Aquaculture Data Management and Governance Strategy

Kassy Raymond, Dr. Deborah Stacey and Dr. Theresa Bernardo

April 2021

1 Introduction

Finding data relevant for analysis usually involves searching for data through open data portals or contacting data owners, government or private sources and requesting data access. Relevant data may be accessed through direct download, file sharing via email or other means or through an application programming interface (API). However, finding rich data sources is often contingent on level of expertise, an individual's network or simply knowing where to look.

In addition to the challenges in finding relevant and high-quality data, data from different sources are often unstable due to differing collection methods, contextual views and granularity. These differences in conceptualization of data make it difficult or impossible to combine data from different sources. Further, critical information about datasets such as transformations to data, collection methods, use of vocabularies and imputations may be poorly described or difficult to find.

Curating data sources and capturing their metadata enables metadata to be used for knowledge discovery through **metadata-driven software design**. Metadata provides information about how the data is collected, how it can be used (licensing), how it is disseminated, where it originated, what the subject matter of the data is and who owns the data for example. Each of these key pieces of information is recorded using standards (Section 2.2 to ensure that data are **F**indable, **A**ccessible, **I**nteroperable, **R**eusable and **S**ecure ('FAIR(S)') [18] [16]. Creation of a graph database based knowledge engine (KE), allows data from different sources to be linked by their metadata, allowing data sources and their metadata to be accessed through an application programming interface (API) which allows automated machine to machine transmission of data.

While the general framework of the KE can be applied to different disciplines of data, domain-specific knowledge is required to create metadata that are **F**indable, **A**ccessible and **I**nteroperable to those that will be querying and using the system. This includes understanding the vocabularies, metrics, codifications, classification systems and ontologies that may be used to describe the data that will flow through the KE. Moreover, security is pivotal in developing the reputation of a trusted broker and in some cases, a repository for data. Therefore, metadata is designed to capture the security and privacy implications of each data source and to ensure that data are only disseminated to those with specific permissions.

This report will outline the overview of the framework used to develop the KE, the architecture of the metadata-driven KE and data specifics about the aquaculture sector using salmon production as the pilot.

2 Data Governance

Data governance is a system that governs access to data, allows for data assets to be quantified under domain-specific categories and involves technologies and processes that define data assets in an organization or more broadly a domain.

As data from different domains use different collection methods, vocabularies and are used in different ways, it is important to have a data governance and management strategy that satisfies the unique needs of those using the data and the data's strengths and weaknesses.

The data governance framework includes the following, adapted from [5]:

- **Data discovery and curation**: Profiling data and tracing data lineage (provenance), and determining vocabularies, ontologies, classification systems and codes used in the data domain
- **Data management**: Creating standard metadata and cataloguing metadata
- **Policies, privacy, security and access**: Understanding data privacy issues, including data politics, access constraints, and licensing

2.1 Data Discovery and Curation

The goal of the data discovery and curation phase is to create a data ecosystem by discovering all relevant sources of data and vocabularies, ontologies and codifications used to describe or represent the data. Data discovery and curation typically involves communicating with domain experts to capture information about key data including where data can be accessed, data politics and key countries, diseases and species. Information about which types of data are publicly available and which are private are also captured during this phase. During the discovery phase data lineage is also captured. This includes recording where data originated, how data are obtained, by whom and whether the data were transformed or changed before or after use. Additionally, this should provide information about how data are accessed (direct download, API, private vs. open) which will be integral in the dissemination of the data in the KE. Once a list of relevant sources are obtained, the data can be profiled. This includes creating broad categorizations of the data. Ultimately, a hierarchical representation of the key categories of data can be created, providing a broad overview of the data ecosystem.

2.2 Data Management

Using what was learned from the data discovery and curation stage, data are managed such that all relevant information about the data are captured in metadata. The outcome from the data management phase is a standardized metadata catalogue of all relevant data for a given domain. This will enable users to find, reuse and access data by querying the metadata catalogue.

