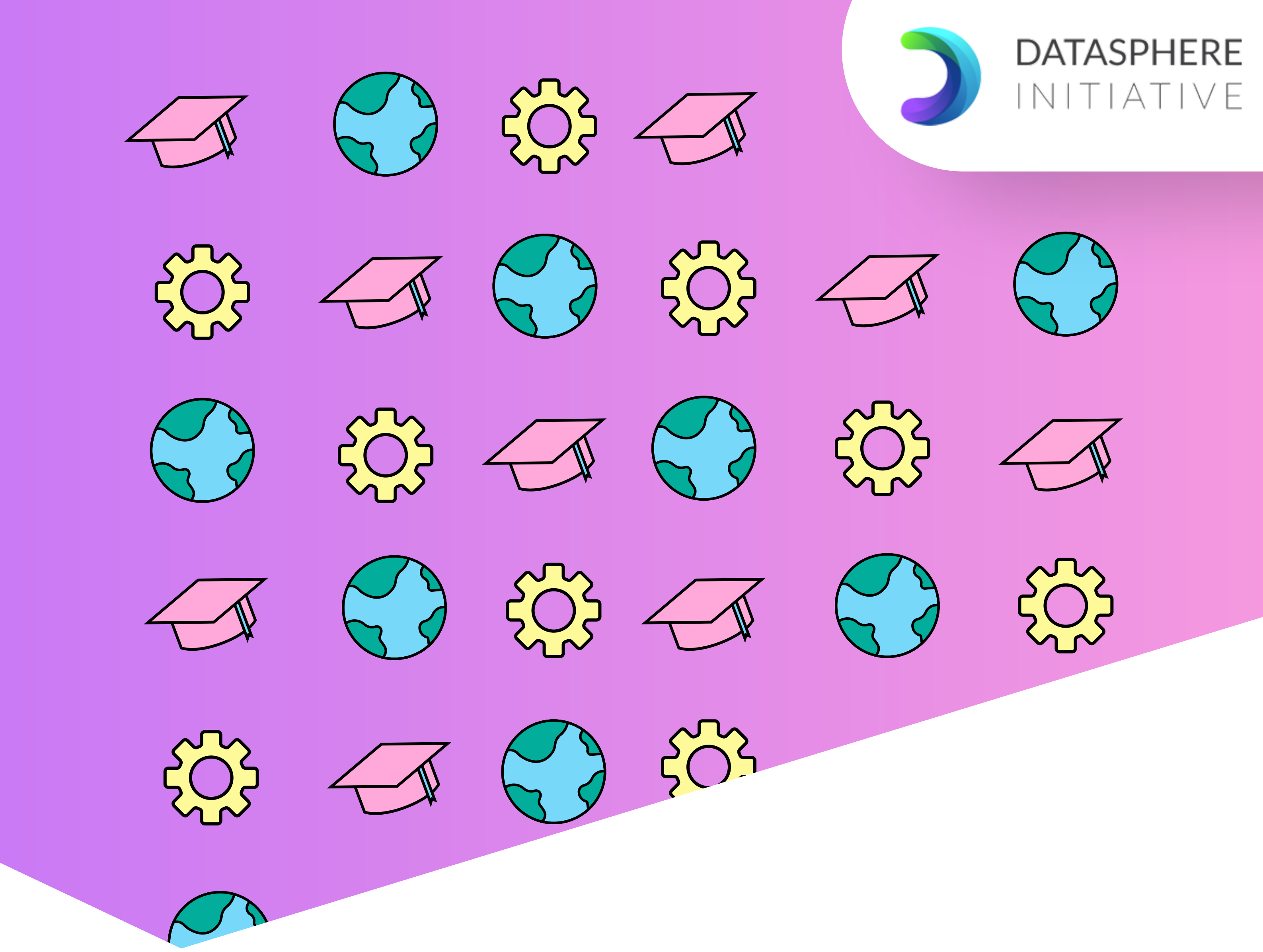
DATASPHERE
INITIATIVE



INTEROPERABILITY BY DESIGN FOR A COLLECTIVELY GOVERNED DATASPHERE

CHALLENGES IN LIVESTOCK POPULATION NAMING CONVENTIONS ANALYZED USING A GRAPH DATABASE MODEL

KASSY RAYMOND

FELLOWSHIP

2021-2022

ABOUT THE DATASPHERE INITIATIVE

The Datasphere Initiative is a global network of stakeholders fostering a holistic and innovative approach to data governance. By cultivating dialogue and connecting communities, the Datasphere Initiative connects sectoral silos and people to build a collaboratively governed Datasphere and responsibly unlock the value of data for all.

For more information, visit www.thedatasphere.org or contact info@thedatasphere.org.

REPORT CITATION AND COPYRIGHT

Raymond, K. (2023) Interoperability by Design for a Collectively Governed Datasphere: Challenges in Livestock Population Naming Conventions Analyzed Using a Graph Database Model. Datasphere Initiative.

ACKNOWLEDGEMENTS

The paper was authored by Kassy Raymond as an output of the Datasphere Initiative Fellowship Program 2021-2022.

The author is thankful to Carolina Rossini from the Datasphere Initiative and Dr. Deborah Stacey (University of Guelph and Global Burden of Animal Diseases) for their careful review of the paper and suggestions for improvement, and ongoing support and mentorship. Thank you to Sophie Tomlinson, Director of Communications at the Datasphere Initiative for her review and input, and to Natalia Loungou for helping with reformatting figures. The author is also thankful to the members of the GBADs Informatics theme who developed webscraping scripts, database tables and the Application Programming Interface that was used to improve the accessibility to the data used in the analysis presented.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [570288-2022].

This paper is based on research funded by (or in part by) the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

The paper has been designed by Natalia Loungou.

ABOUT THE AUTHOR

> *Information about data, the data itself, the interpretations and categorisations of the data by the people who use and create the data and the connections that tie the data together comprise the Datasphere.*

Kassy Raymond

Kassy Raymond is a doctoral student in Computational Sciences with a Collaborative Specialization in One Health in the School of Computer Science at the University of Guelph. Kassy is studying under the supervision of Dr. Deborah Stacey (School of Computer Science) and Dr. Theresa Bernardo (Population Medicine). She is also the Technical Manager for the Global Burden of Animal Diseases (GBADs) Informatics theme.

