scientific data

Article in Press

Check for updates

OPEN

ARTICLE

Enhancing reusability of veterinary epidemiological data by creating contextual metadata

Céline Faverjon[1]✉, Camille Delavenne[1], Servane Bareille[2], Angus Cameron[1], Luis P. Carmo[3], Katharine R. Dean[3], Marco De Nardi[4], Pauline Ezanno[5], Jenny Frössling[6], Wiktor Gustafsson[6], Miel Hosten[7], Anne Meyer[8], Saba Noor[7], Victor H. S. Oliveira[3], Magnus N. Osnes[3], Sebastien Picault[5], Crawford W. Revie[9], Javier Sanchez[10], Inge Santman-Berends[11], Gema Vidal[6], Clazien J. de Vos[12] & Gerdien van Schaik[11,13]

Despite the global adoption of the FAIR guiding principles, significant barriers remain in their practical implementation due to the lack of domain-specific standards for "rich metadata". While generic metadata schemes provide basic information, they often fail to capture the complexity of data quality and the nuances of specialized scientific fields. This gap is particularly evident in veterinary epidemiology, a field characterized by complex, multi-scale data spanning various domains where detailed description of data analysis and modelling tend to be prioritized over raw data description, leading to inconsistent reporting and low data reusability. This paper introduces the first community-developed rich metadata guidelines specifically designed for veterinary epidemiology datasets, accompanied by practical templates and real-world examples. By integrating existing standards with specific guidance on domain-specific attributes and data quality, these guidelines provide a comprehensive, single-source framework accessible to researchers across academic, public, and private sectors. They aim to support researchers and stakeholders in enhancing the reusability of animal health data.

Introduction

Over the last decade, a growing number of governments, funding agencies (e.g., US National Institutes of Health, European Commission's Horizon Europe programme) and individuals have promoted data reusability to maximize research investments' efficiency, enhance reproducibility and transparency, and foster collaboration and innovation. Several tools and guidelines have been developed to support these efforts. The most widely adopted are the FAIR guiding principles (i.e., Findable, Accessible, Interoperable, Reusable) which are intended to, "*act as a guideline for those wishing to enhance the reusability of their data*"[1], and which have become the reference in many organizations for good research data management practices. The FAIR guiding principles highlight the importance of metadata for supporting data reusability and insist on the importance of having 'rich metadata' (i.e., Guiding principles '*F2: Data are described with rich metadata*' and '*R1: meta(data) are richly described with a plurality of accurate and relevant attributes*'). The metadata is a set of data that describes and gives information about other data. The definition of the term is very broad and the guiding principles stop short of providing

[1]EpiMundi, Lyon, France. [2]University of Lyon—French Agency for Food, Environmental and Occupational Health and Safety (ANSES), Epidemiology and Surveillance Support Unit, Lyon, France. [3]Norwegian Veterinary Institute, P.O. Box 64, NO-1431, Ås, Norway. [4]Department of Veterinary Medical Sciences (DIMEVET), Alma Mater Studiorum University of Bologna, Bologna, Ozzano dell' Emilia (Bo), Italy. [5]Oniris, INRAE, BIOEPAR, 44330, Nantes, France. [6]Department of Epidemiology, Surveillance and Risk Assessment, Swedish Veterinary Agency (SVA), SE-751 89, Uppsala, Sweden. [7]Laboratory for Animal Nutrition and Animal Product Quality (Lanupro), Department of Animal Sciences and Aquatic Ecology, Ghent University, Gent, Belgium. [8]Episystemic, Lyon, France. [9]Animal Production and Health Division, UN FAO, Rome, Italy. [10]Centre for Veterinary Epidemiological Research, Atlantic Veterinary College, University of Prince Edward Island, 550 Univ. Ave, Charlottetown, PEI, C1A 4P3, Canada. [11]Royal GD, Deventer, the Netherlands. [12]Wageningen Bioveterinary Research, Wageningen University & Research, Lelystad, the Netherlands. [13]Department of Population Health Sciences, section Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands. ✉e-mail: celine.faverjon@protonmail.com

detailed information on what metadata or rich metadata should look like in practice. They even push the responsibility of deciding which metadata elements are most relevant to each individual scientific community[2].

To help scientists and other stakeholders, several metadata schemes have been developed. These schemes often cover, in some detail, general information relating to digital objects (e.g., title, unique identifier, creators, etc.) useful to meet several of the FAIR guiding principles. The concept of rich metadata is often simply covered by a section entitled, "*Description of the data*", where the users are invited to provide additional information about the data that did not fit in any of the generic categories they have previously defined (for examples see[3–5]), thus leaving plenty of room for interpretation. In addition, the existing generic metadata standards usually do not include consideration of data quality, even though having access to information related to dataset flaws has been identified as a critical issue for data (re)usability and interoperability[6].

Having generic specifications makes sense when the purpose of the metadata scheme is to cover a broad range of datasets. However, it also comes with challenges for users when they try to describe their own datasets and associated limitations. The absence of existing specific guidance on dataset flaws has been, for example, identified as a major barrier to the reporting of data quality[7]. It can be daunting to identify what information is required or useful for describing a dataset and its quality, often pushing authors to report it inconsistently, or to simply overlook this task[8]. This significantly impairs data reusability as the information needed by the person trying to reuse the data is rarely available or sufficiently understandable, especially when datasets are complex. The difficulty is that what makes sense in terms of data description and data quality changes, depending on the nature of the data involved and the specific domain they are related to. To meet this challenge, a growing number of communities are developing their own standards to ensure that the data they generate contain the minimum information required to ensure that they can be easily interpreted and therefore reused. For example, there are minimum information standards relating to microarray experiments (MIAME – Minimum information about a microarray experiment)[9], arthropod abundance[10], or more recently crops data in agriculture[11]; but many are still pending (see, for example, a call to action for food composition data[12] and examples of initiatives in public health[13,14]).

In the field of animal health, and in particular in the veterinary epidemiology, the lack of fit-for-purpose standards has been identified as one of the reasons that could explain the overall low level of FAIR-ness of the data used in this domain, and likely contributes to limiting data reuse[15,16]. Animal health is of key importance not only to ensure sustainable access to safe animal-sourced food, but also to protect public health as highlighted by the emergence of the 'One Health' concept[17]. One of its specificities is that data in veterinary epidemiology are often complex, multi-scale, and span diverse research domains, such as genetics, population dynamics, economics and biology. While epidemiology can be defined as the study of disease in populations and of factors that determine its occurrence, veterinary epidemiology goes beyond and includes the study of other health-related factors such as livestock productivity, economics and biosecurity. Veterinary epidemiology is also used in a wide range of species and production systems and operates at the interface of ecology when wild animal populations or vector-borne diseases are being studied. Therefore, the purposes of the data used in veterinary epidemiology and their associated degree of complexity can vary greatly. This diversity comes with additional challenges when trying to define standards to support data reuse as these standards must, in this context, meet two competing goals: being consistent and flexible. Work conducted by[18,19] has tried to partially fill this gap by defining minimum data standards for laboratory data and wildlife disease research and surveillance, respectively. However, these standards focus on standard data formats (and not metadata) to be used for a very specific kind of veterinary epidemiology data. Guidelines have also been developed to strengthen the reporting of epidemiological studies (see for examples[20,21]). These frameworks focus primarily on describing the analytical framework and its bias, offering only a secondary role to the data. Data are frequently assessed exclusively within the context of the study that uses them. Therefore, there remain significant gaps in terms of mechanisms to produce comprehensive enough information to facilitate the broad reuse of the wide array of veterinary epidemiological data.

The purpose of this work is to propose comprehensive and practical instructions on how to create domain-relevant rich metadata in veterinary epidemiology to support data reuse in this field. In these guidelines, we mostly focus on elements that are specific to veterinary epidemiology, but, for the sake of completeness, we also briefly cover generic aspects of the metadata, by referring to existing standards. Special attention is also given to the description of the dataset quality, as it has often been found to be poorly documented in the field of animal health[15,16,22]. These instructions intend to provide guidance to data owners or data holders working with these data, whether they are researchers creating data as part of research studies, or public or private stakeholders involved in animal health and production and generating data on a routine basis.

Method

Following the work done by[23], we initiated a community-centric model that engaged domain scientists to develop formats for common veterinary epidemiology data types. The objective was to create formatting guidelines and templates that would gather the minimum but sufficient metadata necessary for data interpretation and reuse.

The guidelines and associated templates were developed following four rounds of reviews by members of the community of veterinary epidemiologists. In total, 25 individuals representing 13 institutions provided suggestions at various stages of the development process. This process is detailed below.