Descriptive, technical and access metadata are created and stored in a Json-LD file for each data source. Standard metadata terms are used to ensure that data are findable and interoperable. Descriptive metadata includes information such as temporal range, spatial range, species, accrual method, accrual periodicity, author(s), data contact points, description of data set, data dependencies, collection methods, language, creator, subject and when the data was created. The value for each term is also standard; specific vocabularies and date types are used for each term. Each of the terms used are selected from standard metadata element sets such as Dublin Core Metadata Initiative and schema.org [3] [7]. Technical metadata includes the data distribution (direct download, API or web scraping) and the size of the data. Access metadata articulates who owns the data, licensing and access rights. Each metadata file and term is integral to the knowledge engine architecture as it is the metadata that are driving the discoverability of the data and governing access to it. Therefore, by understanding the key issues and models used in the data domain, metadata creation can be optimized through using keywords which can drive data queries.

2.3 Policies, privacy, security and access

Policies, privacy, security and access for each dataset or database are captured in the metadata. Using the CARE data (Collective benefit, Authority to Control, Responsibility and Ethics) principles as a guide, the overall premise of system security is to ensure that data are used for the Collective benefit of data owners including under-represented groups and that the system honours data sovereignty throughout the entire lifecycle of data use [4]. Private data contributors may also choose to make their metadata open or not. Allowing the contributor to have the Authority to control their data and metadata benefits GBADs in its role as a reputable trusted data broker. In complex cases, GBADs should create private data licensing agreements to ensure that the terms and conditions of private data use protect both parties in the agreement. In particular, licensing agreements of data users must articulate what the private data can be used for, how long the user has access to this data and the conditions for publishing the data in journals. As it is difficult to anticipate the security needs and wants of private data holders, it will be integral to work directly with private data holders to create licensing agreements and terms and conditions.

While it is currently not known which data require more security than others, geographically sensitive data can be aggregated at a higher granularity to protect the identities of those whom the geography represent. It is not yet known if personal identifiable information (PII) will be of relevance in aquaculture so this will be approached on a case by case basis.

2.3.1 Accessing private data through the KE

When a user enters the GBADs KE system they will have access to all open data. However, to access private data or data with specific access rights, users will be required to create a login where they will be required to have an authorized user verify their identity. When the user's identity is verified, depending on their identity and role they may be granted access to private data. The verified user will have to agree to the terms and conditions of the data they have access to including if or under what conditions they can the share the data with colleagues.

3 Overview of Knowledge Engine Architecture

The metadata that are created will be stored in a graph database that will allow common metadata properties to be linked. The graph database will serve as the query engine that will allow users to query based on terms of interest. Figure 1 shows a general overview of how a user may use the KE to access data. In general, a user queries the metadata engine for data using terms of interest. For example, providing the system with the word 'Ethiopia' would return all metadata related to Ethiopia. The user can then read through the description of each dataset and select the data that they would like. The GBADs API either gathers the raw data from an API or from the repository where the data is stored before returning it to the user in their desired format. If desired, a user may also use the GBADs API directly in their programs to get data.

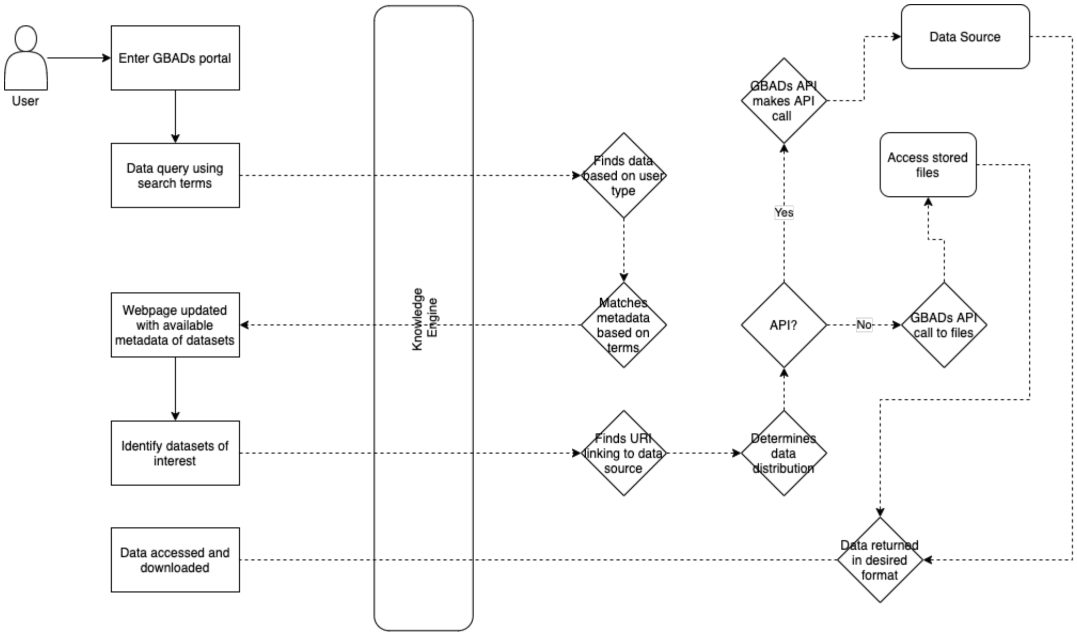
User
Enter GBADs portal
Data query using search terms
Webpage updated with available metadata of datasets
Identify datasets of interest
Data accessed and downloaded
Knowledge Engine
Finds data based on user type
Matches metadata based on terms
Finds URI linking to data source
Determines data distribution
API?
Yes
No
GBADs API makes API call
Data Source
Access stored files
GBADs API call to files
Data returned in desired format

