Funded under the H2020 programme

Monitoring Outbreak events or Disease surveillance in a data science context

Project number: 874850

Horizon 2020

SC1-BHC-13-2019

Type of action: RIA

Deliverable D2.3

Title: Blueprints for epidemiological data collection and management

Due date of deliverable: 30/06/2023

Actual submission date: 30/08/2023

Start date of the project: January 1st, 2020 **Duration: 48 Months**

Leader of the Deliverable: ISS

Dissemination Level	
PU Public	X
PP Restricted to other programme participants (including the Commission Services)	
RE Restricted to a group specified by the consortium (including the Commission Services)	
CO Confidential, only for members of the consortium (including the Commission Services)	

Project Information	
Project Acronym	MOOD
Project Full Title	MOnitoring Outbreak events for Disease surveillance in a data science context
Name of the funder	European Commission – H2020
Call Identifier	H2020-SC1-2019-Single-Stage-RTD
Topic of the call	SC1-BHC-13-2019 Mining big data for early detection of infectious disease threats driven by climate change and other factors
Grant Agreement Number	874850
Project Duration (YYYY/MM - YYYY/MM)	2020/01/01 – 2023/12/31
Project coordinator (Name of institution)	CIRAD
Cirad Project scientific leader (name, RU, e-mail)	Elena ARSEVSKA, ASTRE, elena.arsevska@cirad.fr
Project goals	MOOD aims at using state of the art data mining and data analytical techniques of disease data, Big data, and contextual data originating from multiple sources to improve detection, monitoring, and assessment of emerging infectious diseases (EID) in Europe. MOOD will establish a platform for mapping and assessment of epidemiological and genetic data in combination with environmental and socio-economic covariates in an integrated inter-sectorial, interdisciplinary, One health approach. More precisely, MOOD will develop: 1. The epidemic Intelligence community of practice to identify user needs of end-users i.e. national and international human and veterinary public health organizations; 2. Data mining methods for collecting and combining heterogeneous Big data; 3. A network of disease experts to define drivers of disease emergence; 4. Data analysis methods applied to the Big data to model disease emergence and spread; 5. Ready-to-use online platform tailored to the needs of the-end users and complimented with capacity building and network of disease experts to facilitate risk assessment of detected signals. MOOD outputs will be co-constructed with end-users at public health agencies to assure their routine use during and beyond project duration. They will be tested and fine-tuned on a set of air-borne, vector-borne, multiple-transmission route diseases, including anti-microbial resistance and disease X. Extensive interactions with end-users, studies into the barriers to data sharing, dissemination and training activities and monitoring of the impacts and innovations of MOOD outputs will support future sustainable use.

Key words	Infectious diseases, big data, epidemic intelligence, one health, impact, environmental changes, climate changes, user needs, socio-technical innovation
Project partners (Name of institutions)	CIRAD, ITM, FEM, ETH, INESC ID, ERGO, SIB, INSERM, ULB, KU LEUVEN, UM, SOTON, AVIA-GIS, MUNDIALIS, IRSTEA, OPENGEOHUB, UOXF, ISS, THL, GERDAL, IPHS, ISCIII, ANSES, INRAE, ISID

Executive Summary

The document provides an overview of the legal background, policies, and procedures for accessing and handling epidemiological data on human and animal diseases from international databases in the context of the EU project "MOnitoring Outbreak events for Disease surveillance in a data science context." The project aims to improve the detection, monitoring, and assessment of emerging diseases in Europe by harnessing data mining and analytical techniques. The document also discusses the EU GDPR Regulation (EU) 2016/679, which expands individuals' rights to control data collection and processing, and places new obligations on organizations' controllers and processors. The MOOD framework and platform are developed to identify and manage infectious diseases based on their importance, public-health importance, economic cost, sensitivity to environmental and socio-economic change, and diversity of disease systems. Data is collected from various sources, including official surveillance data, epidemiological data, genomic data, and covariates that drive disease distribution and transmission. The project has established an Ethics and Data Protection Advisory Board to manage ethical aspects and prioritize data collection. The project also uses literature reviews to assess the availability of suitable datasets.

