Digivet's Work Package (WP) 2 is dedicated to ***Creating Smarter Data Systems***.

This WP will explore existing and emerging digital innovations for collecting, managing, using, curating and preserving livestock data, illustrated using three case studies:

1) CS1 - *Food safety* (surveillance and control of foodborne diseases, with special focus on Salmonella Dublin in cattle);
2) CS2 - *Health security* (antimicrobial usage); and
3) CS 3 - *Economic security* (transboundary spread of contagious and exotic diseases of livestock, with special focus on African swine fever).

This deliverable reports the results of "Task 2.1 Data Collection, Curation, Preservation"; and the first sub-task in "Task 2.2 Data FAIRness".

DATA COLLECTION, CURATION, PRESERVATION

We have performed ***an inventory of data sources, datasets, data formats and uses*** for data associated with each of our case studies. The inventory is available at:

https://zenodo.org/record/6322649

Along with the inventory, the unique resource identifier above also contains a summary of the Digivet project and its work packages, as well as a summary of the case studies. This will provide contextualization of the data sources listed.

The inventory lists all data sources identified, in each of the partner countries (United Kingdom - UK, Sweden - SE, Denmark - DK, Norway - NO and Estonia - EE). Within the project, the inventory will give the detailed information needed about each data source to inform the next tasks. We expect many countries in Europe to have similar datasets and similar digitalisation needs; we therefore hope that public sharing of this inventory will allow other countries to identify digitalisation workflows developed in Digivet which can be reused for their purposes.

DIGITALISATION SCENARIOS

The three case studies in DigiVet were chosen to provide a broad representation of the different activities, goals and societal benefits of veterinary public health. These three scenarios will also provide working examples where challenges and opportunities are focused on different steps of the digitalisation workflows built to process data and generate information that can be used by decision makers (data to actionable information continuum) (Figure 1).

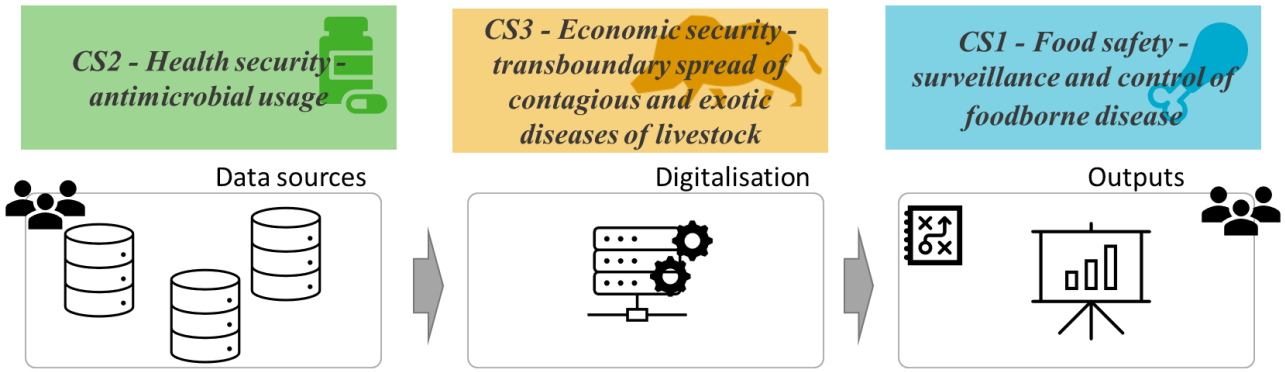

Figure 1. *The data to actionable information continuum can be broadly described into three main steps: data acquisition, analyses, and presentation of outputs to stakeholders. Each of the three case studies in DigiVet, presented in the coloured boxes on top, has stronger focus in one of these steps.*

In ***CS1 - Food safety***, the main achievement will be to build on existing prevalence maps of Salmonella Dublin that are used to guide dairy farmers in Denmark when purchasing new animals (http://www.kvaegvet.dk/). The aim is to calculate the actual risk of transmission of Salmonella Dublin between farms, which would lead to more focussed transmission-risk maps that would be more useful than the current aggregated prevalence statistics that do not distinguish between long-term infection and cycling of reinfection. A major focus of this CS will be the balance between utility and potential sensitivity of outputs that risk identifying individual farms as "super spreaders".

