GBADs informatics strategy, data quality, and model interoperability

K. Raymond* [1, 2], K.E. Sobkowich [1, 3], J.D. Phillips [1, 2, 3], L. Nguyen [1, 2], I. McKechnie [1, 2], R.N. Mohideen [1, 2], W. Fitzjohn [1, 2], M. Szurkowski [1, 2], J. Davidson [1, 2], J. Rushton [1, 4], D.A. Stacey [1, 2] & T.M. Bernardo [1, 3]

(1) Global Burden of Animal Diseases (GBADs) Programme, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, 146 Brownlow Hill, Liverpool, L3 5RF, United Kingdom (https://animalhealthmetrics.org)

(2) School of Computer Science, University of Guelph, Reynolds Building, 474 Gordon Street, Guelph, ON N1G 2W1, Canada

(3) Department of Population Medicine, Ontario Veterinary College, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada

(4) Department of Livestock and One Health, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, 146 Brownlow Hill, Liverpool, L3 5RF, United Kingdom

*Corresponding author: kraymond@uoguelph.ca

Summary

The estimation of the Global Burden of Animal Diseases (GBADs) requires the integration of multidisciplinary models: economic, statistical, mathematical, and conceptual. The output of one model often serves as input for another; therefore, consistency of the model components is critical. The GBADs Informatics team aims to strengthen the scientific foundations of modelling by creating tools that address challenges related to reproducibility, as well as model, data, and metadata interoperability. Aligning with these aims, several tools are under development:

1) GBADs’ Trusted Animal Information portaL (TAIL), a data acquisition platform that enhances the discoverability of data and literature and improves the user experience of acquiring data. TAIL leverages advanced semantic enrichment techniques (natural language processing and ontologies) and graph databases to provide users with a comprehensive repository of livestock data and literature resources.

2) The interoperability of GBADs’ models is being improved through the development of an R-based modelling package and standardising parameter formats. This initiative aims to foster reproducibility, facilitate data sharing, and enable seamless collaboration among stakeholders.

3) The GBADs’ Knowledge Engine is being built to foster an inclusive and dynamic user community by offering data in multiple formats and providing user-friendly mechanisms to garner feedback from our community.

These initiatives are critical in addressing complex challenges in animal health and underscore the importance of combining scientific rigour with user-friendly interfaces to empower global efforts in safeguarding animal populations and public health.

Keywords

Data management – Data quality – Data science – Global Burden of Animal Diseases – Interoperability – Reproducibility.

Introduction

Advances in scientific modelling have bolstered our understanding of the past and our ability to predict future outcomes. Models, simplified images, or representations of reality, while not perfect, can improve our understanding of complex problems, such as climate change, population dynamics, or the spread of a novel virus. In his book, *How to Prevent the Next Pandemic*, Bill Gates says ‘For me, one of COVID’s chief lessons about modeling is the extent to which every model relies on good data, and just how hard it can be to get that data.’ [1].

The estimation of the Global Burden of Animal Diseases (GBADs) incorporates a number of models (economic, statistical, mathematical, and conceptual). These models use data and parameters from diverse sources to create outputs, where the output of one model may serve as input for another. Hence, there is a need for the interoperability in parameters and data that are used across the various models of GBADs. For example, the estimation of the number of animals at a given point or over a period of time is required for calculating both biomass and total economic value.

With the overarching goal of enhancing the discovery and interoperability of data, parameters, and models used and created by GBADs modellers and stakeholders, the

GBADs' information strategy consists of digital informatics tools, designed to support the following core objectives:

1) to make it easier to access data;
2) to ensure downloaded data are available in a usable format, that works with modelling tools that GBADs designed;
3) to improve the robustness, usability, and modularity of modelling tools;
4) to allow members of the community to provide feedback on the functionality, usability, and impact of the system.

These objectives are aligned with key findings highlighted in a review by Antle *et al.* [2], which underscores the importance of improving agricultural data and models through fostering interoperability among models, linking them to data and knowledge products, and implementing cloud-based data management and retrieval systems. Other initiatives such as 50×2030 and CGAIR's GARDIAN platform provide or index agricultural data that may be of use as inputs to GBADs models [3,4]. GBADs builds upon the broader landscape of these initiatives by aiming to improve the quality and interoperability of models, data, and parameters, and facilitating the retrieval and management of these resources.

The GBADs Informatics team, with expertise in veterinary medicine, epidemiology, computer, and data science, conducted a survey to gather insights from successful organisations that operate and sustain global health data gathering systems [5]. Based on these findings, a key recommendation was to spend more time and resources upfront in creating robust and data systems for the future. In line with this advice, we are using cutting-edge technologies to ensure our system is scalable and future proof. This includes incorporating ontologies for more effective searching, mapping relationships using a graph database, and incorporating proven machine learning and Natural Language Processing (NLP) tools into our data gathering system.

This paper aims to present the approaches and tools that support each of the four main objectives, with the ultimate goal of improving usability, interoperability, reusability, and consistency across GBADs models, data, and dashboards. First, we introduce the Trusted Animal Information portaL (TAIL), a system designed to enhance the interoperability, discoverability, and access to data and literature within the GBADs Knowledge Engine. Next, we describe our approach for improving the interoperability, reusability, and reproducibility of models and their outputs through the modularisation of code and standardisation of parameters. Finally, we outline the various tools being developed to gather feedback, engage stakeholders, and quickly incorporate feedback into dashboards. Table I provides a comprehensive overview of all tasks and progress related to each tool discussed.

