Automating Data Unification, Governance, and Analytics in Animal Science: A Cloud-Native, AI-Assisted Framework

Niu Puchun*[1], Liu Enhong[1], Zhang Hanlu[1], Kebreab Ermias[2], Ellis Jennifer[3], Niu Mutian[4], Terry Stephanie[5], Cowley Fran[6], Rode Lyle[7], Hostens Miel[1]

[1] Cornell University, [2] University of California, [3] jellis@uoguelph.ca, [4] ETH Zurich, [5] Agriculture and Agri-Food Canada, Lethbridge, AB T1J 4B1, Canada, [6] University of New England, [7] Sage Nutrition Corp

Metabolism experiments typically use structured or partially controlled conditions and produce large, complex datasets covering animal characteristics, nutrient intake, digestibility, gas emissions, feeding behavior, and production traits. These datasets have grown more complex as researchers routinely collect high-resolution phenotypes and rely more on indirect-calorimetry technologies like respiration chamber and GreenFeed system. Yet inconsistencies in data structure and metadata limit the use of these measurements as reliable phenotypes for genetic analysis and decision support. This creates major bottlenecks when trying to clean, validate, and combine data across studies.

We developed a standardized Excel workbook and a fully automated, cloud-based pipeline for data ingestion, validation, and analysis. Our work was driven by the GEMS project (Accurate Gas Emissions Measures from Cattle with the GreenFeed System), which integrates methane emission, feed intake, and metabolism data from roughly 150 studies worldwide to support the generation of consistent, high-quality methane phenotypes. The workflow runs on Microsoft tools and handles automated ingestion via Microsoft Teams to Azure Blob Storage, through rule-based validation in Microsoft Azure Databricks. We use structured information from study-level data user agreements to define which variables and datasets should be included, allowing the system to check delivered data against what was promised and send notifications when corrections are needed or elements are missing. Once validated, the data are merged with external gas-measurement outputs and organized into analysis-ready tables. Researchers can access these curated datasets in a secure Databricks Clean Room using Python, R, or natural-language AI tools for analysis and modeling, including applications in methane phenotype evaluation, genetic studies, and feed-efficiency research.

While built for GEMS, the modular design can be applied broadly to animal metabolism, greenhouse gas and genetic studies. It reduces manual data handling, improves reproducibility, and accelerates the generation of reliable methane and intake phenotypes for management and genetic analysis.