The team leaders first reviewed existing metadata guidelines and used them as a basis to develop a first draft of the guidelines. The first version was developed as a simple text document, which was initially shared and discussed within a small group of researchers. A second version was then developed and discussed with members of the community during a workshop focusing on Data Stewardship and organized as part of the annual conference of the Society for Veterinary Epidemiology and Preventive Medicine in Berlin in March 2025.

This second round of comments led to a third version of the guidelines and to the development of metadata templates. These templates were developed using a tabular strategy to make it easier for others to review and document the metadata, and check errors[24]. The objective was also to avoid a potential lack of engagement of contributors because of the eventual lack of knowledge on machine-readable data formats, as it could be an issue in this community[16]. The tabular templates were proposed in.csv and.xlsx formats, which are data formats often used by veterinary epidemiologists. This third version of the guidelines and associated templates were shared with the members of the core group and one member of the advisory board of a large Horizon 2020 European-commissioned research project named DECIDE (i.e., Data-driven control and prioritization of non-EU-regulated contagious animal diseases), which gathered 19 partners from the public and private sectors across 11 European countries.

A fourth version was developed after including their comments and was then shared with the rest of the consortium of the DECIDE project and with interested members of the veterinary epidemiology community identified amongst the participants of the workshop organized in Berlin in March 2025. The contributors piloted the guidelines and associated templates within their research groups using different datasets to ensure they were practical and useful for scientists who collect and reuse data. This last round of reviews gathered only relatively minor feedback and led to the version of the guidelines and their associated templates presented in this paper. The tabular strategy used to develop the templates had the added benefit of being easily extendable by all scientists in the community, and convertible later to other metadata formats (e.g., geospatial formats such as the Keyhole Markup Language (KML)) or to a semantic format such as Resource Description Framework (RDF)[25].

Once the guidelines and associated templates were finalized, real life examples were also developed from the DECIDE project to better illustrate how the templates could be filled for two different datasets. The first dataset was related to Polish poultry production data, which had been aggregated and presented a number of data quality issues[26]. The second dataset was related to Norwegian salmonids production data[27]. These two examples are used in the following sections to demonstrate the use of the guidelines.

Results

Data covered by the guidelines. These guidelines do not cover all types of data used in veterinary epidemiology. Specific data, such as those used in statistics or molecular epidemiology research are not considered. Indeed, these specific domains have already made significant advances in terms of data sharing, where standards exist and are already in widespread use (for example, in molecular epidemiology[15,28,29] and in statistics, with the SDMX initiative[30]). The focus of these guidelines is on nonspecific data used in veterinary epidemiology, such as data covering animal populations, disease prevalence, antimicrobial usage, presence or absence of risk factors, economic data, production performance, sensor data or animal movements.

The datasets described using these guidelines can be made publicly available or simply preserved in private repositories. Indeed, creating good metadata should be considered independently from the action of sharing the data with external people. Whatever the data sharing status, the data should be thoroughly described with rich metadata to facilitate their re-use by the members of the organization owning the data, and by others when applicable.

It is also important to highlight the fact that metadata describes a specific digital object. This means that, if only parts of the data maintained in private repositories are made publicly available (for example in open repositories like Zenodo (https://zenodo.org/) or Dryad (https://datadryad.org/)), and sensitive information, for example, is removed before data sharing, then these data should be considered a new digital object with slightly different characteristics than the original data. Therefore, they should be associated with different metadata than the original data. In practice, their metadata will be largely similar, but will slightly vary in terms of, for example, update frequency, and data anonymization.

How and when to use these guidelines. These guidelines should be consulted before creating metadata to have a good understanding of the information that would be required and how it could be structured. The best time to create metadata is during or before the dataset creation or collection process. Steps should be taken to facilitate data reuse from the very beginning to facilitate the task and save resources. Indeed, even if it is always possible to create metadata *a posteriori*, it will be a lot more time-consuming than if considered from the very beginning and integrated into the workflow.

Any metadata created in the context of a data collection process should be stored should be stored together with the data they describe in an appropriate repository. Multiplying the repositories where the data and their metadata are deposited is of interest to increase data longevity and maximize data discoverability. The appropriate repositories to be used depend on the nature of the data, and the rules set by each organization owning the data. However, whatever workflow is decided, it should take into account the evolving services provided by the repository, so that all data versions over time are properly linked.

In these guidelines, the metadata are structured around three components, which are further described in the following sections:

- *General information* – this section refers to basic information associated with any digital object and includes data access rules.
- *High-level description* – this section intends to provide a basic understanding of what the dataset is about, defining what should be included in the description of a dataset to make it meaningful for others.
- *Specific content description* – this section intends to provide a specific content description of the data for someone else to be able to actually (re)use it in an appropriate way[31]. It focuses on detailed information, such as data models, data dictionaries, and data quality (e.g., completeness and accuracy).

Note that these guidelines present the items that the metadata cover but also propose metadata templates based on files in.csv and.xlsx format (see appendices). Real life examples of datasets described using the templates are also available in[26] and[27].

Section: General information. For the sake of completeness, our guidelines include the key elements usually reported in the existing metadata schemes, based on the scheme proposed by the DataCite Metadata Working Group[4]. Some of the items highlighted in this section of our guidelines may be found in slightly different versions or under different names, depending on the platform where the dataset is published and/or the generic metadata scheme used as a reference. The list of generic information to be reported about any dataset is presented below. Such information (as well as some elements from Section "General data description") is generally embedded with the data in formal data repositories and online platforms. However, an example of a template to be filled in order to provide structured general information about a dataset is provided in the files named "1_GeneralDescription" available on Zenodo[32] in case the dataset is not stored in such repository.

- **Title:** A name or title by which a resource is known[4]. A way to identify or refer to the dataset is needed to be able to describe it. It can take the form of a plain language name, which may include the organisation creating the dataset, the purpose or content, and, in the case of a multi-part collection, an identifier of the specific part (e.g., the year or version).

- **Unique Identifier:** the title or name of the dataset should be complemented with a unique identifier in particular if the dataset is published in a data catalogue or a data repository. The unique identifier can be a number (e.g., a digital object identifier (DOI®) that is commonly used to reference journal articles) or a unique text identifier based on the name such as used in some data catalogues, like OpenDataSoft.
- **Creators:** The main researchers involved in producing the data, and/or the authors of the publication, in priority order, including contact details and when relevant their contribution to the data. For instruments collecting data this is the manufacturer or developer of the instrument[4].
- **Publisher:** The name of the entities that hold, archive, publish, print, distribute, or release the resource[4].
- **Contributors:** The institution or person responsible for managing, distributing, or otherwise contributing to the development of the resource[4].
- **Publication date:** Date when this version of the dataset was made available.
- **Funding reference:** Information about financial support (funding) for the resource being registered ideally using URL of the funding reference and grant number[4].
- **Size:** Size of the dataset (e.g., bytes, pages, inches) or duration (extent, e.g., hours, minutes, days) of a resource[4].
- **Version:** the version number of the resource or dataset.
- **Language:** This is the (human) language or languages that the dataset includes. This does not refer to the encoding system used in the data (see "Format" below). While many datasets may be highly quantitative (containing mostly numbers), interpretation of these values requires data dictionaries with text descriptions. Some datasets may also include free text fields recorded in one or more languages.
- **Keywords or subject:** Subject, keyword, classification code, or key phrase describing the resource[4].
- **Format:** Technical format of the resource[4]. This refers to the data format and encoding system used to store the data. It could be a simple XML or CSV file, but other formats could be used, such as OWL/RDF, JSON, BSON, PDF, SQLite, or DB files.

Special attention should be given to the rights associated with the dataset. The rights of the datasets aim at reporting the rules for who can access and reuse the data, for what purposes and under what conditions. It should clearly define the roles and responsibilities, as well as the data reuse policy associated with the dataset. More specifically, it should include the following items:

- **Data owner:** Data owner means a legal person who, in accordance with applicable law, has the right to grant access to or to share certain personal or non-personal data under its control and is concerned with risk and appropriate access to data. For example, in a surveillance scheme, a data collector can be a laboratory, which stores the surveillance results in a governmental database according to legal framework, and where the data owner is the government.
- **Other key stakeholders:** In certain cases, other key stakeholders involved in data governance and access should be reported. For example, the data steward (i.e., oversight role of data management within an organization), the data custodian (i.e., person responsible for the safe custody, transport, storage of the data and implementation of business rules), the data processor (i.e., a third party that processes personal data on behalf of a data controller) and/or data collector (i.e., natural or legal person responsible for creating or capturing data) could be also mentioned there if deemed necessary.
- **License or agreement:** If there is any, the license or data sharing agreement associated with the dataset should be mentioned. If the data are open access, it should be also mentioned there.
- **Access procedures**: This item should describe who makes the decision, what is the process, what are the conditions placed on the data reuse, and how the conditions are communicated and enforced. Information about embargo, classified information or presence of sensitive personal data should also be included if applicable. If the data are classified according to certain security standards, the item should also include a reference to the classification model and what class has been assigned to the data according to that model.

