GBADs Poultry Data Methodology Report

Kassy Raymond*[1] and Deborah Stacey†[1]

[1]School of Computer Science, University of Guelph, Canada

December 2019

1 Introduction

Creating a data inventory and ecosystem is complex. GBADs has conducted a brief pilot study using one production species and one country to explore the opportunities and challenges in the creation of a *"dataverse"* (an open and accessible repository of data, links to data, and associated metadata) capable of creating the foundation for modelling of items such as health loss envelopes.

Selecting poultry as an example, we used Ethiopia to create a pilot study of open data availability. The main goals of this pilot were as follows:

- Determine the types and formats of data available on international, regional, and global levels.
- Determine key stakeholders in the data available including information about who is collecting different types of data, data ownership.
- Explore methodology in data organization for easy identification of differences between data types and gaps in data.
- Present findings in a visualization.

As this pilot study was done under time constraints, the difficulty of obtaining the data provided in the inventory is also provided.

2 Methodology

This pilot data collection focuses on the what and the how. Here, we will outline how we searched for the data in a methodological, well thought-out fashion, why it was important and the barriers encountered.

*kraymond@uoguelph.ca
†dastacey@uoguelph.ca

2.1 THE HOW

1. **How the Pilot was Conducted**

   The data inventory was constructed using guides from the Open Data Institute (ODI). Specifically, a data vocabulary was used to capture the main information provided from the data sets found. This included information regarding data ownership, relevance to GBADs, data accessibility (unpublished, published, open, API or not available), whether there was documentation on the data collection, sources and methodology and to what degree this information was available, keywords, main subject matter of data (population/productivity/health/welfare/environment and price), category (sector/global/national/project), data availability (in years) and whether data were missing.

   Any data related to the poultry industry provided by national, global and regional sources were included in the inventory.

   A data visualization was created that summarized and organized the data found in the search and the trail used to find this data.

2. **Resources Used**

   Data collection of the factors involved in modelling the economic impact of animal disease, production techniques, climate change, demographics, etc. on food production based on livestock, is complex and undertaken at many different levels: global agencies, national agencies and governments, industry, and local producers. Surveying this landscape is a necessary first step in the development of a semi-automated system to monitor, access, and utilise the data. This pilot is that first step and was conducted with basic resources - computers, internet access, and human expertise. The multi-disciplinary composition of the pilot team is indicative of the type of talent that GBADs will have to assemble to successfully scale up this effort. The team consisted of academic personnel in veterinary science, epidemiology, computer science, and public policy. This academic team was assisted by sector partners and experts familiar with the sector under examination. The total number of hours to collect, access, and present the dataverse for poultry in Ethiopia was remarkably small (approximately 30 total hours) for the work accomplished.

3. **Barriers to Data Collection and Assessment**

   The data resources openly available on the Internet and data obtained from private sources have many differences in format, accessibility, quality, and semantics. This exercise highlighted the need for many activities that will have to happen to facilitate GBADs. The data mapping and other data products and tools that will need to be produced will not just be useful to GBADs; they will provide added value to all the communities that use the same data sources and will encourage a community of shared standards and practices.

   Barriers and challenges that have been identified to date include:

- Inconsistencies in terminology between data resources. This points out the need for the adoption and development of vocabularies, ontologies, and data maps that can apply across a wide selection of data sources, both open and private.
- Difficulty assessing the quality of the data - there is a need for *standardization* of methodologies and for the assessment of such methodologies. This would involve the collection of methodologies from all key data stakeholders and the creation of a unique standardized method or ranking system that encapsulates the quality of the data. This can take inspiration from similar work by the World Bank and their quality score. And while part of this quality assessment is rooted in statistics and data science, there is obviously a sensitive political and cultural aspect to assessment that must be handle with sensitivity to all stakeholders from the global to the local level.
- Some of the data found was not open and were only available through a paid membership. This will require a collective effort by the GBADs community to encourage private partners to find a way to share their data in a way that does compromise their own mission and objectives.

THE WHY

1. **Scaling the Methodology**

   A pilot study can help in identifying the challenges to scaling the current methodology to attack the larger problem that GBADs presents. One of the results of the pilot study is the identification of a number of large, open, global resources that have (or will have) mechanisms (i.e. APIs) for access. This indicates that it will be possible to access many data sources without the need to duplicate their storage. *”Just-in-time”* access to data in conjunction with data mapping and quality assessment will allow GBADs to assist stakeholders in the *”collection”* of the data that they need, in the format that they need it in, when they need it for their analysis. It will also allow GBADs to ”schedule” regular reports that will form a set of basic GBADs products.

2. **Finding Gaps and Barriers**

   Identifying barriers is important for understanding how to create a complex multi-variate data structure and assessing the current data. The main gaps and barriers of the data were identified after the data inventory was created. This included comparing the number of data sources found for each main category of data (price, productivity, poultry population, environmental data, human demographics, disease and trade) to determine where resources were lacking. Data sources providing information about disease incidence are lacking; the World Organization for Animal Health (OIE) is currently the only source of data for this category. The data sources for this data are also unclear; this highlights the need to seek sector data that may be more rich in disease data. For example, the National Veterinary Institute of Ethiopia is a vaccine company that may hold valuable information about who is purchasing vaccines and which areas in Ethiopia that these are being distributed to.

There is also a lack of *species-specific* data. While poultry population statistics are available through international, national and regional sources, these numbers seldom include separate information about species or type of chicken included in these values. Further, chicken mortality information is lacking at all levels and when available does not provide information about the age or stage of poultry at death. This information is pertinent in understanding the epidemiology of poultry disease and drawing inferences about common causes of death and the burden of disease.