In ***CS2 - antimicrobial usage***, surveillance officials in Sweden, Norway and Denmark intend to investigate the challenges to access and meaningful use of antimicrobial sales and usage, as an aid to the surveillance of antimicrobial resistance. A major focus of this CS will be the legal and technical challenges in using antimicrobial prescription/sales data, and the potential gaps that exist between the desire for extracting information from the data and the information that is actually available in the data.

In ***CS 3 - transboundary exotic disease***, all partner countries will cooperate to develop a network representation of transmission routes for African swine fever (ASF), which can then be reused in different countries to inform decision making in disease prevention and control. A major focus of this CS will be reusability of software tools between countries that have access to similar but not identical data sources.

Besides providing complementary views of the challenges and opportunities in each of the three broad steps of the data workflows in veterinary public health (Figure 1), the three case studies also represent three different scenarios of digitalisation:

1) In CS1 – *Food safety*, data from regular testing of dairy cattle for Salmonella Dublin will be used to produce risk maps which could potentially be used to provide farmers with an assessment of actual transmission risk, which represents an improvement to the currently available spatially aggregated prevalence statistics. The outputs that CS1 will produce could potentially be made public, but respecting the privacy of stakeholders providing the data will require the development of workflows in which access to the data is granted on a "one-time basis", creating a snapshot risk map based on the data available at a specific time point at a sufficiently aggregated level, followed by a cessation of the access to the data source. Future updates would then require data access to be "turned on" again by the data owners themselves. In this scenario, data is shared once every time updates are to be performed, and we will call this a "***snapshot digitalisation***" scenario, where the challenges are mapped and solved for a "one time access" to the data, and the digitalisation workflow then produces results that reflect that one snapshot of the population at the time of the latest update.
2) In CS2 - antimicrobial usage, summarized snapshots are already shared between governmental agencies once a year. For the data on microbial usage to aid antimicrobial resistance surveillance, however, regular access to the data would be needed. The main challenge to meaningful use is interpreting the data correctly, and drawing the right assumptions regarding its population coverage and representation. All of these are challenges associated with data accessibility, and a meaningful use digitalisation scenario will require addressing the barriers to the establishment of a "***continuous digitalisation***" workflow which respects confidentiality norms.
3) In CS 3 - transboundary exotic disease, all countries will collaborate to develop a network representation of disease transmission routes for ASF which relies on data which should be available in all countries, but which cannot be shared across countries. Collaboration and cooperation across borders, in this case, depends on building "***portable software***", that each country can apply on their own data, to then share results and insights.

The meaning and importance of FAIR issues varies across these scenarios, as we discuss next.

DATA FAIRNESS

The original project description contained a plan to score all datasets identified in Task 2.1 (and listed in the repository linked above) based on the set of 14 metrics defined to quantify levels of FAIRness[1]. This was meant to identify the main gaps in FAIRness to be improved through the workflows developed in the next tasks.

Our desktop review and interviews with stakeholders demonstrated that for many of the case studies, those who analyse the data are only "data custodians", while the data is produced and/or owned by another organisation (which may also be governmental or not). In some cases, even within the data custodian organisation, the workflow is broken among different actors, with data analysts not having direct access or knowledge about data warehouses and the environments where extraction, transformation and loading (ETL) is applied to the data sources.

This disconnection between *data sources* and *datasets* (or rather, specific *distributions*[2]) results in issues such as lack of reproducibility and loss of context. We present the digitalisation challenges that came up within the three case studies, and how they relate to – and could be solved by improving – *findability*, *accessibility*, *interoperability* and *reusability* of data. The issues identified are often normative (data governance) and operational[3], rather than technical. For this reason, we have chosen to provide a general discussion based on these four overarching FAIR principles within the case studies, instead of a specific evaluation of the 14 technical attributes detailed in the framework.