Section: General data description. The purpose of the high-level overview section is to enable somebody interested in data reuse to determine if the dataset is relevant and potentially usable. It provides basic information to understand what the data are about, focusing on items meaningful for data in animal health.

An example of a template to be filled in order to provide a structured high-level description of a dataset used in veterinary epidemiology is provided in files named "2_GeneralDataDescription" available on Zenodo[32]. The template includes examples of possible answers, which are highlighted in yellow within the document.

Context.

- **Domain:** the general domain that the data relate to. It may sometimes overlap with the "keywords or subject" item reported in the previous section, depending on the platform used to publish the data.

  Examples:

- *A broiler production company may have a production database, and the domain might be 'Commercial poultry production'*
- *A veterinary diagnostic laboratory might have data in the domain of 'Livestock diagnostic testing'*
- *A government department may maintain a spatial database of aquaculture production leases with the domain of 'Registration of aquaculture production sites'*

- **Original purpose:** This is the purpose for which the dataset was initially created. Specifying the purpose of a dataset is useful to understand why and how the data were collected and therefore their potential values and biases. When a dataset has been created based on the integration or analysis of multiple other data sources, it is likely that the purpose of the integrated dataset is different from the purpose of the source datasets.

  Examples:

- *Diagnostic, research and regulatory specimen test results generated by a diagnostic laboratory in Poland.*
- *WOAH WAHIS's primary purpose may be to provide information on national disease status to prevent international spread of livestock diseases and facilitate trade in animals and animal products. However, it is based on reports from national animal disease surveillance databases, which may have different original purposes such as disease control and surveillance.*

- **Scope:** This is a definition of what types of data are eligible to be in the dataset, and what types are not. It is related to "What" is included or excluded from the data. In this section, any relevant fact about the data could be also reported. For example, for longitudinal data, reporting important events, which occurred during the time covered by the data and that may have impacted the data could be of interest.

  Examples:

- *Production data from all the Scottish production sites. Please note that the year YYYY was an unusually severe season of jellyfish, which has likely affected the fish mortality rate and maybe other variables that year.*
- *Organic farms and conventional farms with less than 10 pigs are not included in the dataset.*

- **Related documents:** Any digital object that might be connected to the data and could be useful for data reuse could be also mentioned in this section. For example, scientific papers describing or using the data could be cited there. Note that some of these objects could be also used as reference in other sections of the metadata.

Data resolution. This section aims to describe the data dimensionality but focusing only on data in low-dimensional spaces (i.e., limited number of variables per record easy to visualize), which remain very common in veterinary epidemiology. For data in high-dimensional spaces (i.e., number of features, attributes, or measurements is very large—and often much larger than the number of observations), the stakeholders would be invited to consider the addition of other attributes to properly describe their data to others.

- **Subject unit** identifies the subject unit from which individual data items are drawn. This defines what a row is, or in a non-tabular dataset, what defines a unique entry. This definition is essential to properly use the dataset, count records and understand and identify the duplicates. For flat datasets related to epidemiology, this would refer to the epidemiological unit of interest in the dataset such as an individual animal, herd, or region. If there are multiple subject units (e.g., relational database) connected by a hierarchical structure, it should be carefully highlighted and described there. Detailed information on each subject unit can be also provided directly within the data model (see below).

  Examples:

- *A dataset related to poultry production may be structured at the flock level with each flock being associated with a specific farm. A dataset focusing on dairy production could have data relating to individuals, which belong to specific groups of animals, which are themselves in different farms.*

- *A dataset on antimicrobial usage may contain data extracted from veterinary clinics, but the data have been aggregated by region. In that case, the subject unit would be "all the veterinary clinics in a given region".*

- **Spatial scope** is the geographical area covered by the data. It may be a country, an administrative subdivision (state, province, county etc.) or indirectly defined (e.g., the area of activity of the production company and all its contracted farms)[33].
- **Spatial resolution** identifies the spatial unit from which individual data items are drawn. This item may also be thought of as the level of precision used to identify a location[33]. Sometimes individual data items are not linked to any spatial information. In that case, it is important to make clear to future users that there is no spatial information available in the data.

Examples:

- *A dataset on wildlife density may include one estimate for each 5km-by-5km grid cell. The shapefile of reference for the grid is available in XXXX.*
- *Georeferenced farm locations may be recorded in latitude and longitude, (degrees-minutes-seconds format) to the closest whole second, or alternatively, in decimal degrees to the third decimal place.*
- *A dataset on antimicrobial usage may contain data extracted from veterinary clinics, but the data have been aggregated by region. In that case, the spatial unit would be "region defined as NUTS 2 according to the NUTS classification released in 2021* https://ec.europa.eu/eurostat/fr/web/nuts/history*".*
- *A dataset on antimicrobial usage may contain the data collected by veterinary clinics, but there is no information on the spatial localization of these establishments. In that case the spatial resolution of this dataset could be stated as "No spatial information available in the data".*

- **Period covered** is indicated by the start date and the end date. For an operational data set that continues to capture data, there is no (currently known) end date[33], but an update frequency should be reported in the next point.
- **Temporal resolution** refers to the precision used to measure when events occurred. Commonly, an event is recorded, giving a temporal resolution of one day. Some systems may use a timestamp to record when an event occurred (for example, when data were submitted) with a resolution of milliseconds[33]. Regulatory databases often summarize data into reporting periods, and therefore have figures aggregated by week, month, quarter or year.
- **Update frequency and mechanisms** - For operational, continuously growing datasets, data streams and any data receiving regular updates, this item would describe how and when new data are added to the dataset. This may be continuous (with no specified interval) when data are added to the dataset as soon as they are created, or be based on a regular reporting or data processing cycle (daily, weekly, monthly, quarterly or annually (in that case, specifying the period of the year is encouraged))[33]. If any backwards corrections to the previous versions of the data are made during the update, these should be also noted and clearly described in the data lineage section.

Examples:

- *Continuous update: The automatic feeder is recording data every time a cow visits it.*
- *Regular reporting or data processing: New data from production sites are automatically extracted once per month. Data from the previous month are also re-extracted at that time as these data could have been modified or corrected by the data collectors. Data of the previous month are therefore overwritten every month with the most up-to-date data.*

Quantity of data.

- **Number of unique entries** – Number of data points or unique entries present in the dataset ideally for each hierarchical level (e.g., number of farms, number of flocks, number of isolates per antibiotic testing and/or the number of raw antibiotic test results). For fixed dataset, the number of records in each component (e.g., sheet or table) should be documented. For continuously growing operational datasets, indicate the number of records present at a specified date instead.
- **Number of new records added or modified per update** - For operational, continuously growing datasets and data streams, it is useful to provide an estimate of the number of new records or new inserts added or modified per update. This may be exactly known (e.g., *one new submission by each of the 27 EU member states per month*), or estimated (e.g., *about 200 new disease outbreak reports per month*). Note that changes in the number of variables, rather than in the number of records, should be mentioned in the data lineage section.

Data lineage. Data lineage describes the journey data have taken to arrive at the current data set and includes tracking how it moves between formats, software and location, and what processing or changes occur at each step[3,34]. Not only does it explain how the data were captured, but this type of description also provides important

information on how to use and interpret the data, as well as any potential quality issues. (e.g., if several data collectors are involved, maybe their data could be handled separately). For a primary dataset, this may be relatively simple but it can be more complex for secondary datasets that have been integrated or derived from the primary sources.