3. **Data Quality Assessment**

   After examining the data found at different sources and at different levels, it is obviously that there will be a need to assess the quality of the data. Agencies that collect data often provide this kind of information about their data. An example of this type of quality evaluation is the the World Bank's Statistical Capacity Indicator. It is a composite score assessing the capacity of a country's statistical system. It is based on a framework assessing that assess areas such as methodology, data sources, periodicity and timeliness. Many criteria are evaluated in these areas resulting in a Statistical Capacity score on a scale from 0-100. GBADs will need to develop its own set of methodologies and quality metrics to accommodate the need to compare and model situations where data from varied sources are used in combination. Data quality metrics will form the basis for evaluating the risk of using some data in some models.

4. **Identifying Data Flows**

   In open data, the main data flow is from national to regional data sources, and from regional to international sources. The international data sources were the most robust, using data imputation techniques to fill in any missing data. Using the poultry population in 2012 and 2015 as an example, a data flow exercise compared available data from the national to the international level. In 2012, the Federal Agricultural Organization (FAO) reported the total chicken population in Ethiopia to be 40 377 000. This value differs from the amount report by the Central Statistics Agency of Ethiopia (44 893 009). The FAO reported that the total chicken population in 2015 was 60 506 000. The Central Statistics Agency of Ethiopia did not provide any open poultry data for 2015 however, the FAO reported their statistics as 'Official Data', indicating that data were provided from an external source (data was not imputed). This preliminary data flow comparison reveals that while national sources may provide data to international sources, there are clear discrepancies in reported numbers. To perform a more in-depth data flow analysis, access to sector and project data would provide insight on the data quality and trustworthiness of other open data sources.

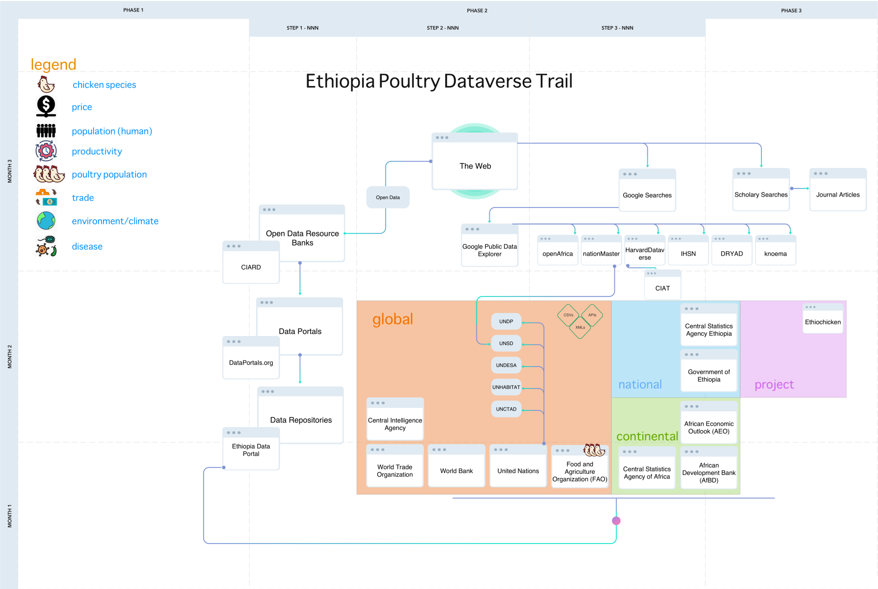
5. **Identifying Partnerships**

   At the sector level, open data is currently not publicly available. Through data flow exercises and collaboration with Ethiochicken, there is clear value in establishing partnerships in the sector.

   Sector data found included qualitative information about the key data stakeholders in this category. This included finding producer companies that have a large presence in the poultry industry in Ethiopia. These companies may keep meticulous records of sales of both poultry and feed which can provide useful information for validation and extrapolation of other data sources. Further, creating partnerships with these companies can enlighten GBADs on issues in the sector that may not be obvious through studies such as the current. Determining active associations and institutes is also important in terms of creating contacts with farmers that may be knowledgeable about the sector in the country of interest.

   Our collaborations with Ethiochicken have given us invaluable information about the need for data that individuals can use and trust.

6. **A Preliminary Dataverse for Poultry and Ethiopia**


PHASE 1
PHASE 2
PHASE 3
STEP 1 - NNN
STEP 2 - NNN
STEP 3 - NNN
MONTH 3
MONTH 2
MONTH 1
legend
chicken species
price
population (human)
productivity
poultry population
trade
environment/climate
disease
Ethiopia Poultry Dataverse Trail
The Web
Open Data
Google Searches
Scholary Searches
Journal Articles
Open Data Resource Banks
CIARD
Google Public Data Explorer
openAfrica
nationMaster
HarvardDataverse
IHSN
DRYAD
knoema
CIAT
Data Portals
DataPortals.org
Data Repositories
Ethiopia Data Portal
global
UNDP
UNSD
UNDESA
UNHABITAT
UNCTAD
CSVs
APIs
XMLs
Central Intelligence Agency
World Trade Organization
World Bank
United Nations
Food and Agriculture Organization (FAO)
national
Central Statistics Agency Ethiopia
Government of Ethiopia
continental
African Economic Outlook (AEO)
Central Statistics Agency of Africa
African Development Bank (AfBD)
project
Ethiochicken