



COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD Qualifying Exam

Tuesday September 26, 2023 at 1PM, In-Person REYN 1101

Kassy Raymond

*Using Graph Databases for the Semantic Enrichment and Discovery of
Heterogeneously Described Metadata and Data:
The Case of Livestock Population Data*

Chair: Dr. Gary Grewal

Advisor: Dr. Deborah Stacey

Co-Advisor: Dr. Theresa Bernardo (Population Medicine)

Non-Advisory: Dr. Daniel Gillis

Non-Advisory: Dr. Zvonimir Poljak (Population Medicine)

Abstract:

Data and metadata originating from National Statistics Organizations (NSOs) and Intergovernmental Organizations (IGOs) are crucial inputs for One Health metrics, encompassing areas like Greenhouse Gas emissions, global nutrition, poverty, and the economic and health impact of animal diseases on both human and animal populations. However, this data is scattered across the web, characterized by diverse descriptions, formats, and semantically similar yet distinct concepts. The implications of these interoperability challenges include difficulties finding and accessing data. No previous studies have explored the potential of semantic search for identifying both national and intergovernmental statistics across time, or for enabling checks for consistency between datasets that share similar but differently described attributes.

Using livestock population data as the case study, the proposed research investigates the use of graph databases as a powerful tool to explore and make data and metadata interoperable. It provides a structured way to assess how data consistency evolves over time when it represents similar but differently categorized concepts. The proposed research will focus on four main objectives:

1. Identifying the requirements for making metadata and data interoperable by reviewing data sources that report livestock population data.
2. Developing, implementing, and evaluating a graph database tailored for the storage, management, and retrieval of diverse metadata.
3. Creating a bottom-up semantic model that reconciles differences in naming and classifications, making it easier to find data that represents similar concepts despite different classifications.
4. Investigating the use of graph databases to store and retrieve external consistency of datasets over time.

By addressing these objectives, this research aims to harness the potential of graph databases to harmonize and improve access to critical data and metadata. This, in turn, can advance our understanding of One Health metrics and promote cross-disciplinary insights through improving the findability and interoperability of data and metadata.