Improved understanding and streamlined acquisition of livestock data

Livestock data, crucial for informed decision making and policy formulation, exists in heterogeneous formats, originates from disparate sources, and uses inconsistent terminology for characterising species, often hindering the ability to attain the FAIR guiding principles (Findability, Accessibility, Interoperability, and Reusability) [6]. To address these issues, we present the GBADs TAIL (Figure 1), a novel system designed to facilitate the acquisition and interoperability of livestock data and literature from diverse sources. This section describes the preliminary system architecture, placing emphasis on the system's ability to reconcile data and metadata inconsistencies, and to enhance semantic search of data beyond traditional keyword queries.

GBADs TAIL: discovering metadata and literature

Consider a user who is interested in finding cattle population data. The user enters the GBADs TAIL and enters a free-text query such as 'cow population in Canada'. The query is sent to the GLoVe (Global Vectors for Word Representation), a NLP software from Stanford University, which recognises terms in the query [7]. The term that is recognised as the species is sent to our ontology, which enriches the term with semantically related terms such as 'cattle', 'bovine', or 'dairy'. The user can then select which of the additional terms to include, resulting in a 'semantically enriched query'.

The semantically enriched query is sent to a graph database, which holds metadata for all data that has been catalogued by GBADs. By matching the terms to those connected to metadata in the graph database, a catalog of relevant metadata are provided to the user, allowing the user to browse for datasets relevant to their search. In parallel, the augmented query is sent to Semantic Scholar, returning literature related to the user's search [8,9].

Finally, the information is displayed on the GBADs TAIL interface, providing users with a catalog of data and literature resources related to their search. The front-end interface is being built with React, a popular framework for dynamic interface development and will eventually integrate knowledge about GBADs models into the portal, creating a truly contiguous network of accessible data resources [10].

Table I outlines the progress made in the ongoing development of the GBADs TAIL interface. Livestock population data from the World Organisation for Animal Health (WOAH), Eurostat, the Food and Agriculture Organization of the United Nations Statistical Database (FAOSTAT), and the Ethiopia Central Statistics Agency is being made accessible in the primary build of the system.

The technical underpinnings

To reconcile the inconsistencies in standardisation, terminology, and content of livestock data and metadata, we present the architecture of a sophisticated back-end system. Using a graph database and ontologies, metadata and data are semantically enriched and made interoperable, allowing for efficient data and literature searches.

Graph database metadata catalogue

The graph database that is employed by the system stores metadata for each of the datasets catalogued by GBADs. Metadata are defined broadly as 'data about data', providing information such as descriptions, temporal and spatial coverage, licensing and ownership information, date of last update, and access instructions [11]. To streamline the efficient discovery and accessibility of data, all metadata are reconciled into a standard format (schema.org) and stored in a graph database (Figure 2) [12]. For example, a user interested in data related to a particular country and species can query the graph database to retrieve all related datasets and their related metadata.

Ontologies

Recognising that datasets are characterised by heterogeneous terminology, searching for data based on explicit keywords fails to retrieve all relevant datasets. Explicit keyword search assumes that users know precisely what they are looking for and does not retrieve semantically similar terms. For instance, in many other data portals, a user searching for data using the keyword 'cattle', would not also retrieve data about 'dairy cattle', 'cows', 'calves', and 'bovine'. We employ a bottom-up approach and use species terms found in data sources that may have varied meanings. Ontologies capture the semantics of data, allowing semantically similar, albeit heterogeneously described terms to be identified.

In our system, ontologies provide vocabularies to describe our datasets, and help to establish clear and shared definitions of concepts, and their associated relationships. By identifying shared definitions and mapping similar concepts, ontologies can be operationalised to enable users to enrich their query with similar terms that may be defined differently. This relieves users from the time consuming, and error-prone task of manually seeking similar terms and their respective definitions themselves.

Once similar terms are identified, these terms can be used to search for resources that are described using different, yet semantically similar, terms. The result yields relevant results that were previously in siloes due to heterogeneity in characterisation, ultimately leading to more FAIR data.

Data quality: internal and external consistency

Beyond data acquisition, assessing data quality is a fundamental task. Data quality is a concept that encompasses a number of dimensions including internal and external consistency [13,14]. External consistency refers to the synchronicity of two datasets reporting the same measurements [13]. For instance, comparing livestock population data reported by FAOSTAT and a National Statistical Organisation offers an indicator of consistency.

To facilitate data quality assessments and reproducibility, we developed a data visualisation tool (Figure 3). This tool allows users to investigate and identify inconsistencies in livestock population datasets from WOAH, FAOSTAT, and national agricultural censuses and surveys. By cross-referencing data between countries, species, and over time, probable inconsistencies between and within datasets have been illuminated [17]. Consequently, datasets can be fully annotated, allowing inconsistencies to be communicated and providing users with an indicator of quality. This systematic approach to data quality reduces the need for data quality to be assessed independently by each modeller, ultimately improving the reproducibility of modelling efforts in GBADs.

Interoperability of models

At the core of GBADs are a set of interconnected animal health and economics models and calculations used for deriving metrics such as the Dynamic Population Model (DPM). Briefly, the DPM is a novel compartmentalised equation-based model that collates information from a variety of data sources (e.g. milk, meat, waste), and at various animal life stages (e.g. neonate, juvenile, adult), to estimate monetary losses. Specifically, output from the DPM is employed to calculate the Animal Health Loss Envelope (AHLE) which is the difference between ideal and current scenarios.