Kassy's research explores the operationalization of data governance principles to improve the quality, discoverability and reusability of animal health and production data in the GBADs knowledge engine. She is interested in exploring data infrastructure to improve and understand data quality and interoperability, specifically using graph databases to map between Open Data sources and identifying the insights that can be drawn using this tool.

Her research is complimented by her interdisciplinary background developed through academic and professional work in biology, bioinformatics, machine learning, data science, and data governance.

Kassy's research is supported by the Natural Sciences and Engineering Research Council of Canada's (NSERC) Doctoral Graduate Scholarship (CGS-D) and through the support of GBADs through funding from the Bill and Melinda Gates Foundation.

She holds a Master of Science in Bioinformatics and a Bachelor of Science in Biological Sciences from the University of Guelph.

This report is an outcome of her Fellowship at the Datasphere Initiative 2021/2022.

TABLE OF CONTENTS

ABSTRACT

Livestock population data is a critical input to estimates of climate change, nutrition, and protein availability, livestock biomass, antimicrobial resistance calculations, and calculating the burden of animal disease on humans and animals. Data must be discoverable, reusable, and interoperable to be combined, compared, and used for these estimates. However, differing classification systems and naming conventions exist between and within data sources raising barriers to the interoperability and discoverability of data resources. Data cataloguing and metadata projects have been established to aid in the discovery of data, but little work has been done in consolidating the classifications of these data alongside relevant data. A graph database framework is presented and implemented using livestock population data from four data sources as a case study. The variability in species naming conventions from the case study are discussed using the concept of the Datasphere and suggestions are made to improve the interoperability of livestock population data from national and international government sources.

1. INTRODUCTION

Livestock population data is a critical input to estimates of climate change, nutrition and protein availability, livestock biomass, antimicrobial resistance calculations, and calculating the burden of animal diseases on humans and animals [1], [2], [3], [4]. Given the vast applicability of these data, they play a pivotal role in decision-making and policy, and tracking progress towards the Sustainable Development Goals (SDGs). The Food and Agriculture Organization of the United Nations Statistical Database (FAOSTAT) is the largest repository of national livestock population data, however it does not provide data at the level of disaggregation required for some estimates (i.e. by sex, utility and breed of animal) [5]. Therefore, researchers often seek other data sources from national or regional aggregators to obtain the data at the level of disaggregation needed for their analyses [4]. However, data that is decentralized across many sources is often difficult to discover. Once decentralized data sources are discovered, the practitioner can also face additional challenges with semantic interoperability due to differing classification systems or categories used to describe livestock populations.

Semantic interoperability is defined as the ability for the meaning of data to be unambiguously exchanged between senders and receivers of data, where the senders and receivers can be machine to machine, machine to human, or human to human [6]. Terminology used to categorize data is subject to differences between data sources and by time and geographic region which presents a barrier in achieving semantic interoperability. This notion is tightly intertwined with the discoverability of data; the way data is categorized informs the discovery of the data since the terms used to classify data inform their contents.

Data cataloguing and metadata projects have been established to aid in the discovery of data, but little work has been done in consolidating the classifications of these data alongside relevant data and metadata. Acknowledging that tools are required to discover and understand distributed databases, Muñoz *et al.* (2017) acknowledged that tools are required for the management of distributed data and created a list of common tasks for ecologists working with data [7]. Here, we adopt the needs identified by Muñoz *et al.* (2017) to fulfil the needs of those working with data in other disciplines and to consider the classification of data [7]:

1. To discover existing dispersed and heterogeneous datasets
2. To discover relationships between datasets that are of potential interest
3. To interpret the semantics of data
4. To be aware of the conditions of access and use

Extending from (2), the relationships that may be of potential interest include understanding how the classification of data may change over time, by data source, country, and within data sources. In addition, as elucidated in the Background section below, since a single international standard does not currently exist, recommendations for improving the interoperability and discoverability of data include using an international standard.

In this paper, as our contribution to the field, we present a novel graph database-based framework for temporal querying of data and metadata which allows for the discovery and cataloguing of data using livestock population data and their classifications as a case study. A graph model was created and implemented in Neo4j, a graph database management system, using livestock population data from FAOSTAT, EuroStat, the World Organization for Animal Health (WOAH) and the Ethiopian Central Statistics Agency (EthCSA). The graph database was then used to analyze the livestock categories used by the data sources selected. Finally, we present a discussion that situates the results in the complex systems comprising the Datasphere, including the governance structures in the creation of standard classification systems for categorizing livestock agriculture data and provide suggestions to improve semantic interoperability and discoverability of data.

2. BACKGROUND

In the data sharing and research community, the advent of the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) has been instrumental in encouraging the sharing and use of standard metadata including publishing data using standards [8]. However, data still resides in different forms, repositories, and portals, and in some sectors common data standards are lacking inhibiting the ability to combine and leverage data. In this section we report on current and past initiatives to improve the interoperability and discoverability of livestock data. Graph databases are then introduced as a tool to improve the interoperability and discoverability of decentralized data.

2.1. THE CURRENT STATUS OF INTEROPERABILITY AND STANDARDS IN LIVESTOCK AGRICULTURE

In the 2014 report by the World Bank and FAO, “Investing in the Livestock Sector: Why Good Numbers Matter”, a key message was that “Data integration, i.e. the use of data originating from different livestock, agricultural and nonagricultural surveys, is essential for the design of effective sector policies and investments” [9]. Combining and integrating data requires data that are interoperable, including consistent species classifications by utility, production system, and sex, or data that are linked to definitions of terms or classifications used. The report recommends that agreed-upon international standards and classifications be used in the collection and generation of livestock data and statistics. However, it does not provide examples of classifications and standards [9].