Keywords

zoonoses, epidemic intelligence, data, data protection, data sustainability

Epidemiological data sources, access procedures and collection

Claudia Cataldo[1], Maria Bellenghi[1], Luca Busani[1], Annapaola Rizzoli[2]

1) *Istituto Superiore di Sanità, Rome, Italy*
2) *Fondazione Edmund Mach, Trento, Italy*

TABLE OF CONTENTS

1. GENERAL POLICIES ON DATA ACCESS AND USE

1.1 Purpose of the document

The present document provides an overview of the general legal background, policies and procedures on access and handling epidemiological data on human and animal diseases from international databases in the framework of the EU project “MOnitoring Outbreak events for Disease surveillance in a data science context (MOOD ID 874850). The MOOD project aims at harnessing data mining and analytical techniques to data originating from multiple sources to improve the detection, monitoring, and assessment of emerging diseases in Europe. MOOD is developing a framework and a platform for data analysis and visualization to enable real-time analysis and interpretation of epidemiological and genetic data in combination with environmental and socioeconomic covariates in an integrated cross-sectoral and interdisciplinary One Health approach; project results are expected to be used routinely during and beyond MOOD. To ensure this continuity, the necessary data must be secured, and made available, and their management must be sustainable.

This version D2_3_V2_30082023 contains all the relevant information concerning the data collected and used in the MOOD project and includes the important outputs of the Demonstration and Interaction with the End Users workshop held in Helsinki on the 28th of June. These parts are integrated with the results of the MOOD internal discussion held in Helsinki on the 29th of June presented in the official minutes of the two meetings.

1.2. Legal Background: The EU GDPR Regulation (EU) 2016/679 (General Data Protection Regulation)

The General Data Protection Regulation (GDPR) is a pan-European data protection law that superseded the EU’s 1995 Data Protection Directive, and all member state laws based on that directive on 25 May 2018. The GDPR expands the rights of individuals to control how their personal data is collected and processed and places a range of new obligations on organizations' controllers and processors) to be more accountable for data protection. The GDPR also gives member states limited opportunities to make provisions or derogations for how the Regulation applies in their country; Ireland has done so via its Data Protection Act 2018, which came into effect on 25 May 2018. GDPR applies to all EU organisations that collect, store, or otherwise process the personal data of individuals residing in the EU, even if they are not EU citizens, and non-EU organizations that offer goods or services to EU residents monitor their behaviour or process their personal data.

The GDPR outlines six data protection principles that summarise its many requirements: Lawfulness, fairness, and transparency; Purpose limitation; Data minimisation; Accuracy; Storage limitation; Integrity and confidentiality. This is the only principle that deals explicitly with security. The GDPR states that personal data must be processed to ensure appropriate security, including protection against unauthorised or unlawful processing and accidental loss, destruction, or damage, using appropriate technical or organisational measures.

The GDPR is deliberately vague about measures organisations should take because technological and organizational best practices constantly change. Organizations should encrypt and/or pseudonymize personal data wherever possible, but they should also consider whatever other options are suitable.

MOOD's compliance with the General Data Protection Regulation (2016/679) is paramount to processing a significant quantity of data. Therefore, task 7.4 (Legal and Data Protection Framework) of the proposal will assess how to implement the relevant clauses from the GDPR, ensuring the protection of natural persons through the processing and free movement of personal data. The National Data Protection laws of the EU Member States will also be analyzed in case specific provisions apply. National transposition of the EU Privacy Directive (2002/58/EC) or any relevant legislation will be covered. Special attention will be put on the legislation of the non-EU partners (located in Switzerland, Serbia, and the United States of America) to ensure the data processing and sharing are in line with the GDPR principles.

1.3 MOOD Data Policy and Ethics

Background

MOOD framework and platform are developed considering a list of infectious diseases selected according to their importance: zoonotic diseases with wild vertebrate hosts, public-health importance, economic cost, sensitivity to environmental and socio-economic change, and diversity of disease systems.