Often, models are coded in programming languages such as R, and use data and parameters. From these components, output data are produced, such as the economic cost of a particular disease on a production system, which can inform policy-decisions. The DPM includes several calculations such as biomass and economic value. By decomposing the model into pieces (modularising the code), components of the model can become interoperable (i.e. be exchanged seamlessly between different programmes). The implications of interoperable models are two-fold. Firstly, individual components of the model can be reused in analyses beyond those imagined for the original task. Secondly, models can be rerun using updated data or parameters without having to build the model from scratch, leading to reproducible and replicable methods [18].

The interoperability of models is therefore essential for the uptake of GBADs methods and to validate and verify our methods. To facilitate interoperability of models, we are developing a package using R and standardising the format and naming conventions of parameters used for modelling. To this end, we envision the development of a user-friendly analysis tool that will permit the chaining of models (encompassing both parameters and data) together in an exchangeable and automatic fashion. The ultimate goal is to improve data sharing capabilities among academics, scientists, health professionals, stakeholders, and policy-makers within the larger purview of an international One Health datasphere [19-21]. In addition, the interoperable design of the modelling pipeline provides an opportunity to use the outputs for a variety of analytical treatments including sensitivity analysis.

R package for GBADs models

For several reasons R was chosen for the modularisation of code, which will result in a GBADs R package (i.e. an ensemble of interoperable modelling components). R is a free, open-source software, encouraging the dissemination of collaborative knowledge and research findings through its extensive ecosystem of user-contributed packages [22]. Supported by a large user base, including its own online journal and academic conference, R packages can be created and deployed by the user community through submission to the Comprehensive R Archive Network (CRAN) [23]. CRAN enforces rigorous standards and policies, with packages being checked daily to ensure operability across multiple major operating systems. Once packages are submitted to CRAN, any

R user can download and use the package freely for their own work, encouraging collaboration. To this end, an R package invites other researchers to use GBADs models and their components, allowing GBADs methods to be adopted by those beyond our direct community.

Interoperability of parameters and modelling outputs

To ensure that parameters can be understood and be reused among models (or components of models), standard formats and naming conventions must be used. In addition, standardised output formats ensure that results from one model can be used as inputs to others (and thus allowing for chaining of models). Work is underway to collect input parameters from existing models, such as the AHLE, and to solicit other parameters that may be useful in future models. This work will provide a dictionary of commonly understood parameters, allowing for a common understanding and standardised naming of parameters used in GBADs models. In addition, the goal is to develop a format for parameters in YAML (Yet Another Markup Language; a markup language that is human and machine readable and is often used for writing configuration files) [24]. To facilitate the largest number of users, parameters described in CSV or JSON will be readily translatable by our shared input packages to the standardised YAML format for use in the model. This standardisation work will facilitate the collection and documentation of model parameters that can be widely shared. It will also generate a wealth of valuable metadata which can serve as an authoritative health informatics resource for GBADs' personnel and other interested parties.

Capacity and community building

Community building amongst all GBADs stakeholders is an important aspect of the Knowledge Engine. The first dimension of community engagement is the philosophy of offering data in multiple formats and from multiple access points (Application Programming Interface, file downloads of CSV files from Cloud storage, GBADs' TAIL, dashboards). The Knowledge Engine strives to facilitate, not dictate, how users work with data and models – through spreadsheets, R, Python, and data analysis software. While this adds a level of complexity to the management of the system, automation techniques readily available in Cloud repositories and systems allow GBADs to provide this level of accessibility and diversity.

Another dimension of community engagement is the concept of constant feedback and improvement. One example of this are the Ethiopia-related dashboards that have been incrementally improved based on feedback generated by local users at the two in-person meetings in Ethiopia [25,26]. Response to feedback has been facilitated by the technological deployment of the dashboards via software containers and an elastic (i.e. dynamic) container deployment system in conjunction with storage of the software in Cloud-based repositories (GitHub) that are engineered to immediately update the dashboards as soon as changes are made in the code [27]. This allows GBADs software developers to respond immediately to feedback so that the user can see if their feedback has been properly interpreted. This ability greatly enhances the utility of in-person meetings with stakeholders since they can see their insights being immediately integrated into the system.

An extension of this philosophy, that will allow feedback to be collected continuously and from all users, is the new integration of dashboards with Slack (a software communications tool) channels used by the GBADs Informatics team to allow feedback from each dashboard to be directed to both the community at large and appropriate GBADs staff for action [28]. All dashboards will be provided with a 'Comments' feature that allows users to submit feedback about the dashboard (and the specific data being viewed) that will be moderated, stored, and displayed (Table I provides high-level tasks associated with this feature). Once the feedback has been reviewed (to protect the system from inappropriate use) and approved on Slack, it seamlessly updates the dashboard, making it visible to the public and creating a continuous community engagement system. All feedback will be stored in a database to allow for the development of analysis programmes to detect patterns that will help in the further improvement of the dashboards based on user interactions. This system is currently in beta testing and its full launch is anticipated for year-end 2024.

Community members need to feel that they are connecting to others and one aspect of this is to know who those others are and what they are doing with the Knowledge Engine. All interactions with the system are being monitored so that GBADs can provide current and longitudinal data about where its community members are and what software tools they are using. These data are accessible by all community members through a dashboard (Figure 4) that displays the location of users (only resolved to city and country to preserve privacy) and the popularity of tools such as the dashboards. Users can observe the reach of GBADs data, tools, and analysis and can be inspired to try out some of the tools that they are not currently using.