Bahlo *et al.* (2019) reviewed current and past standards in livestock agriculture and identified that despite multiple attempts at creating an international standard for livestock agriculture, many have since been deprecated or withdrawn [10]. For instance, in 2000, the International Organization for Standardization published an agricultural data element dictionary (Electronic data interchange information systems in agriculture - Agricultural data element dictionary - ISO11788) but it has since been withdrawn as an international standard [11]. Similarly, in 2013, the World Wide Web Consortium (W3C) formed a Livestock Data Interchange Standards Community Group. The group was intended for dairy and sheep/beef organizations in New Zealand to facilitate the standardization of data about livestock including management, health, and production information and was active for only one year [12].

The Statistics Division of the United Nations (UN) and the Food and Agriculture Organization of the United Nations (FAOSTAT) created an agricultural vocabulary, metadataset, and classification for agricultural censuses. Agrovoc is a multilingual and controlled vocabulary of agricultural concepts and terminology used and created by the FAO. The Agriculture Metadata Element Set (AgMES) is a metadata standard in the agriculture domain created by the FAO in 2010 [13]. However, the current schema is no longer maintained and the adoption of these standards by organizations besides the FAO is not clear.

In addition, the Consortium of International Agricultural Research Centers (CGIAR) Platform for Big Data has created a list of most commonly used Ontologies for Agriculture where species classifications include the NCBI Taxonomy, and Livestock phenotype ontologies including the Animal Trait Ontology for Livestock [14]. These classifications are useful in bioinformatics, and reviewing and phenotypic animal traits, respectively, but do not specifically address the heterogeneity or differences in the meaning of species classifications that exist in national and international databases [15], [16].

Although these efforts provide standards in the livestock agriculture domain, current literature focuses on organizations and initiatives that are developing standards or ontologies, and little work has been done in consolidating standards used in current reports of livestock statistics sourced from national or international sources. The lack of vocabulary or catalog of livestock population categories that are currently used motivated the research presented in this article.

2.2. GRAPH DATABASES

Graph databases are databases that store objects and their relationships as nodes and directional relationships [17]. They provide a dynamic and flexible schema that allows data to be structured in a connected format. Graph databases consist of a graph model which articulates what type of data will be stored and its relation to other data [18]. They consist of the following:

- Nodes: An entity that can be labelled to represent a type or role.
- Relationships: Directional relationships (usually presented as a verb) which connect two nodes.
- Properties: Key-value pairs. Both relationships and nodes can have properties.

Graph databases have been used in ecology, cancer research, and bioinformatics to manage metadata, connect related entities, and to improve semantic interoperability [7], [19], [20]. However, temporal querying (e.g., being able to query over time) is a technical challenge in the creation and implementation of graph databases; it is difficult to design graph systems that allow users to query for information over time [21]. In our implementation, we provide a method to allow for the temporal querying of resources.

3. MATERIALS AND METHODS

The graph database was created in the following stages: (1) selection of data sources and datasets, (2) data and metadata collection, (3) data preparation and category extraction, (4) graph model creation and implementation (5) creation of queries and (6) exploration of query results (Figure 1). Stages 4-6 were developed iteratively; the creation of the queries informed the graph model and its implementation. Therefore, when the model did not support the types of queries required, the graph model was altered until the model supported the developed queries.

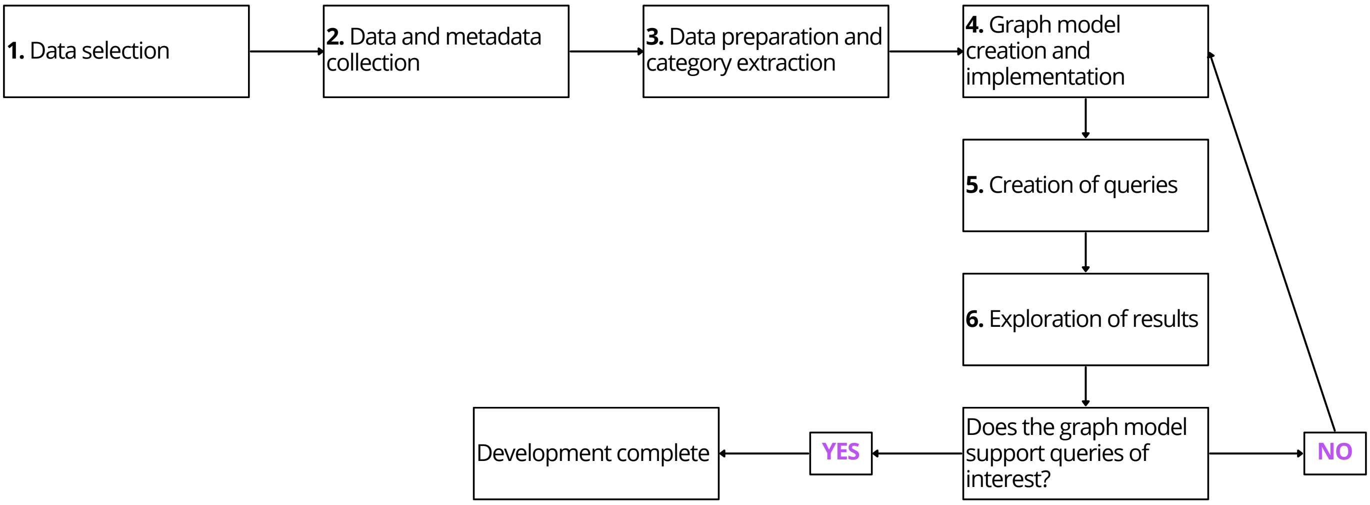

1. Data selection
2. Data and metadata collection
3. Data preparation and category extraction
4. Graph model creation and implementation
5. Creation of queries
6. Exploration of results
Does the graph model support queries of interest?
YES
NO
Development complete

Figure 1: Stages for the development of the graph database.

3.1. DATA COLLECTION AND PREPARATION

3.1.1. SELECTION OF DATA SOURCES AND DATASETS

Livestock population data from FAOSTAT [22], EuroStat [23], WOAH [24], and EthCSA [25] (Table 1) were selected as the case study for the created graph database framework. These sources were selected because they are currently used by the Global Burden of Animal Diseases (GBADs) Programme where Ethiopia is a case study for the calculation of disease burden on animals [26].

