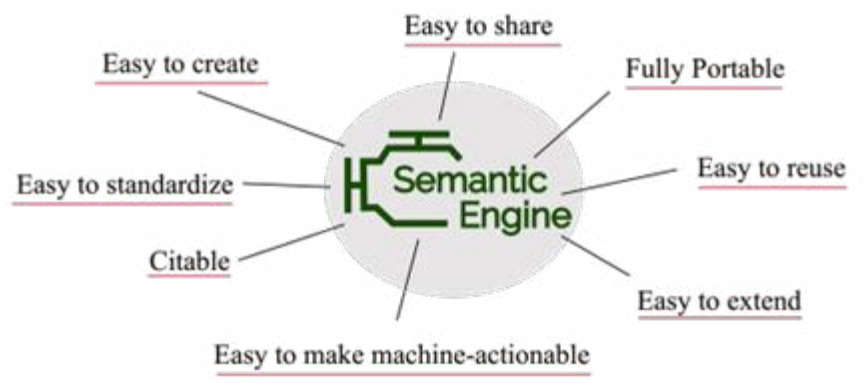

Figure 1. Semantic Engine is a suite of tools that will help researchers write rich contextual data documentation.

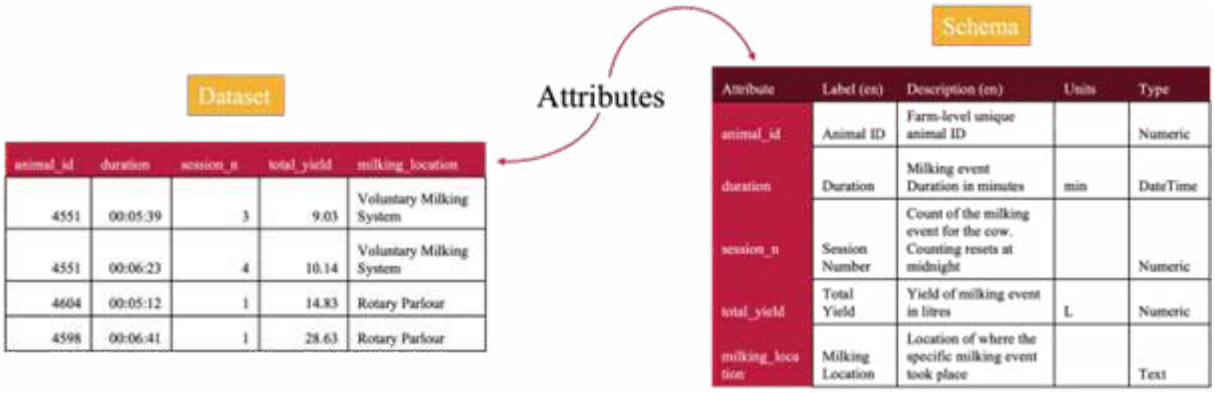

Figure 2. Data must be structured to be understood and a schema describes the structure of the dataset.

will offer a practical solution for researchers looking to maximize the value of their data without the burden of implementing extensive data platform frameworks.

In practical terms, the Semantic Engine offer numerous benefits that cater to various stakeholders and collaborators, including the researchers themselves. These tools will avoid 'mystery' data by providing improved data descriptions, allowing for a clearer understanding of the data's content and context. Additionally, they allow researchers to adjust the level of detail based on specific needs, ensuring that the data is comprehensive yet focused. By using these tools, researchers can deposit high-quality data with less effort, saving time and resources while enhancing the overall value and usefulness of the data.

Another significant advantage of the Semantic Engine is its ability to aid others in utilizing research data effectively. By providing better data descriptions and context, researchers can reduce the time spent supporting individuals who are working with their data, enabling them to navigate and interpret it more efficiently (Figure 2). This is particularly valuable in cross-disciplinary research where data from different domains may need to be integrated. These tools empower researchers to communicate the necessary information clearly, improving data understanding and facilitating collaboration across diverse research fields.

There is an additional benefit, as it also caters to machine consumption of data. Using machine-readable schemas, data can be easily discovered and utilized by automated systems. Publishing these schemas promotes better collaboration and interoperability, allowing researchers to integrate data from various sources seamlessly. The ability to assign a separate DOI to the schema ensures that it can be cited and used by others, contributing to a more robust and efficient scientific ecosystem. By improving data accessibility and usability for machines, these tools pave the way for better science outcomes derived from high-quality and well-documented data.

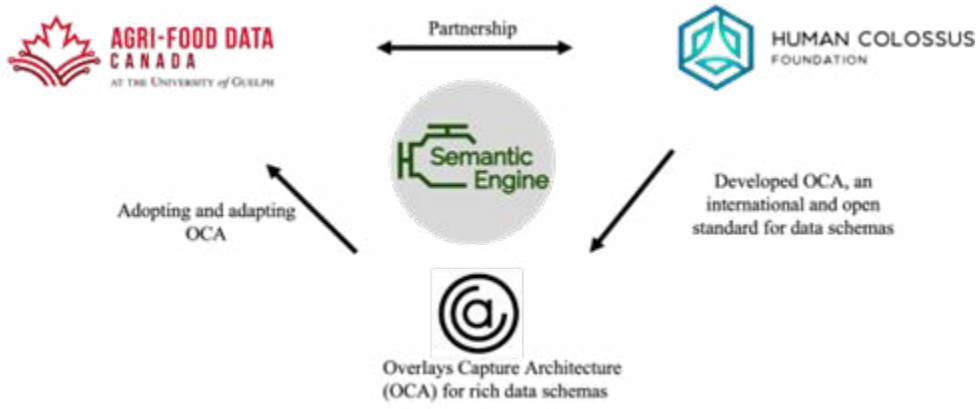

Figure 3. Overlays Capture Architecture, an international open standard developed by the Human Colossus Foundation, is at the heart of Agri-*food Data Canada's Semantic* Engine.