In an analogous manner, developers in the area of animal health and economics can have access to all code driving the Knowledge Engine and the GBADs models and tools.

All GBADs code is open source and developers outside of the GBADs Informatics team are encouraged to use any code available through our GitHub (https://github.com/GBADsInformatics) to develop their own tools, as well as to report bugs and feature requests. A community of developers can greatly add to the impact of the Knowledge Engine and welcome a diversity of voices into the GBADs tool development.

Conclusions

GBADs' methods and data will reach a broader audience due to *1)* the development of GBADs' TAIL; *2)* advances in the standardisation of GBADs' modelling practices; and *3)* the provision of tools to improve community and capacity building. Our tools make data and models more FAIR, facilitating reproducible outcomes and fostering trust in outputs. GBADs' TAIL will be refined and improved based on user feedback, incorporate new languages and will evolve along with advancements in search capability, NLP, and machine learning. The Informatics Team strives to offer better data, for better models, for better decision-making in livestock production. This approach can be adapted to address other complex problems by making advanced technologies more user-friendly and globally applicable.

Acknowledgements

This research is performed in the framework of the Global Burden of Animal Diseases (GBADs) programme which is led by the University of Liverpool and the World Organisation for Animal Health (WOAH). This research is supported through the Grant Agreement Investment ID INV-005366 with the Bill and Melinda Gates Foundation and the UK Foreign, Commonwealth and Development Office (FCDO). GBADs case studies receive additional funding from the following: the European Commission, the Australian Centre for International Agricultural Research (ACIAR), Brooke Foundation and the Food and Agriculture Organization of the United Nations (FAO). A full list of the GBADs collaborators can be accessed here: https://animalhealthmetrics.org/acknowledgements

References

[1] Gates B. (2022). – How to Prevent the Next Pandemic. Vintage, New York, United States of America, 304 pp.

[2] Antle J.M., Jones J.W. & Rosenzweig C. (2017). – Next generation agricultural system models and knowledge products: synthesis and strategy. Agric. Syst., 155, 179–185. https://doi.org/10.1016/j.agsy.2017.05.006

[3] 50×2030 initiative. – A partnership for Data-Smart Agriculture. Available at: https://www.50x2030.org (accessed on 20 November 2023).

[4] Global Agricultural Research Data Innovation and Acceleration Network (GARDIAN). – GARDIAN search engine. Available at: https://gardian.bigdata.cgiar.org/# (accessed on 20 November 2023).

[5] McIver D.J., Patterson G., Stacey D., Rushton J. & Bernardo T.M. (2023). – A horizon scanning exercise of large-scale data aggregation systems to support the creation and sustainability of the Global Burden of Animal Diseases (GBADs) knowledge engine [pre-print]. Social Science Research Network. https://doi.org/10.2139/ssrn.4478794

[6] Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., *et al.* (2016). – The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

[7] Pennington J., Socher R. & Manning C. (2014). – GloVe: Global vectors for word representation. In Proceedings 2014 conference on empirical methods in natural language processing (EMNLP). 25-29 October 2014, Doha, Qatar. Association for Computational Linguistics, Stroudsburg, United States of America, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[8] Kinney R., Anastasiades C., Authur R., Beltagy I., Bragg J., Buraczynski A., *et al.* (2023). – The Semantic Scholar open data platform. arXiv, 2301.10140. https://doi.org/10.48550/ARXIV.2301.10140

[9] Semantic Scholar. – Semantic Scholar API: Overview. Providing a reliable source of scholarly data for developers. Available at: https://www.semanticscholar.org/product/api (accessed on 7 November 2023).

[10] React. – React: The library for web and native user interfaces. Available at: https://react.dev (accessed on 7 November 2023).

[11] Miller P. (1996). – Metadata for the masses. Ariadne, 5. Available at: https://www.ariadne.ac.uk/issue/5/metadata-masses (accessed on 7 November 2023).

[12] Guha R.V., Brickley D. & Macbeth S. (2016). – Schema.org: evolution of structured data on the web. Commun. ACM, 59(2), 44–51. https://doi.org/10.1145/2844544

[13] Edris Abadi R., Ershadi M.J. & Niaki S.T.A. (2023). – A clustering approach for data quality results of research information systems. Inf. Discov. Deliv., 51(4), 337–348. https://doi.org/10.1108/IDD-07-2022-0063

[14] Fox C., Levitin A. & Redman T. (1994). – The notion of data and its quality dimensions. Inf. Process. Manag., 30(1), 9–19. https://doi.org/10.1016/0306-4573(94)90020-5

[15] Food and Agriculture Organization of the United Nations (FAO) (2010). – 2000 World Census of Agriculture: main results and metadata by country (1996–2005). FAO, Rome, Italy, 246 pp. Available at: https://www.fao.org/publications/card/fr/c/d8875104-88cf-55de-91bf-7d9946e4ec8c (accessed on 7 November 2023).

[16] Food and Agriculture Organization of the United Nations (FAO) (2019). – Main results and metadata by country (2006–2015): World Programme for the Census of Agriculture 2010. FAO, Rome, Italy, 404 pp. Available at: https://www.fao.org/documents/card/fr?details=ca6956en (accessed on 7 November 2023).

[17] McKechnie I., Raymond K. & Stacey D.A. – Identifying inconsistencies in data quality between FAOSTAT, WOAH, UN Agriculture Census and National Data.