3.1.2. DATA AND METADATA COLLECTION

Table 1 provides an overview of the data sources, data tables, links, or Application Programming Interface (API) calls to data, methods used to obtain the data, and the date of collection. Livestock population data from the selected data sources was located on online portals. FAOSTAT has two main sources of livestock population; the Crops and Livestock Products dataset and the Emission: Enteric Fermentation dataset. At the time of collection, the FAOSTAT API was deprecated so data was collected via direct download. Data from the WOAH was collected via the GBADs API, where data was originally obtained from an internal contact at WOAH. Relevant data from EuroStat was located using the EuroStat data portal. Once data was located, they were extracted using the EuroStat API [27].

National livestock population data from the EthCSA was obtained from GBADs Informatics scraping scripts, which provides the data in a digitized format. The data are originally available from annual agricultural surveys which are disseminated by the EthCSA in PDF reports [28]. Data from 2005-2020 was available from PDFs which were collected, scraped, and digitized using webscraping programs developed by the GBADs Informatics theme. The digitized reports were formatted in Amazon Web Services database tables and made available via an API.

In the cases of FAOSTAT and EuroStat, metadata was collected via direct download for each data table used. However, WOAH and EthCSA currently do not provide metadata for the data used in this analysis and therefore could not be collected or included.

Data source	Name of dataset or data table	Link to data or API call	Method used to obtain data	Date of collection
Food and Agriculture Organization of the United Nations Statistical Database	Emissions: Enteric Fermentation (EF)	https://fenixservices.fao.org/faostat/static/bulkdownloads/Emissions_Agriculture_Enteric_Fermentation_E_All_Data.zip	Direct download	August 5th, 2022
	Crops and livestock products (QCL)	https://fenixservices.fao.org/faostat/static/bulkdownloads/Production_Crops_Livestock_E_All_Data.zip	Direct Download	August 5th, 2022

World Organization for Animal Health	Livestock population	http://gbadske.org:9000/GBADsLivestockPopulation/oie?species=*&year=<year>&format=file	GBADs API	August 5th, 2022
EuroStat	apro_ec_poula	https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/apro_ec_poula?format=JSON	EuroStat API	September 15th, 2022
	apro_mt_lsequi	https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/apro_mt_lsequi?format=JSON		September 15th, 2022
	apro_mt_lsgoat	https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/apro_mt_lsgoat?format=JSON		September 15th, 2022
	apro_mt_lspig	https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/apro_mt_lspig?format=JSON		September 15th, 2022
	apro_mt_lssheep	https://ec.europa.eu/eurostat/api/dissemination/statistics/1.0/data/apro_mt_lssheep?format=JSON		September 5th, 2022
Ethiopia Central Statistics Agency (EthCSA)	Ethiopia Annual Agriculture Sample Survey: Report on Livestock and Livestock Characteristics	https://github.com/GBADsInformatics/pdfScrapingETHCSA/data	GBADs GitHub	August 5th, 2022

3.1.3. DATA PREPARATION AND CATEGORY EXTRACTION

Once data was collected from the data sources, relevant data was extracted from the tables, and reformatted to be ready to be ingested by the graph database. All data preparation was conducted using the Python programming language. Data preparation programs were developed with support from the GBADs Informatics theme. Only data pertaining to livestock population was extracted from the FAOSTAT Crops and Livestock Products dataset (Table 1); data reporting slaughter, trade and export, and production of animal products was excluded.