The description of data lineage can be communicated in writing and/or by a flow diagram. For secondary datasets, a description of data lineage may be more complex. In this case, data flow diagrams may be particularly useful (see example of data lineage provided with the example dataset from[26]). In both cases, common considerations to note in the data lineage include:

- **Data collection agent and method** - The data collection and transcription method should be specified to ensure that future data users understand the potential biases associated with the data. Indeed, data collection and transcription can be automated (e.g., with sensors) or performed manually. When done manually through human observation, it may be linked to data entry bias and/or inconsistencies (e.g., keying errors such as typographical errors, transposed numbers, and misspelt names). In other cases, such as laboratory test results, knowing the precise laboratory method used is essential to understand the data. Similarly, when the collection agent is a sensor, knowing the actual product and links to the technical description is useful to understand potential bias in the data. Note that this section intends to provide high-level information on how the data were collected. When relevant, more specific details could be provided in the specific content description for each variable.
- **Data processing steps** - The number and nature of processing steps that data undergoes may also impact usability and quality and should be detailed as thoroughly as possible. For example, data cleaning and strategies for handling missing or incorrect data (such as replacing it with interpolated values), duplicated entries, and/or data anonymisation can alter the original data in ways that may not be useful for certain reuse purposes. Similarly, if outliers have been removed, the process used to identify the outliers and reasons for exclusion should be documented.

When organizations or people are involved in specific data collection or processing steps, attaching their function to the tasks they perform can be useful to allow data users to access additional information in case the information provided in the metadata is not sufficient. For long term preservation of the data, providing the analysis scripts or processing workflows used for data processing is the best way to share with others how the data processing was performed.

In some cases, it may be impossible or impractical to provide full details of original data collection systems. For example, the WOAH WAHIS dataset is based on national animal disease surveillance systems from all member countries. There is a large variety and complexity within the national systems, and in many cases, the details of the internal data lineage may not have been published or analyzed. Even if the process is impossible or impractical to describe, it does not mean that this item should be just left blank. Explicitly informing future data users that the data lineage is particularly complex and cannot be fully described is crucial, as it may impact data quality.

Example. Farm-level production data may go through the following steps:

- *Collection agent and method - All information is collected via human observation, except for the room temperature and humidity which are measured via the appropriate devices. The data are transcribed twice (from whiteboard to form once a day, and from form to spreadsheet once a week).*
- *Processing – Data are aggregated across rooms within a barn and entered into an Excel spreadsheet by office staff. No further data cleaning is done after that.*

Section: Specific content description. While the high-level overview helps to determine if the data are relevant and potentially useful for a specific purpose, a detailed description is required to reuse the data for the desired purpose. Despite its technical nature and the effort required for its development, this metadata component remains critical. Indeed, comprehensive knowledge of a dataset's content and quality is crucial for its effective usage.

An example of a template to be filled is provided in the files named "3_SpecificDataDescription" available on Zenodo[32]. It is partially based on the data model developed by the European Food Safety Authority (EFSA) as part of the SIGMA project to collect animal population data[35]. However, we modified this data model to make it more generic and therefore suitable for different types of datasets and to add considerations related to data quality. The template includes examples of possible answers, which are highlighted in yellow throughout the document.

Data model. The data model proposed in this framework is made of three elements, which are further described below.

- **Components of the data and their relationships**
  Clearly identifying any distinct components of a dataset is the first necessary step to document its content and structure. Each component or table of the dataset should be named so that it can be described afterwards. For example, if the dataset is made of multiple spreadsheets or multiple pages within a spreadsheet file, each of them should have a specific name.

Examples of components of a dataset include:

- multiple spreadsheets or multiple pages within a spreadsheet file
- tables within a relational database
- separate Word or PDF documents

When there are multiple tables, the relationship between them should be documented, including how they are linked (e.g., presence of unique identifiers). A more complex relational database will require a more comprehensive description of the relationships between the separate tables, but it will consist of the same essential information. For each table, identify which other tables it is related to, and which identifiers are present in both tables to allow them to be linked. Entity-relationship diagrams (ERDs) are often used to capture this information. A complex integrated database may have dozens or hundreds of tables, making an ERD hard to visualise. In this case, it is often more useful to present a series of partial ERDs that include only the relevant tables for a particular module.

Example:
A farm may maintain separate spreadsheets for:

- *daily observations (feed consumption, mortality, environmental conditions)*
- *production cycle figures (start date, end date, number of animals, average weight at sale, sale price).*

These sheets are related as they both refer to information about the animals in the barns during a production cycle. The link is established by two identifiers that appear in both spreadsheets:

- *The production cycle identifier (e.g., the 2nd cycle starting in 2023: 2023-02)*
- *The barn identifier (e.g., barn 4).*

Using these identifiers, it is possible to link the data between the two spreadsheets to calculate, for example, the average feed conversion ratio for each barn over the cycle.

- **Data elements**
  A data element is defined as a unit of data for which the definition, identification, representation (term used to represent it), and permissible values are specified using a set of attributes[36]. Structured data are often organized in rows and columns, but other approaches may be used (such as key/value pairs grouped into objects or arrays, as used in formats such as JSON or XML). In any case, metadata help to understand and use these data elements (columns of a table, or keys of a structured object).
  A common approach is to use a *data dictionary*, which is often organized as a table listing the attributes in each component. The information required includes:

- *Attribute name* – This is the name used in the data, be it the fieldname in a database, the column header in a spreadsheet or CSV file, or the key in a key/value pair. For example, "*dateOfSampleSubmission*" or "*totalAnimals*". Using a consistent and appropriate naming convention where phrases are written without spaces or punctuation and with capitalized words (e.g., camelCase or PascalCase) is strongly recommended to facilitate readability of the data.
- *Attribute label* - This is how the Name used in the data could be translated in plain language. For example, the label associated with the name "*dateOfSampleSubmission*" would be "*date of sample submission*". Similarly, the label associated with the name "*totalAnimals*" would be "*total number of animals*".
- *Attribute description* – This describes the element identified by a Name and a Label. The description should be as specific as possible, indicating any relevant reference to be mentioned. For example, in the case of results of a diagnostic test, it may include the exact name of the test used and its associated known sensitivity and specificity. When dates, times, timestamps and durations are used, the standards used to report them should be specified such as the time zone if relevant. Using ISO 8601 standard is usually recommended to report these data (e.g., October 31st, 2025 at 6 p.m. should be represented as 2025-10-31 18:00:00.000), but others could be used. In case the data contains "NA" or other values used to indicate the presence of missing values, the meaning of these special values should be also clearly specified there. In general, it is important to clearly distinguish between unknown and zero values.

- Examples:

  - *dateOfSampleSubmission*: the date upon which a sample submission form was completed by the person submitting the sample to the laboratory. The date has a length of ten digits and is presented in the following format: DDMMYYYY (D = day, M = month, Y = year).
  - *totalAnimals*: the total number of live animals of the specified species, of any age on the specified premises at the date of the visit. Missing values are indicated using -1
  - *testResult*: results obtained with the diagnostic test X (sensitivity: 80%, specificity: 92%) for further information about the diagnostic test, follow link X

- *Attribute type* – this indicates the data type of the element. The available data types and terminology vary slightly according to the software and format used, but the main types (and variants) are:
    - Number (integer, float, decimal)
    - Text (string, character, variable-length character)
    - Date (date-time)
    - Logical (Boolean, true/false)
    - Enumerated (a closed list of categorical values nominal or ordinal)
    - Other special field types. Some software may have dedicated field types for specific types of data, such as geographical coordinates, or a universally unique identifier (UUID). When represented in other formats, these values may be represented in one of the main formats listed above (such as text for a UUID, or a pair of float values for coordinates).

Additional information is needed for categorical attributes. Indeed, the meaning of each category should be properly described in a *vocabulary* (sometimes also called a *glossary* or *data catalogue*). The vocabulary can also be linked to an external reference when a standardized vocabulary is used. A *standardized vocabulary* or *thesaurus* is a vocabulary that has been established as a standard within a specific domain or community. When using standardized vocabulary, metadata should include a reference to the existing standard, rather than trying to reproduce it. Indeed, several ontologies or standard vocabularies have already been developed in the animal health sector, including FAANG for the animal genome[37], FoodOn for the food chain[38], ICAR the global standard for livestock data[39], ATCvet code for the classification of veterinary medicines[40], VetSCT for the veterinary extension to SNOMED CT for veterinary health records[41], AGROVOC for agricultural concepts[42]; ATOL, EOL and AHOL for livestock traits, environment and health[43]; AHSO on animal health surveillance[44]; and DECIDE's ontology initially developed for aquaculture and later extended to LivestockHealthOntology (LHO) for livestock diseases[45]. None of the existing standards is perfect because they only cover certain concepts and their quality varies (e.g., inconsistent Identifiers or Uniform Resource Identifiers (IDs/URIs), limited mapping to other vocabularies, sparse documentation, weak maintenance). However, reusing or extending them when required, instead of creating new standards, support data interoperability and improve long-term maintenance and comparison across studies. If a new vocabulary / thesaurus must be developed, it should ideally be done using existing standards to ensure interoperability. The key standards of the domain are the ISO 25964, the international standard for thesauri and interoperability with other vocabularies[46] and the Simple Knowledge Organisation System (SKOS), a common data model for sharing controlled vocabularies via the web in a machine-readable format.