[18] Nicholson G., Blangiardo M., Briers M., Diggle P.J., Fjelde T.E., Ge H., *et al.* (2022). – Interoperability of statistical models in pandemic preparedness: principles and reality. Statist. Sci., 37(2), 183–206. https://doi.org/10.1214/22-STS854

[19] García-Closas M., Ahearn T.U., Gaudet M.M., Hurson A.N., Balasubramanian J.B., Choudhury P.P., *et al.* (2023). – Moving toward findable, accessible, interoperable, reusable practices in epidemiologic research. Am. J. Epidemiol., 192(6), 995–1005. https://doi.org/10.1093/aje/kwad040

[20] Musen M.A., O'Connor M.J., Schultes E., Martínez-Romero M., Hardi J. & Graybeal J. (2022). – Modeling community standards for metadata as templates makes data FAIR. Sci. Data, 9(1), 696. https://doi.org/10.1038/s41597-022-01815-3

[21] Stacey D., Wulff K., Chikhalla N. & Bernardo T. (2022). – From FAIR to FAIRS: data security by design for the Global Burden of Animal Diseases. Agron. J., 114(5), 2693–2699. https://doi.org/10.1002/agj2.21017

[22] Team RDC (2010). – R: a language and environment for statistical computing. CiNii, 1370294721063650048. Available at: https://cir.nii.ac.jp/crid/1370294721063650048 (accessed on 30 August 2023).

[23] Hornik K. (2012). – The comprehensive R archive network. WIREs Comput. Stat., 4(4), 394–398. https://doi.org/10.1002/wics.1212

[24] Sinha V., Doucet F., Siska C., Gupta R., Liao S. & Ghosh A. (2000). – YAML: a tool for hardware design visualization and capture. In Proceedings 13th International Symposium on System Synthesis. 20-22 September 2000, Madrid, Spain. Institute of Electrical and Electronics Engineers, Piscataway, United States of America, 9–14. https://doi.org/10.1109/ISSS.2000.874023

[25] Lemma M., Temesgen W. & Knight-Jones T. (2022). – Global Burden of Animal Diseases: Ethiopia case study stakeholder workshop. International Livestock Research Institute, Nairobi, Kenya, 12 pp. Available at: https://cgspace.cgiar.org/bitstream/handle/10568/125536/GBADs%20ET%20stakeholder%20workshop.pdf (accessed on 7 November 2023).

[26] Asfaw W., Desta H., Temesgen W., Raymond K., Thomas L., Huntington B., *et al.* (2023). – Global Burden of Animal Diseases: Ethiopia case study stakeholder workshop. International Livestock Research Institute, Nairobi, Kenya, 29 pp. Available at: https://animalhealthmetrics.org/wp-content/uploads/GBADs-Ethiopia-2nd-stakholder-worskshop-report-May-11-2023.pdf (accessed on 7 November 2023).

[27] GitHub. Available at: https://github.com (accessed on 11 September 2023).

[28] Slack. Available at: https://slack.com (accessed on 11 September 2023).

Table I

Progress towards tasks involved in the development of GBADs data, modelling, and visualisation tools

Tool	Task	Description	Status
GBADs TAIL	Graph model development	Design the structure (nodes, properties, and relationships) to store and represent metadata that will allow searching on the TAIL interface	Done
	Graph database population	Continuously adding information to TAIL by preparing and loading metadata from various data sources. This ongoing process helps expand and increase the number of datasets that are findable through TAIL over time. Expanding and increasing the number of datasets that are findable through TAIL by continuously adding metadata from various data sources	Ongoing
	Ontology development	Creating a formal representation of our knowledge domain, defining entities and relationships. It involves unique definitions for species classifications and the connections between them based on the way terms are used in the data, rather than trying to impose a top-down approach	Ongoing
	User interface development	Designing and implementing a user-friendly interface that allows users to search, access, and download data and literature. Development includes integrating with ontology, the graph database, and semantic scholar's API, as well as soliciting feedback for continual improvements	In development
	Development of Software Development Kits, middleware, code, and modules	Creating tools that enable different parts of TAIL to work together	In development
	Development of API to interact with the graph database	An API has been developed to acquire information from the graph database	Ongoing
	Development of API to interact with the ontology	Creating an API to interact with the ontology and enhance the functionality of TAIL	In development
	Development of tests	Planning and creating tests to ensure the TAIL application works well. This includes checking the search accuracy, user interface usability, and overall performance	Not started

Data quality visualisation dashboard (Figure 3)	Develop dashboard	Develop a data visualisation tool in Python Dash that allows users to investigate external and internal consistencies in livestock population datasets from the World Organisation for Animal Health, FAOSTAT, and national agricultural censuses and surveys	Done
	Deploy dashboard	The dashboard is put into a software container, is deployed, and is then available via a URL	Done
GBADs R modelling package	Refactor and modularise the DPM R code	Improve the DPM R code by organising it into clear functions and structuring it for consistent use of parameters for robust calculation of the Animal Health Loss Envelope	In development
	Prepare documentation for CRAN	Create detailed documentation for the R package, in preparation for submission to CRAN. This documentation will allow users to understand how to use the package effectively	In development
	Create and run tests	Create a robust set of tests to ensure that the package is reliable and accurate, and that code can be validated	Not started
Standards for GBADs parameters	Create a dictionary of parameters	A dictionary of parameters is being developed to collect the names, descriptions, codes, and biological maximums and minimums of each parameter, for each model	Ongoing
	Establish a template YAML file to ensure interoperability of parameters	Use the dictionary of parameters to create standard templates for configuration of parameters for the DPM. This standard format, coupled with the GBADs R package, will enable the reuse of model configurations, promoting reusability and transparency of GBADs modelling outputs	In development
Ethiopia Sub-National Dashboard Development	Develop dashboard	Design the layout of the dashboard with input from the end-users. Decide on types of data visualisations, prototyping (creating wireframes to visualise the initial design) and building the dashboard in Python Dash	Done
	Deploy dashboard	The dashboard is put into a software container and once deployed, it is available via a URL (https://gbadske.org/dashboards/ethiopia-population)	Done
	Gather feedback from users. Implement the feedback	User feedback on the usability of dashboards was gathered during two stakeholder workshops at the International Livestock Research Institute. The feedback was used to improve the design and functionality of the dashboard (i.e. aesthetic changes, improved usability and functionality, improvements to speed) and will continue to be incorporated in future developments	Ongoing