3.2. GRAPH DATABASE CREATION AND IMPLEMENTATION

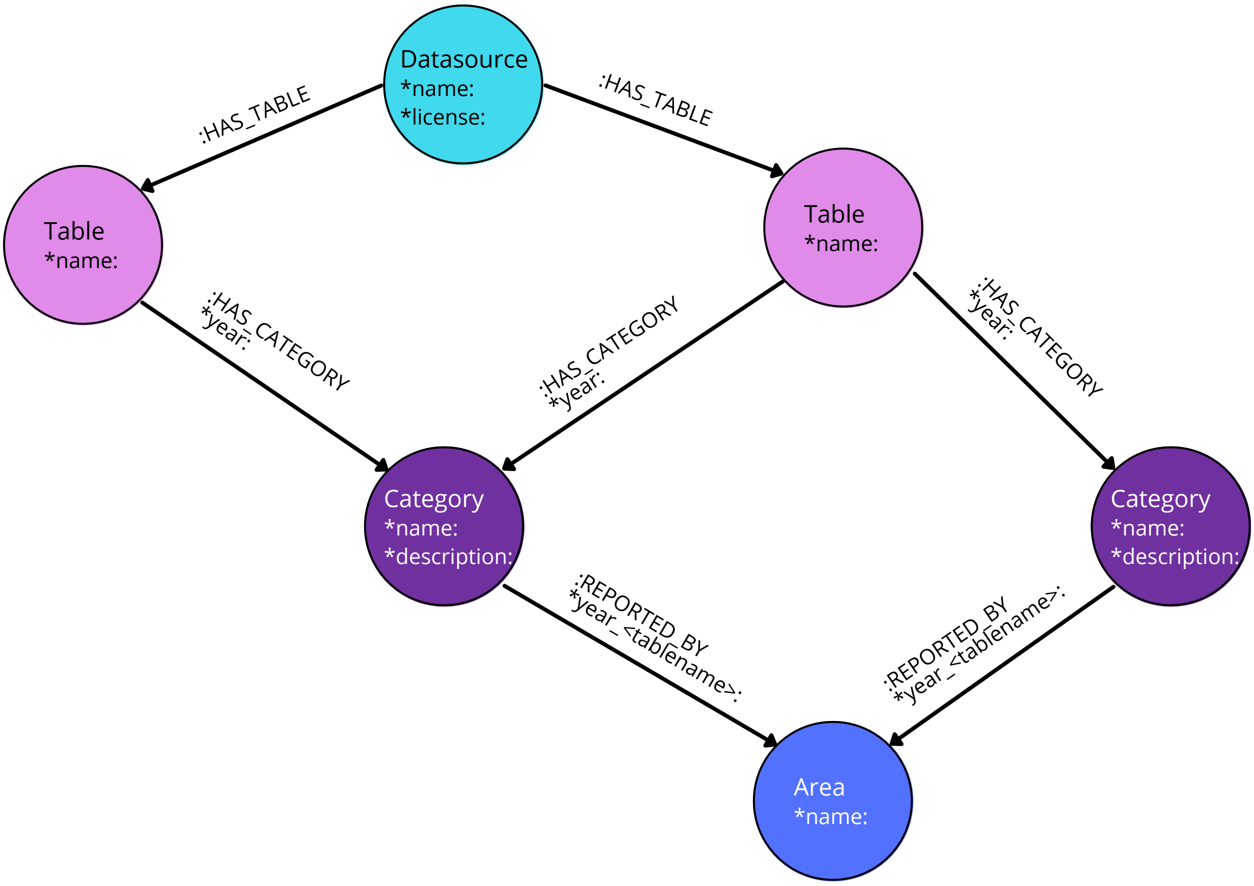


Figure 2: Graph data model that connects data sources, to data tables, to categories, and categories to areas (countries or regions) that report these categories in a given table, and data source. Property names denoted by asterisks (*).

The graph model was implemented by loading the prepared data and metadata into Neo4j, a graph database management system. Data loading was conducted using scripts developed in python using the Neo4j Official Driver. The resultant graph database consisted of 640 nodes and 16,025 relationships.

3.3. QUERIES

Queries were designed to obtain information about the metadata and data contained in the graph, and their relations to each other using the four needs adopted from Muñoz et al. (2017) [7] as outlined in the Introduction:

1. To discover existing dispersed and heterogeneous datasets
2. To discover relationships between datasets that are of potential interest
3. To interpret the semantics of data
4. To be aware of the conditions of access and use

The queries were first posed as a series of questions (*Q*) or tasks for the graph database Queries were developed using the Cypher Query Language (referred to as 'Cypher'), which allows users to access data from the Neo4j graph database. The developed queries were executed through the Python Neo4j Official Driver. Example outputs for each query are provided in Appendix A.

3.3.1. NEED 1: TO DISCOVER EXISTING DISPERSED AND HETEROGENEOUS DATASETS AND NEED 2: TO DISCOVER RELATIONSHIPS BETWEEN DATASETS THAT ARE OF POTENTIAL INTERESTS

Q1: Get a list of all categories that are used in the datasets

Cypher query:

```
MATCH (n:Category)
RETURN n.name
```

The output here is all the categories that are used in all datasets. Given the length (338 categories) of the output, the full output is available on GitHub.

Q2: Given a category/livestock species, what are the data sources that report that category and for which years?

Cypher query:

```
MATCH (n:Category)-[r]-()-[]-(d:Datasource)
WHERE n.name =~ '(?i)<name_of_category>.*'
RETURN n.name, d.name, r.year
```

The above query provides all data sources that report a category for the given category. (?i) provides the output regardless of case (lower case or upper case).

Q3: Which countries report a given livestock species and for which years?

Cypher queries:

```
MATCH (n:Category)-[r]-(a:Area)
WHERE n.name =~ '(?i)<name_of_category>.*'
RETURN n.name, a.name, r.<year_property>
```

The above query provides the countries that report a given livestock species and the years that they are reported.

Q4: For a given country and livestock species, which datasets are available and which years does the country report that species to each data source?

Cypher query:

```
MATCH (n:Area {name: <name_of_country>})-[r]-(c:Category)-[]-()-[]-(d:Datasource)
WHERE n.name =~ '(?i)<name_of_category>.*'
RETURN n.name, c.name, d.name, r.<year_property>
```

The above query provides all datasets that provide data for a given country that reports livestock population for a given livestock species, and the years that the country reports that species to each data source.

3.3.2. NEED 3: TO INTERPRET THE SEMANTICS OF DATA

Q5: What is the definition of a given species category?

Cypher query:

```
MATCH (n:Category {name: <name_of_category>})
RETURN n.definition
```

The above query provides the definition of a category of interest. In some cases, the definition is not available so the query does not yield a result. However, the graph does support the storage of category definitions.

3.3.3. NEED 4: TO BE AWARE OF THE CONDITIONS OF ACCESS AND USE

Q6: For a given data source, what are the associated licences or access restrictions?

Cypher query:

```
MATCH (n:Datasource {name: <name_of_datasource>})
RETURN n.license
```

The above query provides the licences associated with a given data source. In some cases, the licence is not available so the query does not yield a result. However, the graph does support the storage of links to licences.

Additional queries were constructed to gather additional information about the categories used by data sources:

Q7: Return the number of unique categories stored in the database

Cypher query:

```
MATCH (n:Category)
RETURN COUNT(n)
```

The output is 338, meaning that there were 338 unique species conventions among all data sources stored.

Q8: Given x named countries, return the common species categories

For instance, if we were looking for the common species between two countries:

Cypher query:

```
WITH ['<country_1>, '<country_2>'] as names
MATCH (a:Area)
WHERE a.name in names
WITH collect(a) as areas
MATCH (c:Category)
WHERE ALL(a in areas WHERE (c)-[:REPORTED_BY]-(a))
RETURN c.name
```

The above query provides the species that are reported by both countries.

4. DISCUSSION AND CONCLUSION

4.1. GRAPH DATABASE FRAMEWORK

The graph database framework has provided a method to discover datasets and has provided a catalogue of categories used in decentralized datasets that allows for temporally querying of categories within a graph database system. The exercise resulted in a graph-based data catalogue that stores metadata and stores naming conventions used by data sources. The graph model and implementation provides a catalogue that allows users to discover data sets that are related based on the species classification, while fulfilling the needs of information systems providing metadata for decentralized and heterogeneous datasets. In addition, the ability to query temporally allows users to understand which countries report a given species and to which data source. This will allow users the opportunity to fill data gaps that occur when data is missing in one data source for a given species and country but is available in another.

While a case study with only four data sources was used in this analysis, the database model supports the additions of datasets from other data sources. In addition, by expanding the graph model with additional nodes and relationships, different data can be added such as documents, journal articles, statistical acts. By expanding the model, relationships can allow data to be connected to related literary sources or metadata and allow for users to extract this related information at once.