CS1 - Food safety

Only Denmark will be taking part in this case study, and two main sources of data will be used: register of cattle farms and cattle movements; and regular herd-level Salmonella testing results. The details given in the survey and relevant for the FAIR discussion are presented in Table 1. The cattle registry is mandatory in all European countries, and many also have data regarding regular Salmonella testing in cattle farms. To exemplify the potential application of the case study in other countries, details for these two data sources are also listed in Table 1 for Sweden and Norway.

***Table 1**. Data sources applicable to case study 1, food safety.*

Country	Data source name	short description	Data source type	Data owner	Data custodian
DK	CHR database	Registry of herds, and cattle births, deaths and movements	SQL database	DVFA	SEGES
DK	S.Dublin test data	Farm-level test data for Salmonella Dublin	SQL database	SEGES	SEGES
NO	Individual cattle register	Registry of herds, and cattle births, deaths and movements	SQL database	Food Safety Authority (NFSA)	NFSA
NO	LIMS at NVI	Salmonella test results, all negative for S. Dublin so far.	SQL database	Submitter of sample	Veterinary Institute (NVI)
SE	CDB	Registry of herds, and cattle births, deaths and movements	SQL database	Board of Agriculture	Board of agriculture / SVA
SE	SVALA	Laboratory Information management System (LIMS)	SQL database	Submitter of sample	National Veterinary Institute (SVA)

[1] https://www.go-fair.org/fair-principles/ (findability, accessibility, interoperability and reusability)

[2] A "distribution" is a specific representation of a dataset. A dataset might be available in multiple serializations that may differ in various ways, including natural language, media-type or format, schematic organization, temporal and spatial resolution, level of detail or profiles (which might specify any or all of the above). Data Catalog Vocabulary (DCAT) - Version 2 (https://www.w3.org/TR/vocab-dcat-2/)

[3] Innovation and Big Data in Health Surveillance, available HERE.

UK	Cattle tracing system (CTS)	Online database of all bovine animals in Great Britain	Relational database	UK government	British Cattle Movement Service
UK	Salmonella database	Database of positive Salmonella samples from all animal species	Unknown	UK government	Animal and Plant Health Agency

A cattle register including location and information about all owners of cattle, as well as an individual recording of all cattle births, deaths and movements between herds (even herds of the same owner) is mandatory in all European Member States (EC 1760/2000 and EC 911/2004). In Denmark, this database is owned by the Danish Veterinary and Food Administration (DVFA) (https://www.foedevarestyrelsen.dk/). However, the database itself is curated by CGI (https://www.cgi.com/dk/en) on behalf of SEGES, with unrestricted access granted to DFVA under Danish law.

A specific scheme for surveillance of Salmonella Dublin has been implemented in Denmark since 2002 through SEGES (https://en.seges.dk/), who administers the scheme. In 2008, this was expanded to include a control plan with legislative measures restricting trade abilities of farms conditional on their Salmonella Dublin status introduced by the Danish Veterinary and Food Administration (DVFA) (https://www.foedevarestyrelsen.dk/), and in 2020 the. Accordingly, each cattle herd in Denmark has a "Salmonella status level" of either 1 (Salmonella free) or 2 (Salmonella positive), which is generated primarily based on routine antibody tests of bulk tank milk and is publicly visible at farm level via the online central livestock register (CHR; http://chr.fvst.dk). This binary classification was introduced in 2021 (https://www.foedevarestyrelsen.dk/SiteCollectionDocuments/Foder-%20og%20foedevaresikkerhed/Mikrozoonoser/Bekæmpelsesplan%20september%202020.pdf), although farms had previously been categorised as levels 1 (no antibodies detected), 2 (positive antibodies), or 3 (bacteriology positive), so in effect the more recent strategy combines levels 2 and 3. Although these Salmonella status levels are in principle publicly available, it is only possible to manually extract the current status of one herd at a time: batch downloads and historical data are not supported. It is also only possible to see the overall herd status: specific test results (principally antibody levels of bulk tank milk, but also including some individual animal antibody results and bacteriology results) are not publicly available. However, the private (farmer-owned) organization SEGES has agreed to share the data with researchers from the University of Copenhagen (KU) for the specific purpose of working on the Salmonella Dublin eradication scheme, which covers a number of research projects at the University of Copenhagen. The availability of Salmonella data for other countries varies: for the UK positive bacteriology samples are recorded in accordance with the notifiable status of the disease, but antibody statuses are not routinely recorded, and in Norway, samples are recorded but have all (so far) been negative for Salmonella Dublin.