Epidemiological data, big data, and contextual data (drivers) needed for the early detection, monitoring, and assessment of these diseases have been identified in the early stage of the project. Later on, the data were collected, including (1) official surveillance data for each of the identified diseases at the local, national, continental, and global scale; (2) epidemiological data gathered through text mining of social media, internet queries, and news media; (3) genomic data related to the relevant pathogens; (4) point prevalence data for anti-microbial resistance (AMR) in food animals. and (5) covariates that drive disease distribution and transmission.

Two categories of data will be required to feed MOOD modelling activities: training data for the target (dependent) variables (for example, diseases and disease vectors data) to be modelled and the covariate data that provide the independent variables used to drive the modelling process. For example, environmental indicators, disease, or vector hosts. The disease data were drawn from several sources, web-scraping and text mining from several sources, published literature and open-access reports released by public health agencies (i.e., ECDC, WHO, EFSA). A survey among the MOOD partners among the public health Institutes, to map the availability and sharing policies of the national data about the selected diseases was carried out.

Identifying the covariates needed for modelling activities within MOOD required the precise definition of which data. A very wide range of different data types and sources were analysed to ensure the availability of covariates.

Detailed specifications for both the dependent and independent variables are needed to ensure that the models are fed with consistent and appropriate inputs. For each type of variable, a single data source should be chosen according to data availability and quality. This is particularly important for denominators such as population and major climatic variables such as temperature, for which many possible sources exist.

To manage the ethics aspects, the MOOD project has established an Ethics and Data Protection Advisory Board with two external ethics advisers specialised in data protection and research ethics, and an ethics report is regularly produced by the external ethics advisers to independently report how the ethical issues have been handled.

Data Prioritisation

The prioritization of the data to be collected was set considering the diseases listed in the MOOD project, namely:

- Influenza A;
- tick-borne encephalitis and Lyme borreliosis
- West Nile and Usutu
- Chikungunya, dengue and Zika
- Tularaemia and Leptospirosis
- Antimicrobial-resistant bacterial

- Unknown pathogens (disease X)

Covariates and other relevant data

Prioritization of the covariate selection is a four-stage process to identify: a) what covariates are required by project analysts – for modelling the introduction, spread and establishment of the target diseases; b) what datasets are already available to partners within existing archives; c) what covariate data are needed that are not already available, and should be acquired; and d) if the necessary data cannot be acquired, establishing whether proxies can be used or derived from existing datasets.

In parallel, the literature reviews implemented for the listed diseases provided information on the drivers for each disease. These are then screened to assess whether suitable matching datasets are already held by project partners, after which they will need to be sourced, acquired, and processed to project standards.

Some of the prioritized disease data were also collected through text mining. The prioritization process for these methods also concerns identifying search terms, described in the following section. The genomic data collection supports the phylodynamic analyses in WP4. The prioritization of genomic selection depends on the aims of the specific study that will use the genome sequences. Phylogenetic characterization is generally not subject to sampling bias and often requires the most similar genome sequences to the ones of interest. Population genetic inference however uses models that assume random sampling and therefore require a selection process. If the aim is to reconstruct spatial spread, e.g. using phylogeographic methods, then the location of sampling is a critical determinant of the collection process. In this case, genomic sequences are often down-sampled in locations for which many sequences are available. This type of prioritization can be guided by prevalence estimates for the pathogen under investigation. Finally, suppose the genomic reconstructions involve time-calibrated phylogenies for rapidly evolving viruses. In that case, sampling time information needs to be taken into account, as well as spatiotemporal biases in sampling if the spatial component is also of interest.