Comments feature in dashboards	Develop a comment feature in dashboards	Develop code that allows users to share thoughts, comments, and suggestions directly within the dashboard and hence with the developers. This allows users such as modellers or government officials to easily provide feedback for incorporation into subsequent versions	Done
	Develop a mechanism to send comments to Slack for approval	To protect against inappropriate use, the comments are sent to Slack (a software communications tool) for approval. This task involves developing and implementing a tool called a 'bot' to allow individuals to review and either approve or deny comments, ensuring a moderated and appropriate environment	Done
GBADs usage statistics dashboard (Figure 4)	Set up data tracker	Configure a system to keep track of the cities and countries from which users are accessing the dashboard. Decide how to measure this, where to save this information, in what format and where to save this information in AWS	Done
	Enable dashboard access	Develop code that allows the dashboard to securely access data from AWS. The data is anonymised and only accessible to authorised developers	Done
	Create and design dashboard	Plan how the dashboard will look, choose what information to show and the data visualisations that are used, and develop the dashboard in Python Dash	Done
	Share dashboard	The dashboard is put into a software container and once deployed, it is available via a URL (https://gbadske.org/dashboards/users)	Done
Verification and validation	Accessibility testing	All interfaces (websites, dashboards etc.) must meet accessibility standards set by AODA. AODA is the standard in the Canadian province housing the GBADs Informatics team	Ongoing
	Unit and integration testing	All code (currently Python and R) will undergo a set of standardised tests using best practices for the language	Planning is ongoing

AODA:	Accessibility for Ontarians with Disabilities Act	FAOSTAT:	Food and Agriculture Organization of the United Nations Statistical Database
API:	Application Programming Interface	GBADs:	Global Burden of Animal Diseases
AWS:	Amazon Web Services	TAIL:	Trusted Animal Information portaL
CRAN:	Comprehensive R Archive Network	URL:	Uniform Resource Locator
DPM:	Dynamic Population Model	YAML:	Yet Another Markup Language