All data sources listed in Table 1 are contained within relational databases kept by governmental organizations (or on their behalf) in fulfilment of legislative requirements, as a necessary source of information for the design and implementation of measures of animal disease prevention and control, and/or to record the results of these activities. All of them contain sensitive information about animal holdings and their owners, and are therefore protected under strong privacy protection regulations, with only specific elements of these data being made publicly available in ways that typically prohibit batch downloads of large datasets from multiple farms.

The applicability of the FAIR principles in this scenario are discussed in Table 2.

***Table 2**. Data Fairness and the case study 1, food security.*

Find-able	Meta-data for the data sources listed in Table 1 is not systematically documented nor indexed. These data sources are not meant to be public, but findability could be improved within the governmental setting. In none of the partner countries there is central documentation of the data sources available, along the food chain, which could contribute to food security. Different governmental organizations are aware of the data sources they own or maintain as data custodians, but there is no systematic way to identify complementary information in sister organizations within the same country.
Access-ible	Accessibility to the datasets is low for both humans and machines. For all the datasets listed in Table 1 authentication protocols are not in place outside the owner/custodian organizations. Data users are often not part of the process of extracting the data, and generally the queries and filters applied to extract and flatten the data are not transparently documented to these users, making it hard for machines to reproduce the process. For humans, the lack of meta-data documentation makes it hard to interpret the data at hand. A particular concern in this case study is providing data analysts with access to a snapshot dataset, while preserving, for the extracted flat table, the same data governance as the source database.
Inter-operable	The information in the cattle registry database is similar across European countries, yet a common knowledge representation language is not in place. The legislation establishes which type of events must be documented, but not the structure of the datasets, which varies greatly across countries. Documentation and comparison of these structures would be needed for models to be portable. For Salmonella testing data, surveillance methods are usually documented only as narrative descriptions of the programs in place, and changes in practices over the years are not systematically documented. The information following the testing data and the structure of the data varies considerably between different Laboratory Information management Systems, particularly for microbiological (culture) data. Documentation and comparison of these structures would be needed for models to be portable, although using primarily antibody test data may increase comparability between countries. As a result, interoperability is very low across institutions, as well as over time within the same institution.
Reusable	The main barrier to reusability in this case is a concern with the privacy of data owners, and a concern with the impact of detailed information generated from the original data. In the case of the risk maps, in particular, although the information about Salmonella Dublin status levels is public, there is a concern that making lower-level test result data available and transforming this information into a detailed risk map could lead to shaming and blaming of specific dairy owners interpreted as presenting a high transmission risk.

CS2 – Antimicrobial usage

The Norwegian Veterinary Institute has access and experience analysing antimicrobial usage (AMU) data. In this case study, their goal is to improve the evidence generation further, attending specific needs from those designing surveillance, such by incorporation of trend analyses and early aberration detection. The

goal is also to investigate how these data could be combined with other data sources, such as for instance results from surveillance of mastitis and antimicrobial resistance (AMR).

Sweden also has a program of AMR surveillance, but AMU have not been systematically analysed in animal health, and access to these data is restricted.

Data is mostly not shareable across countries, but partners in DigiVet will collaborate in developing workflows to digitalise their respective data sources and produce actionable information in aid of AMR surveillance. An exception is the Danish antimicrobial usage data available here (ttps://vetstat.fvst.dk/vetstat/login). This example of public sharing will be explored in this case study in comparison to the other countries.

Their identified relevant data sources are listed in Table 3.