To expand the utility of the graph database to stakeholders, a front-end interface can be developed to allow users to interact with the graph outside of the Neo4j Bloom or Browser interface and to be able to support data in languages other than English. A user-interface would allow for users without a technical background to explore the data via the graph visualization, aiding in the discoverability of resources. Currently, the graph is available via code in GitHub and requires users to create and load a Neo4j instance on their local machine or in Neo4j Aura. Ongoing funding is required to host the graph database in the Cloud and create a query-able user-friendly interface, and host the interface on the web.

While the graph database model supports the ability to store definitions alongside categories, in some cases the definitions of the categories were unavailable and therefore could not be stored. **Therefore, the categories currently stored in the database are subject to ambiguity and are not semantically stable.** For the graph model to better support achieving semantic interoperability, the definitions of the categories must be available. The consequence of this results in semantic mismatches between terms, or multiple meanings for similar terms. For instance, the element bovine animals may include buffaloes but in other cases may not. Therefore, when data is compared or combined it is subject to potential misuse.

4.2. THE DATASPHERE AND RECOMMENDATIONS FOR INTEROPERABILITY BY DESIGN

The concept of the Datasphere, which is defined by La Chapelle and Porcincula (2022) as "*the complex system encompassing all types of data and their dynamic interactions with human groups and norms*", provides a vantage point for demonstrating the need for an exploration of the complex systems surrounding the creation and use of standards for livestock population reporting [29]. 338 unique species categorizations among the four data sources used in the case study illustrates that even among international and national organizations, a common standard is difficult to employ. The multi-faceted stages in the creation of standards including legal documents providing supporting documentation support the need for supporting documents to be interoperable with data sources and related metadata, as well as discoverable alongside the data resources. Here, we provide a discussion on the complex systems defining data standards, and their interactions to provide recommendations for improved interoperability in livestock data.

4.2.1 STAKEHOLDERS AND GOVERNANCE STRUCTURES IN NAMING CONVENTIONS OF LIVESTOCK DATA

Each data source employed their own standard naming conventions; in some cases (as was the case for WOAH), these standards changed across time presumably due to changes in the needs of the granularity of data categories. While standards are necessary to facilitate the interoperability of data and subsequent reuse within and between sectors, they must be designed in an inclusive manner to prevent the unintended exclusion of categories that may be important in cultures outside of the western-centric lens.

In addition, another core aspect in the creation of standards and classifications of livestock data from the country-level include laws and governance structures that regulate how statistics are reported at a national level; statistical laws may be present both at the national and regional level where they mandate the categories of data to be collected and the frequency of collection. For instance, member states of the European Commission are required by law to report livestock statistics to EuroStat biannually. The categories for livestock statistics are outlined and defined in Regulation (EC) No 1165/2008 Annex I and II [30]. However, WOAH is an Intergovernmental Organization comprising 182 Members, where each Member has an "obligation to submit information on their animal health situation" [31]. In addition, National Authorities from Members are obligated to submit annual reports on animal populations among other disease information. As one of the main goals of WOAH is to collect data related to disease information, classifications of livestock data are likely reflective of the organization's internal data needs. Members may also be mandated to collect data using national classifications where categories and definitions may be controlled by statistical acts or agricultural laws, as is the case for EU member states.

These national classifications may differ from those required by WOAH resulting in the need to map categories to combine or integrate data from WOAH and other data sources.

The governance structures within organizations and national statistics agencies mandate the reporting structures, and thus classifications and terms used in livestock population data. Therefore, policies are directly influencing the semantic interoperability of resultant data. These governance structures must be considered when seeking to improve the discoverability and interoperability of data. If each actor is mandating different classification systems and naming conventions for data, this creates data silos and it is much more difficult to find shared meaning between datasets that are being collected for similar or shared purposes ultimately impacting the ability to reuse, combine, and leverage data for societal good.

4.2.2 RECOMMENDATIONS FOR INTEROPERABILITY BY DESIGN TOWARDS A COLLECTIVELY GOVERNED DATASPHERE

The complex and dynamic nature of the naming conventions used is exhibited in the analysis through the heterogeneity in categories used to represent semantically similar categories; for interoperability to succeed, a bottom-up approach could be used to ensure interoperability standards are used at each stage in the data's life cycle. In this sense, our recommendations for the key stakeholders in the creation of standards - which include policy-makers, internal committees, government statistics offices, and those involved in the digitization and dissemination of data - include:

1. **Before enforcing new standards, or suggesting the use or creation of ontologies or other semantic mechanisms to improve the interoperability and subsequent reuse of data, *determine how governance structures impacting the enforcement of standards in data from national surveys and censuses might differ between different countries and organizations.***

   Learning about how governance structures impacting the enforcement of standards operate within a country, and between different countries will allow for an understanding of how standards can be operationalized. For instance, enforcement of new standards may not consider how current standards reflect cultural categories for livestock (for instance, population of livestock used for offerings) that may not be consistent across countries. In addition, the way in which countries collect, store, and disseminate statistics may differ between countries. An understanding of the governance structures between countries and organizations can therefore be used to improve the uptake of new standards or recommendations, and/or improve standards to be more inclusive of the unique needs of different countries and organizations.

2. Provide a root-cause of difficulties reported in interoperability and failures or non-coherence to internationally developed standards:

Providing a root-cause analysis of difficulties reported in interoperability and failures or non-coherence to internationally developed standards by either research communities or organizations such as the FAO could provide insight on the systems around the creation of standards and stakeholder roles in facilitating interoperability and discoverability of data at each level of the data life cycle. For instance, categories are used in data collection forms or questionnaires, methodology for collection, legal and policy regulations including statistical and agricultural acts, digitization of data, and dissemination of data; data categories are also introduced during the aggregation of national data and depending on the reason by which the data is collected. For standards to succeed in enabling interoperability and reuse of data outside of the original purpose of collection, each stage must be interoperable with the other which will require a multi-stakeholder approach.