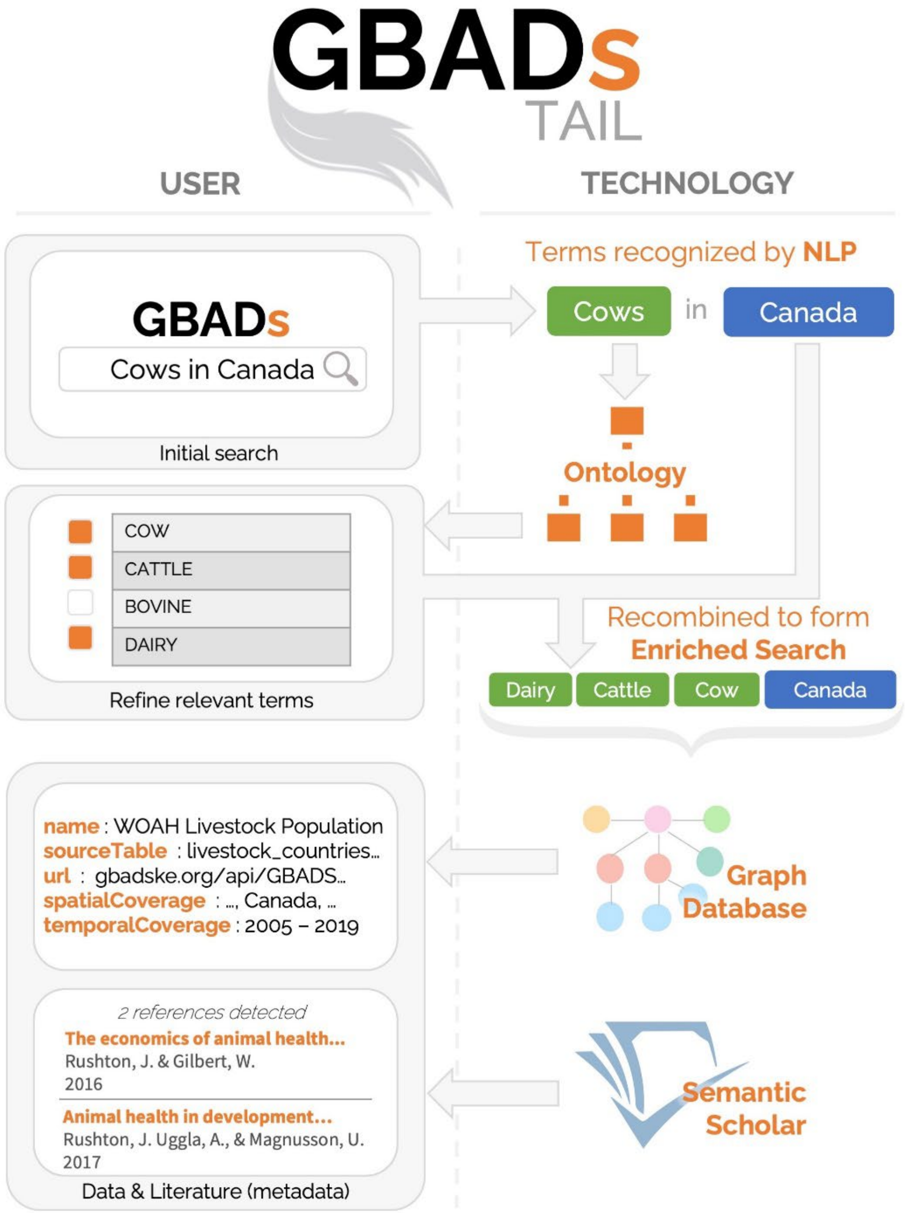

GBADs
TAIL
USER
TECHNOLOGY
GBADs
Cows in Canada
Initial search
Terms recognized by NLP
Cows
in
Canada
Ontology
COW
CATTLE
BOVINE
DAIRY
Refine relevant terms
Recombined to form
Enriched Search
Dairy
Cattle
Cow
Canada
name : WOAH Livestock Population
sourceTable : livestock_countries…
url : gbadske.org/api/GBADS…
spatialCoverage : …, Canada, …
temporalCoverage : 2005 – 2019
Graph
Database
2 references detected
The economics of animal health…
Rushton, J. & Gilbert, W.
2016
Animal health in development…
Rushton, J. Uggla, A., & Magnusson, U.
2017
Semantic
Scholar
Data & Literature (metadata)

Figure 1

Schematic representation of GBADs TAIL operational workflow

Terms from a user’s free-text query are recognised by NLP and species terms are enriched with semantically related terms using an ontology. The user then refines their search by selecting related terms resulting in an enriched search. The enriched search is sent to the graph database and semantic scholar and relevant metadata is presented to the user

GBADs:	Global Burden of Animal Diseases
NLP:	Natural Language Processing
TAIL:	Trusted Animal Information portal
WOAH:	World Organisation for Animal Health

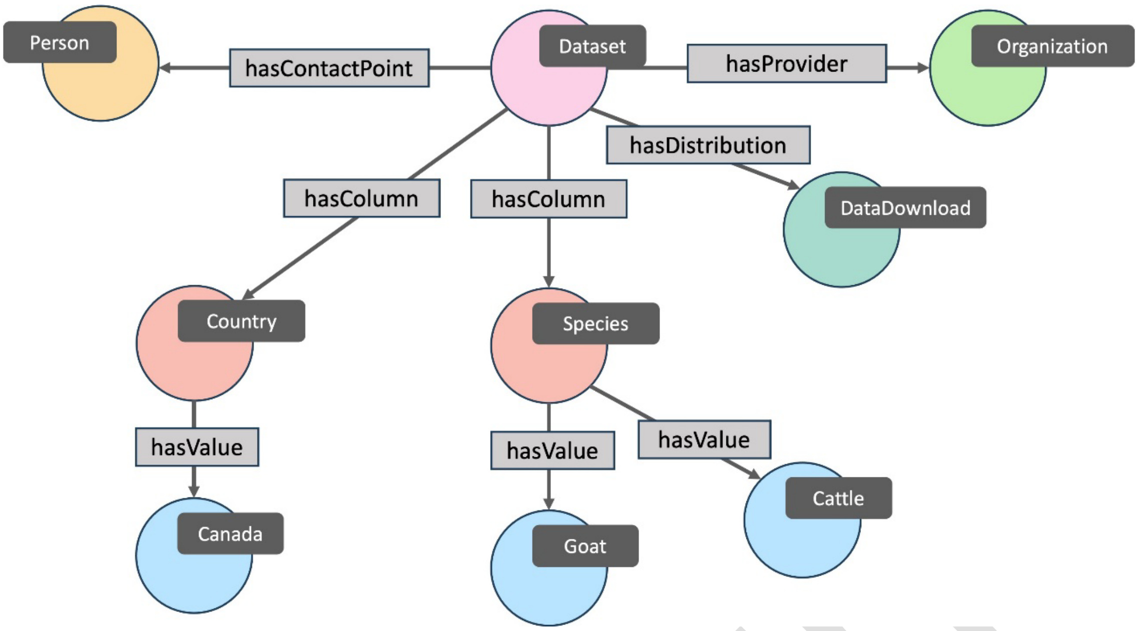
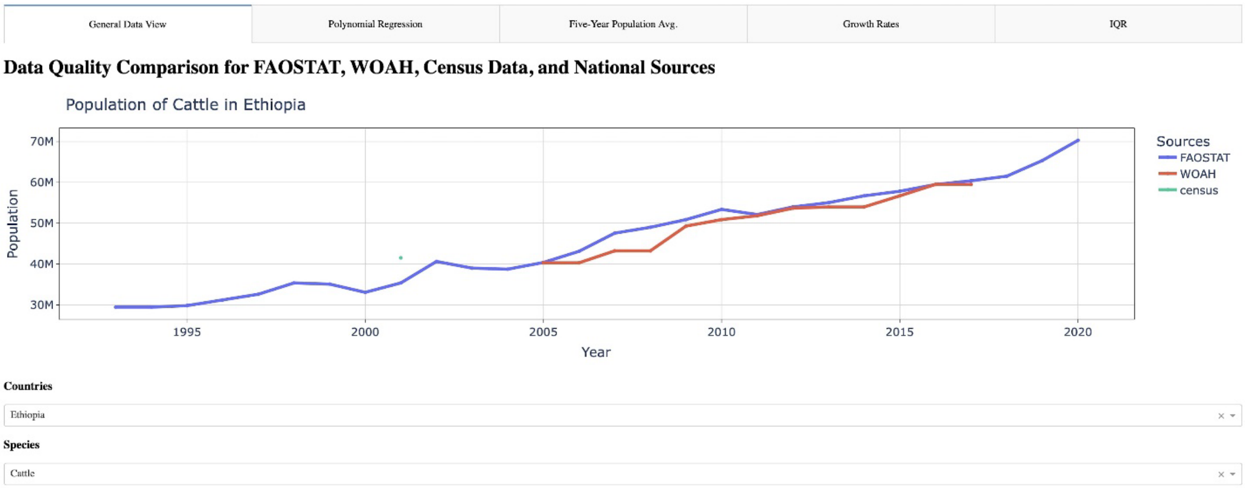