***Table 3**. Data sources applicable to case study 2, antimicrobial usage.*

Country	Data source name	short description	Data source type	Data owner	Data custodian
NO	Mastitis test results	Microbiolocical test results including resistance	Unknown	Tine	Tine
NO	NormVet	Test results from AMR surveillance	SQL database	Food Safety Authority (NFSA)	Veterinary Institute (NVI)
NO	VetReg	Veterinary prescriptions and veterinary use of medicines	SQL database	Norwegian Food Safety Authority (NFSA)	NFSA
SE	FOTA	Antimicrobial sales	SQL database	eHealth agency	
SE	Djursjuksdata (DAWA)	Antimicrobial usage	SQL database	Board of Agriculture	Board of Agriculture
SE	SVALA	Antimicrobial resistance	SQL database	National Veterinary Institute (SVA)	National Veterinary Institute (SVA)
DK	VetStat	Antimicrobial prescription and sales	Relational database	Danish Veterinary and Food Administration (DVFA)	Danish Veterinary and Food Administration (DVFA)
DK	DANMAP	Danish Integrated Antimicrobial Resistance Monitoring and Research Programme	Unknown	Technical University of Denmark	Technical University of Denmark

As in case study 1, all data sources listed for case study 2 (Table 3) are relational databases kept by governmental organizations. In this case, however, a greater number of organizations in the same country are involved, and accessibility across organizations is a challenge.

The main barrier to meaningful use of the data is not just accessing the data, but interpreting its context, in particular across animal species. Antimicrobials can be purchased by veterinarians who then prescribe/administer them to individual animals. How and when antimicrobial prescriptions, sales and actual administration to animals are registered varies greatly across countries, and can vary within country between livestock and companion animals, or even private and state veterinarians. These differences are not well documented, and in some cases the database structures are not well suited to capture enough details to meaningfully interpret whether records represent prescription, sale/purchase or actual administration.

The applicability of the FAIR principles in this scenario are discussed in Table 4.

***Table 4**. Data Fairness and the case study 2, antimicrobial usage.*

Find-able	Meta-data for the data sources listed in Table 3 is not systematically documented nor indexed. Challenges of findability are the same described for case study 1, but worsened here by the multitude of institutions which can hold data along the continuum.
Access-ible	Accessibility to the datasets is low for both humans and machines. As described in case study 1, data access is not based on specified protocols, and there is little documentation of the process of extraction and transformation applied when generating specific datasets for sharing across institutions. An exception to this is in Denmark, where an online interface to the underlying VetStat database is provided for use by any Danish (or EU) citizen to download antimicrobial prescription/sales data at the level of farm and active ingredient.
Inter-operable	All datasets evaluated lack an explicit knowledge representation. The context of the data registered is not fully clear to data users, especially when data is shared across agencies. More concerning, however, is the potential *loss* of context details during data recording. During interviews with stakeholders it became clear that some of the data structures have not been optimized for all animal species. There are inherent differences between the practices of prescription and usage across livestock and companion animals. When data structures are not built to accommodate that, users will make their own assumptions about how to summarize or transform the data to fit into the format of registries, and these individual decisions cannot be reproduced or corrected to bring back the lost context. Understanding the meaning of data, and how – or even whether – different sources can be linked will be the main challenge to address in this case study. Privacy issues, the challenges described above and how they differ across countries, and the lack of harmonized agency responsibilities and database structures across countries makes cross border interoperability virtually impossible.
Reusable	Data are not reusable and in this case even digitalisation workflows developed may not be reusable across countries. Collaboration across DigiVet partners will however focus on understanding the potential for meaningful use despite all these barriers.

CS 3 - transboundary exotic disease

This case study is driven by the decision needs, rather than the data available. Modellers across the partner countries will collaborate to develop software tools that generate a network representation of disease transmission risk between commercial pig herds based on inputs related to animal movements and (where relevant) wild boar mediated spread. This network representation can be used along with pseudo-anonymised herd-level data to inform ASF spread models to address the decision-making needs of surveillance officials in countries under different risks of ASF introduction. As the tools are developed, data needs will be identified, and the availability of the data sources mapped.