3. Legal documents and policies reporting definitions for categories should be interoperable:

Given that legal documents and policies may mandate and provide definitions of the categories used in livestock population reporting, these documents should be machine-actionable and linked to datasets to allow for the semantics to be available. The documents themselves and definitions can then be used to determine whether data sets are interoperable. The graph model presented could be expanded to include linkages to these documents and allow this information to be available to data users.

Exploring the classification of data from a bottom-up approach has demonstrated the need for a more in-depth analysis of classifications for livestock data at a national scale. The heterogeneity that exists from the international or regional organizations that were analyzed demonstrates the difficulties in adhering to a standard in the sector, perhaps due to the lack of a common standard in the sector, but also aligns to other claims that adherence to standards does not ensure interoperability [32]. To make suggestions to improve the interoperability of the categorizations of data, the governance structures involved in the creation of these categories must be considered.

REFERENCES

[1] R. Goodland and J. Anhang, "Livestock and climate change: what if the key actors in climate change are... cows, pigs, and chickens?," Livest. Clim. Change What Key Actors Clim. Change Are Cows Pigs Chick., 2009, Accessed: Dec. 28, 2022. [Online]. Available: https://www.cabdirect.org/cabdirect/abstract/20093312389

[2] B. Catry, H. Laevens, L. a. Devriese, G. Opsomer, and A. de Kruif, "Antimicrobial resistance in livestock," J. Vet. Pharmacol. Ther., vol. 26, no. 2, pp. 81–93, 2003, doi: 10.1046/j.1365-2885.2003.00463.x.

[3] M. Herrero et al., "Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems," Proc. Natl. Acad. Sci., vol. 110, no. 52, pp. 20888–20893, Dec. 2013, doi: 10.1073/pnas.1308149110.

[4] B. Huntington et al., "Global Burden of Animal Diseases: a novel approach to understanding and managing disease in livestock and aquaculture," Rev. Sci. Tech. Int. Off. Epizoot., vol. 40, no. 2, pp. 567–584, Aug. 2021, doi: 10.20506/rst.40.2.3246.

[5] "Data dissemination | Food and Agriculture Organization of the United Nations." https://www.fao.org/statistics/databases/en/ (accessed Dec. 28, 2022).

[6] T. Benson, "Sharing Data," in Patient-Reported Outcomes and Experience: Measuring What We Want From PROMs and PREMs, T. Benson, Ed. Cham: Springer International Publishing, 2022, pp. 67–83. doi: 10.1007/978-3-030-97071-0_7.

[7] V. Méndez Muñoz et al., "Analysis on the Graph Techniques for Data-mining and Visualization of Heterogeneous Biodiversity Data Sets," in Complexis 2017, Porto, Portugal, Apr. 2017, pp. 144–151. doi: 10.5220/0006379701440151.

[8] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Sci. Data, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18.

[9] A. Zezza and C. Azzarri, "INVESTING IN THE LIVESTOCK SECTOR Why Good Numbers Matter", Accessed: Dec. 28, 2022. [Online]. Available: https://www.academia.edu/21032972/INVESTING_IN_THE_LIVESTOCK_SECTOR_Why_Good_Numbers_Matter

[10] C. Bahlo, P. Dahlhaus, H. Thompson, and M. Trotter, "The role of interoperable data standards in precision livestock farming in extensive livestock systems: A review," Comput. Electron. Agric., vol. 156, pp. 459–466, Jan. 2019, doi: 10.1016/j.compag.2018.12.007.

[11] "ISO 11788-2:2000," *ISO*. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/02/64/26400.html (accessed Sep. 15, 2022).

[12] "Livestock Data Interchange Standards Community Group," *W3C Community & Business Groups*. https://www.w3.org/community/livestockdata/ (accessed Sep. 15, 2022).

[13] "Agricultural Metadata Element Set (AgMES) | AIMS." http://aims.fao.org/standards/agmes (accessed Sep. 15, 2022).

[14] "Ontologies for Agriculture," *CGIAR Platform for Big Data in Agriculture*. https://bigdata.cgiar.org/ontologies-for-agriculture/ (accessed Sep. 15, 2022).

[15] L. M. Hughes, J. Bao, Z.-L. Hu, V. Honavar, and J. M. Reecy, "Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species," *J. Anim. Sci.*, vol. 86, no. 6, pp. 1485–1491, Jun. 2008, doi: 10.2527/jas.2008-0930.

[16] "Home - Taxonomy - NCBI," *National Library of Medicine, National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/taxonomy (accessed Sep. 15, 2022).

[17] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, Inc., 2015.

[18] "What is a Graph Database? - Developer Guides." https://neo4j.com/developer/graph-database/ (accessed Sep. 15, 2022).

[19] R. Renner and G. Jiang, "Challenges in Using a Graph Database to Represent and Analyze Mappings of Cancer Study Data Standards," AMIA Summits Transl. Sci. Proc., vol. 2020, pp. 517–526, May 2020.

[20] N. Swainston et al., "biochem4j: Integrated and extensible biochemical knowledge through graph databases," PLOS ONE, vol. 12, no. 7, p. e0179130, Jul. 2017, doi: 10.1371/journal.pone.0179130.

[21] A. Debrouvier, E. Parodi, M. Perazzo, V. Soliani, and A. Vaisman, "A model and query language for temporal graph databases," VLDB J., vol. 30, no. 5, pp. 825–858, Sep. 2021, doi: 10.1007/s00778-021-00675-4.

[22] "FAOSTAT." https://www.fao.org/faostat/en/#data (accessed Feb. 02, 2023).

[23] "Web Services - Eurostat." Accessed: Feb. 02, 2023. [Online]. Available: https://ec.europa.eu/eurostat/data/web-services

[24] "Home - WOAH - World Organisation for Animal Health." https://www.woah.org/en/home/ (accessed Feb. 02, 2023).

[25] "Ethiopian Statistics Service." http://www.statsethiopia.gov.et/ (accessed Feb. 02, 2023).

[26] "Ethiopian Case Study – GBADS – Global Burden of Animal Diseases." https://animalhealthmetrics.org/case-study-ethiopia/ (accessed Feb. 02, 2023).