Examples:

- *An element called "language", could be associated with the label "the language in which the respondent answered the questionnaire", and defined as "language of the respondent based on the ISO 3166-1 two-letter language code* - https://www.iso.org/iso-3166-country-codes.html"
- *A spatial unit called "NUTS" could be associated with the label "Nomenclature of territorial units for statistics" and defined as "NUTS as defined in the classification published by Eurostat in 2021* - https://ec.europa.eu/eurostat/web/gisco/geodata/statistical-units/territorial-units-statistics "

- **History**
  Metadata should include a change log documenting the date an alteration was made and any significant changes in the data collection, processing, storage, structure, dictionary, or vocabularies that may have an impact on the interpretation or analysis of the data. For example, a change in the data dictionary may result in a field having a different definition before and after the change date, which might need to be considered for the analysis. Any information on the impact of that change for future data analysis should be made as explicit as possible.

Examples:

- *Before the change date, the indices in column XXX started at 1, while after the change date they start at 0. Old data have not been recorded meaning that indices 1 before the change date should be considered equivalent to indices 0 in the data collected after the change date.*
- *After the change date, the category "Equidae" from the catalogue XXX was divided into two categories: "horse" and "donkey".*
- *Data collected by the farmer association Y were added to the dataset after the change date.*

Data quality. Data quality is an important consideration when deciding whether existing data are suitable for reuse for a specific purpose[47]. The first complexity arises from the fact that the definition of "good data quality" depends on the nature of the data being considered and its intended use. Providing an exhaustive description of data quality in the case of generic data reuse is likely out of reach. The purpose of this framework is therefore not to provide a thorough description of what is meant by good data quality, but rather to ensure that the metadata offers at least basic information related to data quality, allowing users to understand the key biases of the data.

The second issue comes with the wide variations reported in the list of criteria to be used for data quality assessment and in their definitions, making standardized evaluations of data quality very challenging[48–51]. For example, the term "validity" can be used to represent concepts such as "accuracy" or "completeness", which can also be seen as two distinct concepts. To overcome these issues, our framework focuses solely on generic quality criteria, independent of the task to be performed with the data, and on information not already covered in other sections of the framework. For example, the concept of "Timeliness" was not considered, as it is very specific to the task to be accomplished with the data. Moreover, when "timeliness" is understood as the period covered by the data, the concept is already covered under the item "Data resolution > period covered" of this framework. When understood as the general update frequency of the dataset, it would be covered by the item "Data resolution > Update frequency and mechanisms". When understood as the time between an event and its actual reporting (e.g., time lag between sampling and analysis by a laboratory), it would be often more appropriate to report it under the description of the corresponding attribute (e.g., "table.population > Attribute name > Attribute description"). In any case, it does not need to be repeated here.

In this framework, data quality is described based on four key items, which are further described below: data completeness, data accuracy, reliability, and quality assurance measures. These categories are not intended to support quantitative scoring, but to structure qualitative reporting. However, if additional activities have been performed to assess other data quality criteria, documenting them in the metadata would be considered of added value and should be encouraged. Such additional information could avoid future users from repeating data quality assessments which have already been performed. Moreover, if the quality of the data has changed at a certain time point in the dataset, it would be also important to highlight it in the description of the data quality. The files named "3_SpecificDataDescription.xslsx" and "3_SpecificDataDescription_DataQuality.csv" available on Zenodo[32] propose templates to document data quality information in a standardized way.

- **Completeness**
  Data completeness can be defined as the extent to which data are of sufficient breadth, depth and scope for the task at hand[52]. When the task at hand is unknown, completeness can be defined as a measure of the proportion of expected data that is actually present. It can be evaluated in two ways:

  - *Coverage* – measures the completeness at the *record level*. It can also be seen as the sampling fraction of the data. For example, if there are known to be x units of interest in the population, but the dataset only has data on y units, the coverage is y/x. This may be easy to calculate if reliable data on the population exists (e.g., all the farms from a given region were sampled). However, the true size of the population of interest may not be known. In that case, the coverage may either be unknown or only estimated especially if it changes over time due to the data collection process.
  - *Missing data items* – measure the completeness at the *column or field level*. For these items, the proportion of records with missing values should be calculated.

  Examples: Registered aquaculture sites may be audited periodically.

- *Coverage* - an audit dataset may contain audit results on only 20% of the true total number of sites before 01/02/2025, but it reached 70% after that date due to addition of a new data source.
- *Missing data items* – within that audit dataset, information about the person in charge of data collection might be missing in 20% of the records.

- **Accuracy**
  Data accuracy or correctness can be defined as the extent to which data are correct and/or certified by a third party[52]. It can be also understood as the inverse of the error rate, which provides an estimate of the proportion of the data that contains errors, including invalid or incorrect values. For the purpose of this framework, we restrict the definition of data correctness to invalid and incorrect data that were not removed during the initial data cleaning phase.
  Invalid values are data that are not consistent with the others in terms of format or content. They are due to data entry errors or system glitches. For example, a capital letter 'O' used instead of the digit zero in a field intended to store numbers would be considered as an invalid value. A text value that is not a valid member of the defined code list or vocabulary in a categorical field would also be considered invalid. Well-designed systems with fields proposing pre-typed entries for selection and quality assurance systems (documented in the *business rules*) should contain no *invalid data*. Systems that allow free-text entry of categorical values are, however, prone to typographical, capitalization and punctuation inconsistencies, requiring careful data cleaning before analysis starts. In any case, the proportion of invalid present in the shared data should be documented. If invalid data has been corrected before data sharing, the correction process should be documented in the data history section.

  Incorrect data are a different type of error where the data may have a valid value, but those values do not reflect the true state. Incorrect data may be due to various issues, including data capture and data entry errors resulting from typographical mistakes or selection of the incorrect categorical value. Identifying *incorrect data* is more difficult because it often appears to be correct. Quantifying them implies comparing the available data with a known point of truth or biological value, which may not always be easily available. Therefore, in practice, incorrect data can often be just defined as "*unknown values*".
  Examples: Registered aquaculture sites may be audited periodically.

- *Invalid data – the dataset contains no invalid data. However, be aware that the data contains three free-text fields to document the evolution of disease events. The content of these fields cannot be validated.*

- *Incorrect data – The dataset contains data from Belgium; however, 2% of the latitude and longitudes coordinate pairs reported are located in Poland. In addition, 1% of the cattle appear to have an age largely above the life expectancy of a cow (>50 years) indicating probable incorrect data.*

- **Reliability**
  Reliability provides an estimate of the confidence we can put in the results. In this framework, it covers the concepts of *data consistency* and *data credibility*.
  Data consistency represents the extent to which data are compatible or coherent with similar data[52]. This refers to internal and external discrepancies present in the data.

  - *Internal discrepancy* means that the data present in the dataset are not coherent, but the truth is unknown. For example, partial duplicated entries might be present in the data, but there is no way to identify which entry should be used for analysis.
  - *External discrepancy* focuses on the comparison between the data present in the dataset and those available elsewhere. In general, external data discrepancy may be difficult to assess in a standardized way, especially for users who lack in-depth knowledge of the data context. However, if such information is available, it should be reported.

Data credibility refers to the objective and subjective components of the trustworthiness of data. In that case, biases are suspected but the truth is unknown. Indeed, a data value may be valid, and it may reflect what the generator of the data intended, but it still may not reflect the truth. For example, for human input into a database, some analysis and judgement bias may be involved.