Potential datasets have been listed by the different partner countries in DigiVet. In Sweden, a project is ongoing to develop a risk assessment of the risk of introduction of ASF into the wild board population. The transmission risk network in Sweden will make use of public datasets, such as meteorological data, road

networks, terrain maps and records of traffic accidents involving wild animals in order to assess the risk of indirect transmission of disease between commercial pig herds mediated via wild boar. Data directly related to the wild boar populations will be concerned with wild animals. As these animals are not owned by any individual, these data are generally more open and accessible than data from livestock animals.

In DK, UK, NO and EE the data sources will focus on transmission risks via movements of the domestic pig population – and SE will also use these data sources in addition to those listed above. These data sources are more closely related to those described for case studies 1 – they are not shareable across partner countries, but all countries are likely to have similar data sources, although these data, their metadata and knowledge representations are not well documented and indexed.

Cross-border collaboration is an essential component of preparedness against transboundary exotic disease. As data sources of animal health and animal production registries are not shareable across countries, we suggest that the focus here should not be on the FAIRness of the data, but on the construction of FAIR models which can be shared across countries, allowing resulted to be compared.

We discuss our idea of a *FAIR data analysis* model in Table 5.

Table 5. ***FAIR data analysis models***, *adapted from the FAIR data principles*[4].

Find-able	Just as described in the FAIR principles for data, model *codes* and *metadata* should be easy to find for both humans and computers. All code, along with metadata describing the inputs and outputs should be stored in a globally unique and persistent identifier (F1), with version control. These metadata should contain a description of the model language, usage, target population, required data and model outputs (F2); clearly and explicitly include the identifier of the model and version available (F3); and be registered or indexed in a searchable resource (F4).
Access-ible	We can again apply the FAIR principles by interpreting data, in this case, as the data analysis *codes*, and *metadata* as all necessary descriptions to understand in which contexts the model is applicable, which kind of outputs it can provide, and what data sources structures are necessary to feed the analyses. (A1) *codes* and *metadata* are retrievable by their identifier using a standardised communications protocol. The *programming language* used should be open, free, and universally implementable. (A2) Metadata are accessible, even when the *codes* are no longer available (or out of date)
Inter-operable	We propose that for models to be interoperable, the data used to develop the software don't need to be stored and made public along with the models, but the (I1) *content and structure of the data sources* needed to feed the model must be documented using a formal, accessible, shared, and broadly applicable language for *knowledge representation*. The metadata should therefore contain knowledge representations of the data *inputs* and *outputs* using (I2) vocabularies that follow FAIR principles.
Reusable	Lastly, models should be made reusable by making sure that codes, as well as knowledge representations of inputs and outputs, are (R1) richly described with a plurality of accurate and relevant attributes, including in particular, (R1.1) clear and accessible data usage license, (R1.2) detailed provenance and (R1.3) adherence to domain-relevant community standards.