[27] “Web Services - Eurostat.” https://ec.europa.eu/eurostat/data/web-services (accessed Feb. 02, 2023).

[28] “SURVEY REPORTS.” http://www.statsethiopia.gov.et/our-survey-reports/ (accessed Feb. 06, 2023).

[29] L. Porciuncula, and B. D. L. Chapelle, “Hello Datasphere — Towards a systems approach to data governance,” The Datasphere Initiative, Feb. 28, 2022. https://www.thedatasphere.org/news/hello-datasphere-towards-a-systems-approach-to-data-governance/ (accessed Dec. 28, 2022).

[30] Regulation (EC) No 1165/2008 of the European Parliament and of the Council of 19 November 2008 concerning livestock and meat statistics and repealing Council Directives 93/23/EEC, 93/24/EEC and 93/25/EEC (Text with EEA relevance). 2014. Accessed: Sep. 15, 2022. [Online]. Available: http://data.europa.eu/eli/reg/2008/1165/2014-01-10/eng

[31] “World Organization for Animal Health: Who we are.” https://www.woah.org/en/who-we-are/ (accessed Sep. 15, 2022).

[32] S. Madnick and H. Zhu, “Improving data quality through effective use of data semantics,” Data Knowl. Eng., vol. 59, no. 2, pp. 460–475, Nov. 2006, doi: 10.1016/j.datak.2005.10.001.

APPENDIX A: EXAMPLE GRAPH OUTPUTS

This Appendix provides example outputs to queries articulated in Section 3.4. Queries. All outputs are provided in table form since this is the output provided by Cypher.

Q2: Given a category/livestock species, what are the data sources that report that category and for which years?

Example query:

```
MATCH (n:Category)-[r]-()-[]-(d:Datasource)
WHERE n.name =~ '(?i)Chick.*'
RETURN n.name, d.name, r.year
```

The output to the query using Chick as the example species is presented in Table 2.

n.name	d.name	r.year
Chickens	FAOSTAT Production: Crops and livestock products	[1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020]
Chicks used for laying	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012]
Chicks of turkey (fattening)	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021]

Chicks of mixed meat-laying breeds	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022]
Chicks of meat broiler breeds (selection)	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022]
Chicks of meat broiler breeds (fattening)	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021]
Chicks of laying hen breeds (selection)	EuroStat	[1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022]
Chicks of laying hen breeds (laying)	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021]

Chicks of goose (fattening)	EuroStat	[1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021]
Chicks of duck (fattening)	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021]
Chicks of Guinea fowls (fattening)	EuroStat	[1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022]

Table 2: Example output with ‘Chick’ as <name_of_category> for Q2.

Q3: Which countries report a given livestock species and for which years?

The output to the query using Chicken and year_Stocks as they year property as the example species is presented in Table 3. Given the number of countries that report chicken population (241 countries), we provide only a subset of the example output.

The year properties were determined using:

```
MATCH (n:Category)-[r]-(a:Area)
WHERE n.name =~ '(?i)Chicken.*'
RETURN DISTINCT keys(r)
```

Resulting in this query:

```
MATCH (n:Category)-[r]-(a:Area)
WHERE n.name =~ '(?i)Chicken.*'
RETURN n.name, a.name, r.year_Stocks
```

n.name	a.name	r.year_Stocks
Chickens	Zimbabwe	[1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020]
Chickens	Zambia	[1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020]
Chickens	Yugoslav SFR	[1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991]

Table 3: Subset of example output from Q3 using Chicken as the <name_of_category> and year_Stocks as the year property.

Q4: For a given country and livestock species, which datasets are available and which years does the country report that species to each data source?

The output to the query using Goat as the <name_of_category> and Canada as the country is presented in Table 4.

The year properties were determined using:

```
MATCH (n:Category)-[r]-(a:Area {name: 'Canada'})
WHERE n.name =~ '(?i)Goat.*'
RETURN DISTINCT keys(r)
```

Resulting in the query:

```
MATCH (n:Area {name: 'Canada'})-[r]-(c:Category)-[]-()-[]-(d:Datasource)
WHERE c.name =~ '(?i)Goat.*'
RETURN d.name, r.year_WOAHpopulation, r.year_Stocks, r.year_UNFCCC, r.year_FAOTIER1
```

Q5: What is the definition of a given species category?

Example query:

```
MATCH (n:Category {name: ‘Chickens’})
RETURN n.definition
```

Q6: For a given data source, what are the associated licences or access restrictions?

Example query:

```
MATCH (n:Datasource {name: ‘FAOSTAT Enteric Fermentation’})
RETURN n.license
```

Q7: Return the number of unique categories stored in the database

The output of Q7 is provided below:

COUNT(n)
338

Table 7: Example output from Q7.

Q8: Given x named countries, return the common species categories

The output to the query using Canada and Ethiopia as

```
WITH ['Canada', 'Ethiopia'] as names
MATCH (a:Area)
WHERE a.name in names
WITH collect(a) as areas
MATCH (c:Category)
WHERE ALL(a in areas WHERE (c)-[:REPORTED_BY]-(a))
RETURN c.name
```

c.name
Camelidae
Dogs
Equidae
Layers
Other commercial poultry
Sheep / goats
Camels
Cattle
Cattle, dairy
Cattle, non-dairy
Goats
Horses
Mules and Asses
Mules and hinnies
Sheep
Sheep and Goats
Swine

Swine, breeding
Swine, market
Eggs Primary
Eggs, hen, in shell
Cattle and Buffaloes
Chickens
Mules
Pigs
Poultry Birds
Milk, Total
Milk, whole fresh cow
Milk, whole fresh goat
Beef and Buffalo Meat
Fat, cattle
Fat, pigs
Fat, sheep
Hides, cattle, fresh
Meat, Poultry
Meat, cattle
Meat, chicken
Meat, pig
Meat, sheep
Offals, edible, cattle
Offals, pigs, edible

Offals, sheep,edible
Sheep and Goat Meat
Skins, sheep, fresh
Bees
Birds

Table 8: Example output from Q8 using Ethiopia and Canada as example countries.

thedatasphere.org