Examples:
- *Internal consistency:*

  - *10% of the mortality data present internal consistency issues: the percentage of mortality does not align with the category of mortality the flock has been assigned to (e.g., Poultry flock mortality was sometimes reported as 90%, but was at the same time labelled as "very low" in another column). This is an internal discrepancy between these two pieces of information and we are unable to know which one is true.*
  - *200.000 fish were reported dead at the end of the month, yet no fish were reported to be present at that location.*

- *External consistency:*

  - *Some latitude and longitude of the cases reported in the data do not match with the official description of the outbreak location. Similarly, some of the symptoms reported in the data are not aligned with the symptoms expected in this production type. In particular, drop in milk production has been reported by 1.2% of the beef farms present in the data.*
  - *The total volume of slaughtered fish reported by salmon farmers in [Region/Area] for [YYYY–YYYY] is [A] tonnes. For the same period, the volume reported to the Norwegian Tax Administration is [B] tonnes. This discrepancy ([A–B] tonnes; [(A–B)/A × 100]%) indicates an external inconsistency between data sources. Moreover, the dataset contains values that are inconsistent with regulations. For X% of locations, fish are reported present within one month of the end of the previous production cycle, whereas the regulations require a minimum two-month fallowing period between cycles.*

- *Credibility: A field that provides clinical diagnosis may be considered relatively reliable when a veterinarian provides the data, but less so when it is provided by a farmer without veterinary training. Similarly, a diagnostic test performed in difficult field conditions could be considered less credible than a test conducted in an accredited laboratory.*

- **Quality assurance measures**
  This final section describes any measures in place to ensure the quality of the data. Examples include one-off (for static datasets) or periodic (for operational datasets) data assessment and cleaning, or elements of system design, such as:

  - Business rules or built-in automated data validation steps which may have been applied to the data. Business rules are a set of rules related to data quality that are applied to the data. They are important to document how to reuse the data properly. For example, a business rule might be that all values above a certain threshold are considered abnormal and are replaced by a constant or set to missing. Whenever possible, business rules should be expressed and shared in a machine-actionable format (e.g., script), or at the minimum with a strict formalism using mathematical, statistical or conditional explicit formulas to ensure they are univocal.
  - Feedback systems which provide the data generator with an opportunity to validate the data contained in the dataset. For example, within online survey systems, the mechanisms put in place to validate the data and ensure internal consistency could be reported here.

Discussion

This paper presents the first community-developed rich metadata guidelines tailored to veterinary epidemiology datasets with ready-to-use templates and real-life examples. Thanks to the large number of stakeholders involved in the process, the guidelines should be easily understood by the members of the community (i.e. researchers in university but also scientists working in public or private organizations), and target the key information needed to support internal and external data reusability in this domain. The guidelines provide a single, comprehensive source of information, so researchers do not need to consult the metadata literature, which is often complex and overly technical for newcomers to the field.

If having good metadata increases the value of the collected data, creating metadata primarily benefits the data-owning organization by ensuring institutional memory, making internal data reuse easier and more cost effective, and therefore maximizing the value they get from the data they collect. Therefore, creating rich metadata, even if only in a private repository, should be driven by the direct benefits they provide to the data owner, and not by the willingness to share data with external stakeholders. In the modern era where data are becoming big and complex and essential for running all kinds of operations, relying solely on people's memory or availability to be able to work with data are a critical risk that public and private organizations should not be taking anymore. Having strong organizational incentives to support the creation of metadata is therefore critical to help individuals engage in the process even if they may not have a personal benefit for doing so. For example, in the public sector, including Data Management Plans (DMPs) requirements in grant contracts as done by numerous funding bodies is a first step, but, this should be associated with a specific allocation of resources to not strip them of their meaning and truly achieve data sustainability[53]. At a smaller scale, it would be also worth requiring metadata creation from students as part of their first projects during their studies and internships to make metadata creation a habit right from the start. This would be part of the development of their data literacy, which must be good enough to allow them later to handle several issues. Developing different career advancement criteria in academia, for example based on a dataset citation index, may also help researchers to dedicate time to this activity. In the private sector, the role of data as a valuable resource to better manage production processes, identify new opportunities or leverage business partnerships is not new[54,55]. We can only encourage stakeholders in the animal production business to invest more in good data management practices to be able to fully embrace this opportunity. In some special cases, researchers may want to reuse non-scholarly data for conducting research, but the data owner is not willing to invest in the creation of metadata. In that context, we believe the responsibility of generating the best possible metadata is with the researchers, who are intending to generate and share scientific knowledge based on this data. It is indeed part of their ethical responsibility as scientists to carry out and document their work in a transparent and accurate manner[56]. In this situation, it could be interesting to investigate whether sharing back and cross validating with data owners the metadata created during the research process could be considered as an added benefit, and whether this practice could help improve the accessibility of non-scholarly data for scientists.

Data sharing is not a prerequisite to the creation of metadata, and metadata can be published even when data sharing is restricted. Data sharing is indeed often complex especially in the agriculture domain because of privacy and confidentiality issues (see for examples[57,58]). On the other hand, metadata sharing is easier as it is not meant to contain any sensitive information and it still provides a high degree of "FAIRness" even in the absence of FAIR publication of the data itself by facilitating data discovery and including clear rules regarding the process for accessing the data[1]. The work conducted by[53] presents a number of recommendations on how and where to share metadata that can be applied to the case of data in veterinary epidemiology. We emphasize their recommendations of a two-stage approach using Git to continuously catalog data and documentation, and a well-established research repository like Zenodo to describe the Git repository with the catalog and point at its URL.

These guidelines intend to facilitate the process of creating rich metadata in veterinary epidemiology, but it is worth noting that it will be always more challenging to use them post-hoc, i.e., after data production, than concurrently with the data production[53]. We therefore cannot stress enough the importance of DMPs to make sure the creation of metadata is embedded in the workflow and is not an extra task raised only at the end of a project. Moreover, it should be clear that these guidelines will not necessarily help improving metadata quality scores as defined by some frameworks such as the one developed by the official portal for European data[59]. Indeed, such frameworks assess the quality of metadata only via a selected list of FAIR guiding principles focusing on concepts that are easy to evaluate in a standardized and domain-agnostic manner. They do not consider the "richness" of metadata and how they can be effectively used to support data reuse, as this task is necessarily domain dependent and a lot more difficult to achieve. The development of domain-specific guidelines such as those proposed in this paper is a step forward to be also able to assess this aspect of metadata quality, but more work would be needed to develop a formal evaluation process.

During the development of these guidelines, the researchers often navigated between two competing goals: providing detailed enough guidance while remaining flexible to adapt to a large range of datasets. Some authors would have preferred to go into even more detailed reporting, while others found the proposed list of items too long and overwhelming and may have preferred to prioritize some items. The version presented in this paper was deemed to provide a suitable balance between both objectives. It is important to note that these guidelines intend to be a first version that may evolve over time to better meet the needs of the professionals working in veterinary epidemiology, and stakeholders should feel free to adapt them to their specific or evolving needs. Indeed, additional items may need to be added to make specific kinds of data understandable by others and prevent potential misuse of the data. In general, when it comes to rich metadata, there is never too much information available so adding more should never be considered as an issue. However, to do so, we strongly recommend continuing using a structured approach and building on existing standards to ensure any new information added can be easily retrieved from the metadata by any stakeholders. Moreover, some parts of the guidelines

could be extended to be handled in a more standardized manner. For example, the guidance provided in this study to report data quality is a first step but could be further refined to better support good data quality check practices. At this stage, several definitions were kept broad on purpose to cover a large diversity of situations, which may lead to different interpretations and therefore different assessments. For example, assessing internal data consistency in practice is likely to remain challenging in the absence of more detailed information on how to proceed. To allow for such evolution and ensure the sustainability of these guidelines, this version is available in a version control platform[60], which is associated to a long-term data repository[32]. New versions developed should ensure they refer to these first versions and that differences are clearly highlighted to help stakeholders navigate potential multiple versions of the guidelines and associated templates.

For inexperienced professionals in our community, the guidelines may appear complex and lengthy at first. However, a more thorough examination will demonstrate that it would be difficult to reuse a dataset without having access to the basic information listed in the guidelines. During the development process, several rounds of discussions and reviews have led to retaining only the critical elements and removing anything that was agreed to be unnecessary. Some compromises had to be made between the authors to lead to these choices. It is expected that similar disagreements will also be raised by other professionals in our community. However, we expect that the potential barriers to the adoption of these guidelines are less related to their content, but rather to the true or perceived lack of resources and incentives to fill the templates. These concerns echo with what has been reported elsewhere when considering barriers to data sharing and the adoption of good data management practices in related domains[8,61–64]. First, it is important to highlight the fact that none of the information gathered through these guidelines is really new in the sense that anyone who is working with the data should be aware of it. Therefore, the information already exists "somewhere", the only effort required is to document it once in a standardized and centralized way. Moreover, not every type of data would require the same amount of work to generate metadata. For example, the development of sensor data in modern agricultural systems partially solved the question of metadata creation for these data which usually come with robust data schemas and quality descriptions. However, information related to the context of use of these technologies remains rarely available even if technologies are being developed to fill this gap[65]. In that context, only an ad-hoc collection of contextual metadata would be necessary to ensure proper data reuse as other information should be already made available by the manufacturer of the tool. Then, regarding the lack of resources to generate metadata, there is likely a biased perception that current practices are not costing resources, which has been shown to be largely false at least in the academic world[66,67]. As said by Mons in 2020[68]: "*if data are treated properly, researchers will have significantly more time to do research*". There is therefore a need to change the narrative from "the time lost by creating metadata" to "the time saved by creating metadata" in this community.