[4] https://www.go-fair.org/fair-principles/

Conclusions and way forward

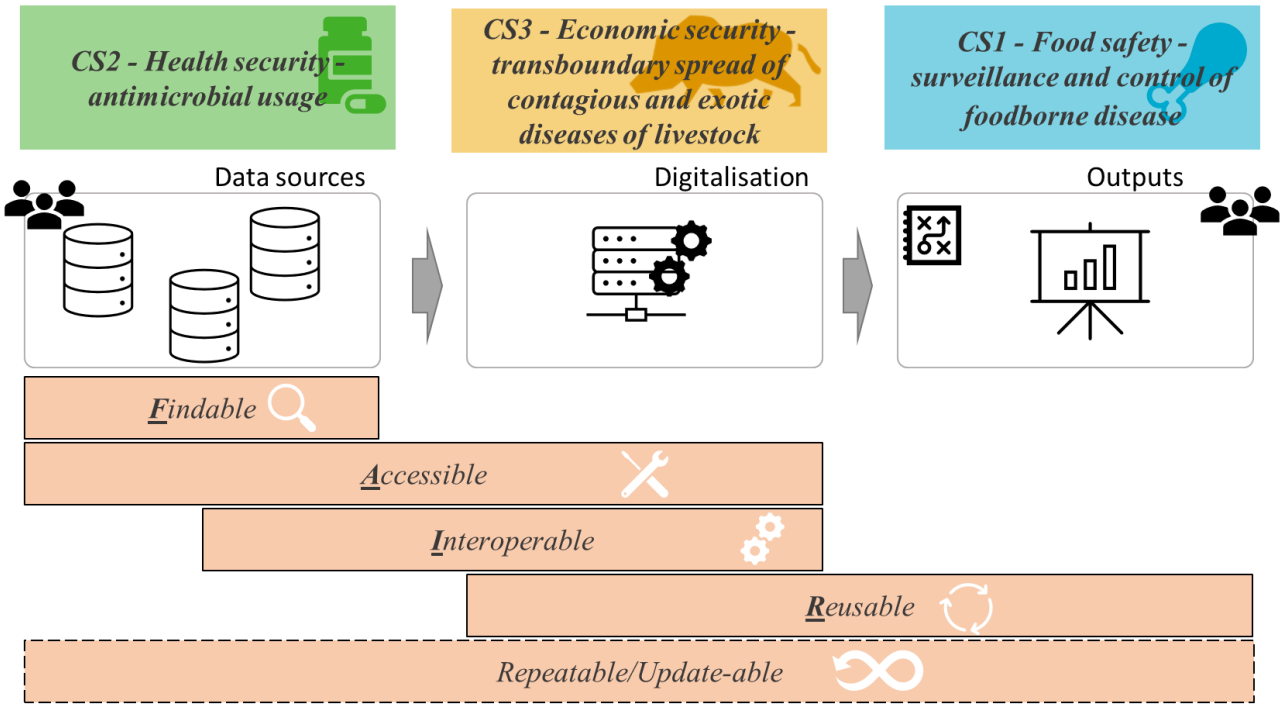

Figure 2. *DigiVet case studies and the importance of the FAIR data principles along the data to actionable information continuum. Building future-proof digitalisation workflows (that can be reused over time, and even updated without loss of compatibility) will require improvements along the entire continuum.*

The FAIR principles have a pivotal role in the management and stewardship of scientific data[5], and are generally more applicable to situations in which data openness and sharing can be promoted.

While the 14 specific FAIRness attributes listed in the framework are mainly focused on technical solutions to achieve data findability, accessibility, interoperability and reusability, the general four principles can be helpful in discussing and addressing normative and operational challenges in governmental settings, and to guide establishment of sustainable, future-proof digitalisation workflows in any settings. Privacy barriers do not need to result in "silo-ing of data".

Case studies 1 and 2 are literally *data-driven*, that is, we are working in data sources we already know exist, and trying to develop workflows for meaningful use and continuous generation of actionable information for surveillance activities already in place. In these scenarios, findability is not a bottleneck.

Accessibility and interoperability, for humans as well as machines, are both impaired by the absence of documented data structures and knowledge representations that can provide interoperability. These issues impact even the historical interoperability within the same institutions, making it impossible to compare data and results across the years.

Lack of accessibility and interoperability also reduces the potential for data reuse outside its primary purpose, within the owner agency or by other actors. Building workflows that are shareable would avoid duplication of work, and allow collective growth of knowledge and digitalisation capabilities. This is the

[5] https://www.nature.com/articles/sdata201618

promise of the linked data model. The important lesson to incorporate into digitalisation workflows is that data doesn't have to be shared in order for it to be "linked". If knowledge models are made explicit, the digitalisation workflows can be adapted and reused across agencies and countries, and results compared.

In CS3, which is question (rather than data) driven, reusability, or rather portability, led to the conclusion that FAIRness should be applied to software rather than data.

As we studied the current scenario for each case study, it became clear that barriers to FAIRness are usually dealt with mostly as a "single use" scenario, with issues related to finding, accessing, and integrating the data being solved once when the data are needed, but not implemented workflows that can be reused and updated over time.

In the next steps of the DigiVet project, we will focus on developing future-proof digitalisation workflows. CS2 will focus on data findability, accessibility and meaningful use. CS1 will investigate technical solutions to data accessibility that respect data governance. CS3 will develop reusable software that can be used to support disease spread models. In all case studies, adoption of the linked data model principles will be useful to improve interoperability and workflow sustainability.