Overlays Capture Architecture

To create the Semantic Engine, ADC is partnering with the Human Colossus Foundation to adopt its work on Overlays Capture Architecture (OCA) as the underlying schema standard and adapt it to the agri-food research area (Figure 3).

Overlays Capture Architecture is an extensible, flexible, international, open, and machine-accessible standard for data schemas (Knowles, 2022). From a table representation of a schema, an OCA schema splits each feature into a separate layer and each layer is a separate file (written in a machine-readable format) that recognizes the capture base, or the foundation of the schema describing the data set (Figure 4). Layers are added to the schema to provide more detail, making it easier to understand and use data collected and structured according to the associated schema.

Overlays Capture Architecture schemas offer a flexible and collaborative approach to schema development by allowing multiple contributors to enhance a schema individually. This decentralized approach enables each contributor to work on specific aspects or components of the schema without disrupting the underlying structure. Furthermore, the ability to add new elements or modify existing ones in overlays ensures that schemas can be continuously improved and refined throughout the lifecycle of a dataset, promoting long-term data quality and adaptability to evolving requirements.

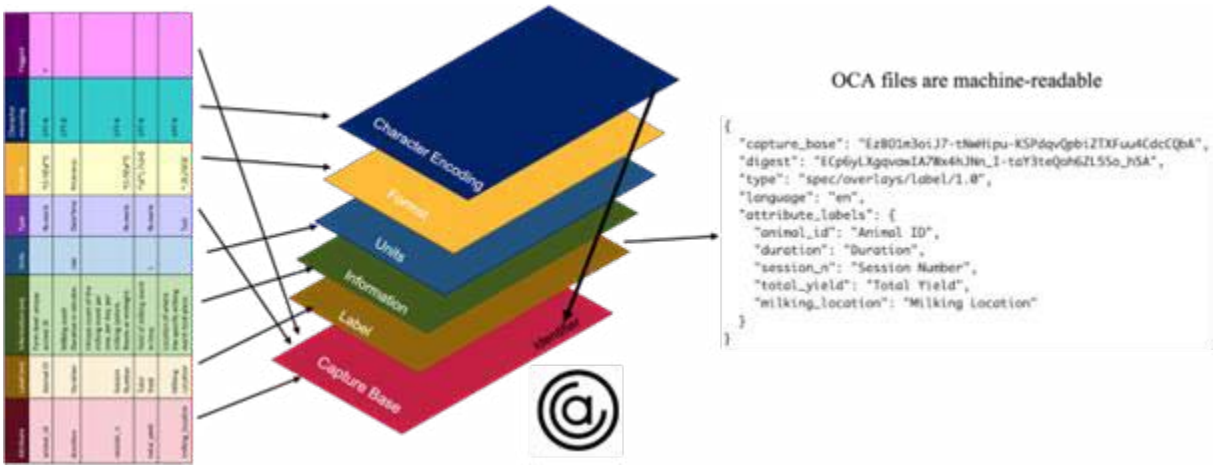

Figure 4. Overlays Capture Architecture expresses rich data schema in a series of distinct overlays that are machine-readable.

Additionally, OCA schemas support internationalization through the creation of language-independent overlays, which enables the translation of schema elements into different languages without altering the schema's core structure. This feature enhances interoperability across different linguistic contexts and enables the seamless exchange of data between diverse systems.

Applications for the agri-food sector

The benefits offered by ADC through the Semantic Engine and OCA are applicable for researchers in the agri-food sector and directly extensible to members of the International Committee for Animal Recording (ICAR) in numerous ways. Firstly, OCA allows for the inclusion of downstream data validation rules through schemas, ensuring data quality and consistency. This feature is particularly valuable for ICAR members as it enables the incorporation of ontological terms and data standards endorsed by the committee, facilitating data interoperability and harmonization. By leveraging OCA and its integration capabilities, ICAR members can enhance the international agri-food sector by promoting standardized and consistent data management practices.

ADC's development of a comprehensive collection of tools and a robust data ecosystem further benefits ICAR members. These initiatives address the challenges faced by the international agri-food sector, such as barriers to data documentation and data sharing. ADC's tools simplify the process of data documentation, making it more accessible and efficient for researchers to capture and share valuable data.

Agri-food Data Canada is developing a data ecosystem to serve agri-food sustainability through a powerful collection of tools that will reduce barriers to data documentation, ease data sharing, and support the international agri-food sector's data needs. This data ecosystem will foster collaboration and cooperation among stakeholders, enabling seamless data integration and enhancing the overall data needs of the agri-food sector.

References

Knowles, P, 2022. Overlays Capture Architecture (OCA) (1.0). Zenodo. https://doi.org/10.5281/zenodo.7707467.

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.,* 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 3(1).