New technologies currently being developed can help address some of the barriers identified to the creation of rich metadata such as the lack of resources. In particular, the integration of artificial intelligence (AI) into metadata management processes has the potential to address many of the challenges faced by traditional systems, particularly in the areas of metadata generation, reuse, and lifecycle management[69]. These technologies can be particularly useful when metadata needs to be created post-hoc and many documents describing the data already exist and can be used as reference to generate rich metadata. Indeed, AI models can only be of use when the information already exists and is accessible in a digital format somewhere. Moreover, if AI models are effective at pointing out statistical anomalies, they are still struggling to identify issues of data quality which are domain-dependent (e.g., differencing a normal from an abnormal value is difficult without a deep contextual understanding of the data). Therefore, our guidelines are unlikely to become obsolete because of AI technologies, and they would rather be used as a source of information by these tools. AI may also help address the fact that rich metadata can probably only be unified up to a certain point. Indeed, our guidelines propose several items to be filled, but the content of each of them remain mostly textual and is likely to vary greatly among datasets depending for example on their type or structure. AI tools could help turn these partially structured data into fully structured data that can be more easily queried while not impairing human usability of the metadata collection templates. If AI provides powerful tools to save resources and support metadata creation, the users always need to keep in mind that such tools may also lack consistency, validation and may compromise data privacy and security[69]. They should therefore be used with great caution.

We chose to propose the templates associated with these guidelines in.csv and.xlsx files formats to facilitate their adoption by a community not yet very engaged in the process of making their data reusable. However, these formats are not satisfactory from a machine-readable perspective nor for long-term or large-scale data sharing. They should be considered a first but important step towards greater data FAIRness in the domain of veterinary epidemiology. Indeed, looking at the FAIR continuum proposed by[70], our guidelines mostly focus on steps 1 to 3 (i.e., Metadata exist in a human and basic nonproprietary machine readable format), providing a stepping stone for our community to reach step 4 (i.e., "linked data framework") in the future. The future development of these guidelines may therefore propose more advanced formats to better support machine-readability such as Resource Description Framework (RDF), JSON-LD or ontology-based systems, which can help automate metadata creation and make them more consistent and reusable. Such a stepwise process has proven successful, for example by the Cattle Barometer, a data-driven decision support tool which was initially developed based on direct data sharing using spreadsheets and CSV files received by email and manually uploaded and cleaned[71]. Later versions transitioned to an ontology-driven pipeline mapping records to the Livestock Health Ontology for creating interoperable and fully automated dashboards[72,73]. The workflow is documented in the DECIDE *Cattle barometer tutorial*[74]. Work on more advanced format for the guidelines has already been started and the first steps are available on various platforms (see for example GitHub[75]). Our guidelines are a first step in the transition to more advanced data management practices, but to go further, more awareness campaigns and fit-for-purpose training will be needed to increase the overall skills of the community on these questions. Increasing collaboration and relationships with computer scientists may also help to facilitate this transition.

Data Availability

All the templates associated with the paper are available on a version control platform 59, which is associated to a long-term data repository [60].

Code availability

No custom code was used.

Received: 6 March 2026; Accepted: 25 August 2026;
Published: xx xx xxxx

References

1. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
2. Jacobsen, A. *et al*. FAIR Principles: Interpretations and Implementation Considerations. *Data Intell*. 10–29 https://doi.org/10.1162/dint_r_00024 (2020).
3. Dublin Core™ Metadata Initiative. Metadata Terms.
4. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. https://doi.org/10.14454/mzv1-5b55 (2024).
5. Data Catalog Vocabulary (DCAT). (2024).
6. OECD. *Responding to Societal Challenges with Data: Access, Sharing, Stewardship and Control*. **342** https://doi.org/10.1787/2182ce9f-en (2023).
7. Callahan, T., Barnard, J., Helmkamp, L., Maertens, J. & Kahn, M. Reporting Data Quality Assessment Results: Identifying Individual and Organizational Barriers and Solutions. *eGEMs* **16** https://doi.org/10.5334/egems.214 (2017).
8. Tenopir, C. *et al*. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE* e0229003 https://doi.org/10.1371/journal.pone.0229003 (2020).
9. Brazma, A. *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
10. Rund, S. S. C. *et al*. MIReAD, a minimum information standard for reporting arthropod abundance data. *Sci. Data* **6**, 40 (2019).
11. Tenorio, F. A. M. *et al*. Filling the agronomic data gap through a minimum data collection approach. *Field Crops Res.* **308**, 109278 (2024).
12. Blumberg, K. L., McKillop, K., Pehrsson, P. R. & Fukagawa, N. K. Call to Action: A Need for Community-Driven Minimum Information Standards for Food Composition Data. *Am. J. Clin. Nutr.* https://doi.org/10.1016/j.ajcnut.2025.06.027 (2025).
13. Gregory, A. *et al*. WorldFAIR Project (D7.1) Population Health Data Implementation Guide. https://doi.org/10.5281/zenodo.7887385 (2023).
14. Sinaci, A. A. *et al*. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. *Methods Inf. Med.* **59**, e21–e32 (2020).
15. Meyer, A., Faverjon, C., Hostens, M., Stegeman, A. & Cameron, A. Systematic review of the status of veterinary epidemiological research in two species regarding the FAIR guiding principles. *BMC Vet. Res*. 270 https://doi.org/10.1186/s12917-021-02971-1 (2021).
16. Delavenne, C., Cameron, A., van Schaik, G., Frössling, J. & Faverjon, C. Reusability challenges of livestock production data to improve animal health. *Sci. Data* **12**, (2025).
17. Zinsstag, J., Schelling, E., Waltner-Toews, D. & Tanner, M. From "one medicine" to "one health" and systemic approaches to health and well-being. *Prev. Vet. Med.* **101**, 148–156 (2011).
18. Schwantes, C. J. *et al*. A minimum data standard for wildlife disease research and surveillance. *Sci. Data* **12**, 1054 (2025).
19. Kloeze, H. *et al*. A minimum data set of animal health laboratory data to allow for collation and analysis across jurisdictions for the purpose of surveillance. *Transbound. Emerg. Dis.* **59**, 264–268 (2012).
20. O'Connor, A. M. *et al*. Explanation and Elaboration Document for the STROBE-Vet Statement: Strengthening the Reporting of Observational Studies in Epidemiology – Veterinary Extension. *Zoonoses Public Health* **63**, 662–698 (2016).
21. Plate, K., Knüppel, S., Hornbacher, A., Greiner, M. & Müller-Graf, C. Development of a tool for rapid assessment of the evidence of human observational epidemiological studies with focus on risk of bias in the context of public health and risk assessment. *EFSA Support. Publ.* **21**, 8925E (2024).
22. McKechnie, I., Raymond, K. & Stacey, D. Identifying Inconsistencies in Data Quality Between FAOSTAT, WOAH, UN Agriculture Census, and National Data | Data Science Journal. https://doi.org/10.5334/dsj-2024-044 (2024).
23. Crystal-Ornelas, R. *et al*. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. *Sci. Data* **9**, 700 (2022).
24. Lortie, C. J., Vargas Poulsen, C., Brun, J. & Kui, L. Tabular strategies for metadata in ecology, evolution, and the environmental sciences. *Ecol. Evol.* **12**, e9245 (2022).
25. W3C RDF Working Group. RDF - Semantic Web Standards. https://www.w3.org/RDF/ (2014).
26. Śmiałek, M., Delavenne, C. & Faverjon, C. Production and health performance of a selection of polish poultry broiler flocks (Ross308) between 2018 and 2023. *Zenodo* https://doi.org/10.5281/zenodo.17406208 (2025).
27. N Veterinary Institute, Oliveira, V. H. S., Osnes, M. N. & Dean, K. R. Salmonid mortality and losses in Norwegian aquaculture. https://doi.org/10.5281/zenodo.17791861 (2025).
28. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
29. Cernava, T. *et al*. Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis. *Environ. Microbiome* **17**, 33 (2022).
30. SDMX community. https://sdmx.org/.
31. Fairchild, G. *et al*. Epidemiological Data Challenges: Planning for a More Robust Future Through Data Standards. *Front. Public Health* https://doi.org/10.3389/fpubh.2018.00336 (2018).
32. Faverjon, C. EpiMundi/faverjon-2026-vet-epi-contextual-metadata-guidelines. *Zenodo* https://doi.org/10.5281/zenodo.18889286 (2026).
33. Zeginis, D., Kalampokis, E., Palma, R., Atkinson, R. & Tarabanis, K. Statistical Challenges Towards a Semantic Meta-Model for Data Integration and Exploitation in Precision Agriculture and Livestock Farming. *7th Int. Workshop Semantic Stat. SemStats2019 Co-Located 18th Int. Semantic Web Conf. ISWC2019* https://doi.org/10.3233/SW-233156 (2019).
34. Mitchell, S. *et al*. FAIR data pipeline: provenance-driven data management for traceable scientific workflows. *Philos. Transact. A Math. Phys. Eng. Sci*. **380**, (2022).
35. Aminalragia-Giamini, R. *et al*. SIGMA - Population Data Model. https://doi.org/10.5281/zenodo.7520891 (2023).
36. Catalogue - ORIONKnowledgeHub. https://aginfra.d4science.org:443/web/orionknowledgehub/catalogue?p_p_auth=Rc853tgT&p_p_id=49&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&_49_struts_action=%2Fmy_sites%2Fview&_49_groupId=92976941&_49_privateLayout=false.