Figure 2

Example of the graph database model that stores metadata for a dataset

The graph database stores all GBADs' metadata and connects data according to commonly reported species and countries allowing for users or machines to query for all datasets related to livestock species and countries of interest

GBADs: Global Burden of Animal Diseases



Data Quality Comparison for FAOSTAT, WOAH, Census Data, and National Sources

Countries

Ethiopia

Species

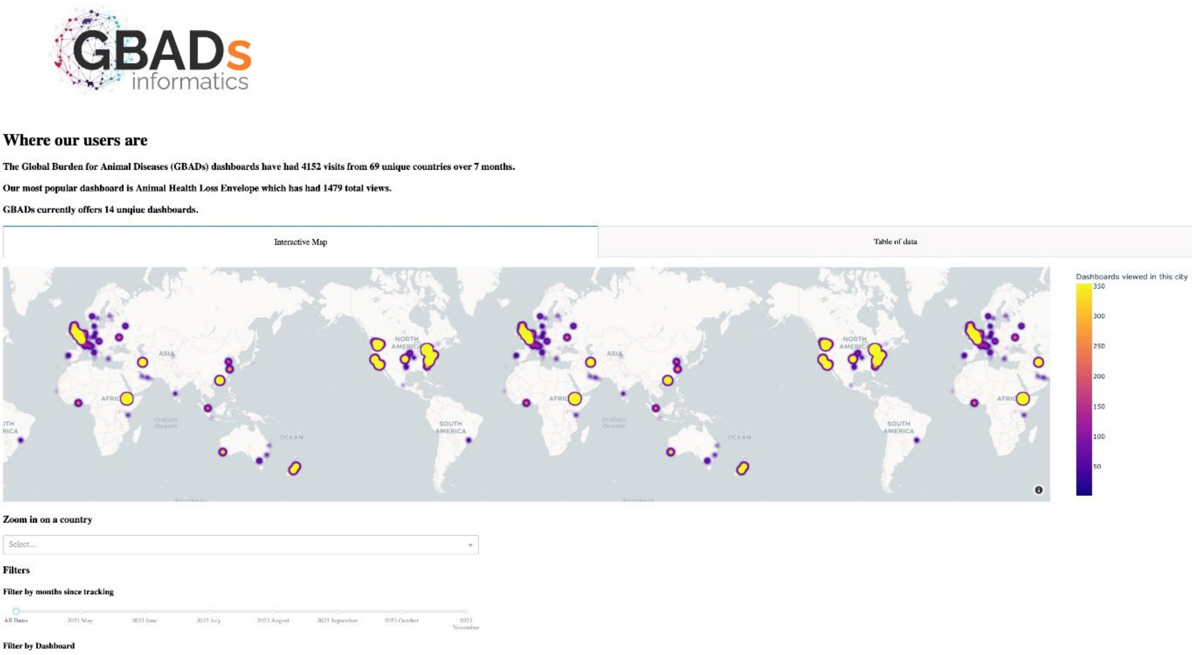
Cattle

Figure 3

The GBADs Data Quality Visualiser Tool showing the population of cattle in Ethiopia

The tool has five options: General Data Viewer, Polynomial Regression, Five-Year Population Average, Growth Rates and Interquartile Range. Each tool tab allows the user to visualise data in unique ways to spot inconsistencies. Data from WOAH, the FAOSTAT crops and livestock products (Live animals: QCL) dataset, and census data [15,16] are visualised. Note that there is only one year of census data and therefore it appears as a single green dot

FAOSTAT:	Food and Agriculture Organization of the United Nations Statistical Database
GBADs:	Global Burden of Animal Diseases
IQR:	Interquartile Range
WOAH:	World Organisation for Animal Health



Where our users are

The Global Burden for Animal Diseases (GBADs) dashboards have had 4152 visits from 69 unique countries over 7 months.

Our most popular dashboard is Animal Health Loss Envelope which has had 1479 total views.

GBADs currently offers 14 unqiue dashboards.

Zoom in on a country

Filters

Filter by months since tracking

Filter by Dashboard

Figure 4

GBADs Informatics Usage Statistics dashboard

The dashboard shows where users are and the number of views on each GBADs Informatics resource such as dashboards and APIs

APIs: Application Programming Interfaces
GBADs: Global Burden of Animal Diseases