Figure 1: Knowledge engine overview. Flowchart of a non-verified user accessing data through the GBADs data portal.

Disease	Measurement	OIE Reportable
Infectious Salmon Anaemia (ISA)	Number of fish infected	Yes
Salmonoid Alphavirus (SAV)	Number of fish infected	Yes
Sea lice	Lice per fish	No
Piscirickettsiosis (Chile)	Number of fish infected	No

Table 1: Salmon diseases of interest.

4 Aquaculture Specific Data Governance and Management

As mentioned, while the general framework of the data governance and management strategy is consistent across domains, the data ecosystem varies. Preliminary data discovery and curation of aquaculture data was guided by conversations with aquaculture experts Hernan Rojas [15] and Crawford Revies [14]. These conversations provided background information about the salmon production sector and the aquaculture industry in general. Using salmon as the pilot species, the following were discussed: salmon production systems and practices, key production countries, diseases of particular interest or concern, reportable diseases, methods of vaccination and treatment, feed, water quality and water temperature. Here we report on and discuss the data resources that were collected and how they are expected to vary between different countries, the types of data that may be available, potential challenges of obtaining aquaculture data and strategies in discovering and curating data.

In general, the aquaculture sector broadly consists of fish, shrimp and oyster farming. Focusing on salmon, the main production countries are Norway, Chile, Western Canada, Faroe Islands, and the United Kingdom. For these countries, data related to mortality, disease (Table 1), vaccination, antibiotic treatment, salmon weight, water quality and feed intake are reported on a regular basis during both freshwater and saltwater stages of production. The periodicity of recording is dependent on what is being measured and on the country where production is taking place. In addition, data related to trade, production and salmon

Title	Description	Citation	Distribution	Access	Spatial Coverage
BarentsWatch	"Weekly historical data of sea lice status on site level for salmonoids at sea including sea temperature, historical data of lice countermeasures, medicamental treatments (bath and feed), mechanical removal, cleaner fish and suspected / confirmed status for notifiable fish diseases (PD and ISA)."	[2]	Direct Download (csv); API	Open	Norway; lat/lon of production sites
Statistics Canada - Industry sea lice counts at BC marine finfish aquaculture sites	"The DFO Sea Lice Audit Report shows the results of DFO's random quarterly lice audits, which coincide with one of the licence holder's numerous scheduled counts. To assure quality, farm staff count lice on 50 per cent of the selected fish and DFO staff count lice on the other half."	[9]	Direct Download (csv)	Open	British Columbia, Canada
FishStatJ - Software for Fishery and Aquaculture Statistical Time Series	"FishStatJ... include[s] datasets on production, trade and consumption. Data can be extracted and aggregated according to different level of details and international standard classifications. It consists of a main application and several workspaces that include the datasets."	[6]	Database available through downloading the FishStat application	Open	To be determined; all countries that report to FAO
Norwegian Directorate of Fisheries: Statistics for aquaculture	Production and losses, percentage of sale, overview of livestock, juvenile fish production, number of hatched juveniles of Atlantic salmon and rainbow trout	[12]	Direct download (excel spreadsheet)	Open	Norway

Table 2: Preliminary data discovery and curation of country specific salmon production data.

consumption are of particular interest. Data was reported to be scarce in countries that are not main producers [14], however, we have not yet tried to access or search for data from other countries.