37. Harrison, P. W. *et al*. FAANG, establishing metadata standards, validation and best practices for the farmed and companion animal community. *Anim. Genet*. 520–526 https://doi.org/10.1111/age.12736 (2018).
38. Dooley, D. M. *et al*. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Sci. Food* 23 https://doi.org/10.1038/s41538-018-0032-6 (2018).
39. ICAR. Global standard for livestock data.
40. Norwegian Institute of ublic Health. ATCvet code (2026).
41. College of Veterinary Medicine - Virginia-Maryland. VetSCT (2025).
42. FAO. AGROVOC Multilingual Thesaurus (2026).
43. INRAE. Ontologies for Livestock.
44. Dórea, F. C. *et al*. Drivers for the development of an Animal Health Surveillance Ontology (AHSO). *Prev. Vet. Med*. 39–48 https://doi.org/10.1016/j.prevetmed.2019.03.002 (2019).
45. Noor, S. & Hostens, M. Livestock Health Ontology (2025).
46. ISO. ISO 25964 – the international standard for thesauri and interoperability with other vocabularies.
47. Oliveira, P., Rodrigues, F. & Rangel Henriques, P. A Formal Definition of Data Quality Problems. *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005* (2005).
48. Chen, H., Hailey, D., Wang, N. & Yu, P. A Review of Data Quality Assessment Methods for Public Health Information Systems. *Int. J. Environ. Res. Public. Health* **11**, 5170–5207 (2014).
49. Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **20**, 144–151 (2013).
50. Li, L., Peng, T. & Kennedy, J. A Rule Based Taxonomy of Dirty Data. *GSTF Int. J. Comput*. (2011).
51. Birkegård, A. C. *et al*. Building the foundation for veterinary register-based epidemiology: A systematic approach to data quality assessment and validation. *Zoonoses Public Health* **65**, 936–946 (2018).
52. Wang, R. Y. & Strong, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **12**, 5–33 (1996).
53. Fraga-González, G. *et al*. Affording reusable data: recommendations for researchers from a data-intensive project. *Sci. Data* **12**, 258 (2025).
54. Grimaldi, D., Fernandez, V. & Carrasco, C. Exploring data conditions to improve business performance. *J. Oper. Res. Soc.* **72**, 1087–1098 (2021).
55. Popovič, A., Hackney, R., Tassabehji, R. & Castelli, M. The impact of big data analytics on firms' high value business performance. *Inf. Syst. Front.* **20**, 209–222 (2018).
56. ALLEA - All European Academies. *The European Code of Conduct for Research Integrity*. (ALLEA - All European Academies, DE, 2023).
57. Jouanjean, M.-A., Casalini, F., Wiseman, L. & Gray, E. Issues around data governance in the digital transformation of agriculture: The farmers' perspective. *OECD Food Agric. Fish. Pap.* https://doi.org/10.1787/53ecf2ab-en (2020).
58. Gavai, A. K. *et al*. Agricultural data privacy: Emerging platforms & strategies. *Food Humanity* **4**, 100542 (2025).
59. European Data Portal. Metadata Quality Assurance. https://data.europa.eu/mqa/methodology?locale=en (2025).
60. Faverjon, C. faverjon-2026-vet-epi-contextual-metadata-guidelines. *GitHub* (2026).
61. Houtkoop, B. L. *et al*. Data sharing in psychology: A survey on barriers and preconditions. *Adv. Methods Pract. Psychol. Sci*. 70–85, https://doi.org/10.1177/2515245917751886 (2018).
62. Hughes, L. D. *et al*. Addressing barriers in FAIR data practices for biomedical data. *Sci. Data* 98 https://doi.org/10.1038/s41597-023-01969-8 (2023).
63. Perrier, L., Blondal, E. & MacDonald, H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. *PloS One* https://doi.org/10.1371/journal.pone.0229182 (2020).
64. Rainey, L., Lutomski, J. E. & Broeders, M. J. FAIR data sharing: An international perspective on why medical researchers are lagging behind. *Big Data Soc.* https://doi.org/10.1177/20539517231171052 (2023).
65. Basir, M. S., Zhang, Y., Buckmaster, D., Raturi, A. & Krogmeier, J. V. Meta Ag: An automatic agricultural contextual metadata collection app. *Smart Agric. Technol.* **12**, 101073 (2025).
66. Badolato, A.-M. Cost-Benefit analysis for FAIR research data – Policy Recommendations. *Ouvrir la Science* https://www.ouvrirlascience.fr/cost-benefit-analysis-for-fair-research-data-policy-recommendations-2 (2019).
67. European Commission. *Cost-Benefit Analysis for FAIR Research Data: Policy Recommendations*. https://data.europa.eu/doi/10.2777/706548 (2018).
68. Mons, B. Invest 5% of research funds in ensuring data are reusable. *Nature* **578**, 491–491 (2020).
69. Yang, W., Fu, R., Amin, M. B. & Kang, B. The Impact of Modern AI in Metadata Management. *Hum.-Centric Intell. Syst.* **5**, 323–350 (2025).
70. García-Closas, M. *et al*. Moving Toward Findable, Accessible, Interoperable, Reusable Practices in Epidemiologic Research. *Am. J. Epidemiol.* **192**, 995–1005 (2023).
71. Bokma, J. *et al*. European veterinary barometer for Bovine Respiratory Diseases: a tool showing diagnostic test results and geolocation of respiratory tract samples from cattle. *in European Buiatrics Congress and ECBHM Jubilee Symposium* **2023**, 179–181 (2023).
72. Noor, S. *et al*. Advancing precision livestock farming through ontology-driven interoperable health data management: extending the Livestock Health Ontology (LHO) for enhanced disease surveillance. in *11th European Conference on Precision Livestock Farming, Proceedings* 571–580 (The Organising Committee of the 11th European Conference on Precision Livestock Farming (ECPLF), 2024).
73. Noor, S., Bokma, J., Pardon, B., van Schaik, G. & Hostens, M. Agri semantics: developments to improve data interoperability to support farm information management and decision support systems in agriculture. in *Smart farms: improving data-driven decision making in agriculture* 75–96. https://doi.org/10.19103/as.2023.0132.05 (Burleigh Dodds Science, 2024).
74. Bokma, J., Noor, S., Pardon, B. & Hostens, M. Cattle barometer (2023).
75. VEMO: veterinary epidemiology metadata ontology (2026).

Acknowledgements

All the people in the different research teams involved tested the templates on their data. Special thank you to Guita Niang from INRAE, Robert Johansson from the Swedish Veterinary Agency.

Author contributions

C.F. conceived the study, developed the theoretical framework and performed the experiments. C.D. aided in the development of the framework and implementation of the experiments. C.F. supervised the study. K.R.D., V.H.S.O. and M.N.O created the salmon case study. C.D. created the poultry case study. G.v.S. acquired the funding and did the project administration. All authors took part of the development of the guidelines and associated templates and provided critical feedback and helped shape the research, analysis and manuscript.

Funding
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them. This work was specifically supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 101000494 (Data-driven control and prioritisation of non-EU-regulated contagious animal diseases [DECIDE]) and by the European Partnership on Animal Health and Welfare (EUPAHW) under the grant agreement No 101136346.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to C.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.