Country specific data sources that are summarized in Table 2. Data can also be accessed through databases that include data from multiple countries. Data from Chile are not publicly available but are available through government and private sources. Piscirickettsiosis, a bacterial disease affecting salmon is endemic only in Chile and is not an OIE reportable disease. Therefore, it is likely that this data will be particularly hard to access without a close trust relationship.

4.1 Data from Literature

Information about the salmon production sector can also be found in literature; data from journal articles provides insight about commonly used data resources, data availability and the ability to scrape tables when open data are unavailable. Systematic reviews will be useful in revealing data sources used to inspire publications, models and decision making. In particular, the following papers have been of use in the work so far: [1], [8], [10], [17], [11] and [13].

4.2 Future Directions

Future work should extend on the preliminary data discovery by curating more information about the data sets already discovered, explore and curate data from model outputs, explore FISHSTAT, continue data discovery phase to increase knowledge of data resources and speak to economic experts about production, trade and export of salmon and aquaculture in general. GBADs Informatics should also collaborate with themes working on aquaculture systematic reviews as an avenue to scrape data from journal articles. Further, public access to OIE aquaculture data is not available through WAHIS; annual reports provide data however access to raw data files would allow us to avoid scrapping the data from pdf files. Once an adequate amount of data is catalogued, a metadata analysis can be conducted to reveal information about differences in naming conventions, spatial granularity, data reporting structures, ontologies and vocabularies for example. Additionally, metadata analyses can provide insight about areas of data richness and where data are lacking.

References

[1] Jennifer L Bailey and Sigrid Sandve Eggereide. "Mapping actors and arguments in the Norwegian aquaculture debate". In: *Marine Policy* 115 (2020), p. 103898.

[2] *BarentsWatch: Portal to coastal and sea areas in the north.* URL: https://www.barentswatch.no/en/.

[3] DCMI Usage Board. "DCMI metadata terms". In: (2008).

[4] Stephanie Russo Carroll et al. "The CARE Principles for Indigenous Data Governance". In: *Data Science Journal* 19.1 (2020).

[5] Evren Eryurek et al. *Data Governance: The Definitive Guide: People, Processes, and Tools to Operationalize Data Trustworthiness.* English. Paperback. O'Reilly Media, Mar. 2021. ISBN: 978-1492063490.

[6] *FishStatJ - Software for Fishery and Aquaculture Statistical Time Series.* Accessed: 2021-04-21. URL: http://www.fao.org/fishery/statistics/.

[7] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. "Schema. org: evolution of structured data on the web". In: *Communications of the ACM* 59.2 (2016), pp. 44–51.

[8] Rita Hannisdal et al. "Anti-sea lice agents in Norwegian aquaculture; surveillance, treatment trends and possible implications for food safety". In: *Aquaculture* 521 (2020), p. 735044.

[9] *Industry sea lice counts at BC marine finfish aquaculture sites.* Accessed: 2021-04-21. URL: https://open.canada.ca/data/en/dataset/3cafbe89-c98b-4b44-88f1-594e8d28838d.

[10] Audun Iversen et al. "Production cost and competitiveness in major salmon farming countries 2003–2018". In: *Aquaculture* 522 (2020), p. 735089.

[11] Tróndur J Kragesteen et al. "Optimal salmon lice treatment threshold and tragedy of the commons in salmon farm networks". In: *Aquaculture* 512 (2019), p. 734329.

[12] *Norwegian Directorate of Fisheries: Statistics for aquaculture.* Accessed: 2021-04-21. URL: https://www.fiskeridir.no/English/Aquaculture/Statistics/Atlantic-salmon-and-rainbow-trout.

[13] Ruth BM Pincinato et al. “Factors influencing production loss in salmonid farming”. In: *Aquaculture* 532 (2021), p. 736034.

[14] Crawford Revie. Personal Communication; Meeting with GBADs Informatics. Apr. 7, 2021.

[15] Hernan Rojas. Personal Communication; Meeting with GBADs Informatics. Mar. 31, 2021.

[16] D Stacey et al. “From FAIR to FAIRS: Data Security by Design for the Global Burden of Animal Diseases (GBADs)”. In: *Agronomy Journal - American Society of Agronomy* (2020).

[17] Lars Helge Stien et al. “Governing the welfare of Norwegian farmed salmon: Three conflict cases”. In: *Marine Policy* 117 (2020), p. 103969.

[18] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.