1.4 Sustainability in Post-project Data Management

Covariates and Diseases

MOOD's ethos requires that analytical outputs are produced in a way that is as automated and streamlined as feasible. This means that covariate datasets that have a sustainable supply, with scripted processing, will be prioritized over those that require manual updating and pre-processing. Any acquisition or processing costs will also be factored into an assessment of the sustainability and feasibility of use. Conventional disease data are provided by MSs as surveillance mandatory activity and regularly updated at least once a year. For disease data acquired by text mining, the supply is provided by public domain web-based sources which, barring legislative restrictions, will always be available. Also, by their nature, text mining searches are largely automated, which also promotes the sustainability of supply and update. As for covariates datasets, genomic datasets that have a sustainable supply, with scripted processing, will be prioritized over those that require manual updating and pre-processing.

2. DESCRIPTION OF DATA SOURCES AND DATABASES

2.1 European Centre for Disease Prevention and Control (ECDC)

2.1.1 General description of ECDC

ECDC was established in 2004 (Regulation (EC) No 851/2004) to enhance the capacity of the Union and the Member States to protect human health through the prevention and control of communicable diseases in humans.

The ECDC's mission is to identify and assess current and emerging threats to human health from communicable diseases, making related information easily accessible. In addition, ECDC provides science-based recommendations and support in coordinating the response at the Union and national levels, especially in case of cross-border interregional and regional events.

The ECDC works in collaboration with competent bodies of the Member States for data collection, exchange and in case of transboundary outbreaks or emerging diseases.

ECDC, according to current EU and International legislation, is the owner of copyright and database rights and contents. Information and documents made available on ECDC web pages and for which ECDC owns the copyright are public and may be reproduced, adapted and/or distributed, totally or in part, irrespective of the means and/or the formats used, provided that ECDC is always acknowledged as the source of the material. Such acknowledgement must be included in each copy of the material. Citations may be made from such material without prior permission, provided the source is always acknowledged.

2.1.2 European surveillance system for infectious diseases (TESSy)

TESSy is The European Surveillance System for communicable diseases. It is managed by the ECDC which is responsible for collecting, storing, and disseminating surveillance data on communicable diseases at the European level. TESSy is a Web application that enables nominated users to upload surveillance data from their country to the TESSy database, which is the repository for these data. The TESSy Web interface also enables users from the participating countries and ECDC, to access historical data in the form of standard reports and ad hoc queries, through the Query data feature. The data are collected according to the Commission Implementing Decision (EU) 2018/945 of 22 June 2018 on the communicable diseases and related special health issues to be covered by epidemiological surveillance as well as relevant case definitions and the technical operating guides on reporting data to TESSy by member states and the metadata set that contains the TESSy metadata set specifications together with a summary of the current data sources.

The data collected are publicly available in different forms:

Dashboards display data which users can visualize in graphs, figures and maps. These are frequently updated during outbreaks.

Interactive databases allow users to select sets of data and visualize them in tables and maps. Some have fixed, static datasets, while others have dynamic datasets based on annual surveillance data regularly updated.

ECDC datasets are directly downloadable in various formats (e.g. .xlsx, .csv, .json) on selected diseases and topics, such as COVID-19, monkeypox, antimicrobial consumption, measles, rubella, and West Nile virus.

Maps are available online and as downloadable images, for selected diseases such as COVID-19, chikungunya, cholera, dengue, measles and rubella, and vectors. (e.g. mosquitoes, biting midges, ticks, sandflies).

ECDC also provides additional data for research purposes upon written request (https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy)

Request for Tessy data for research purpuse

In accordance with the "Policy on data submission, access, and use of data within TESSy"[1] and in order to process a request for an extraction of case-based/aggregated data from TESSy for research purposes/tasks in the public interest, information regarding applicant details, researcher CV, purposes of the use of Tessy data, description of the requested data . An ECDC data disclaimer provide further details on data use and dissemination including legal implication and responsibilities.

2.1.3 ECDC-Vaccine Tracker

The data presented in the Vaccine Tracker are submitted by European Union/European Economic Area (EU/EEA) countries to ECDC through The European Surveillance System (TESSy) once a week on Tuesdays. EU/EEA countries report aggregated data on the number of vaccine doses distributed by manufacturers to the country, the number of first, second, additional and unspecified doses administered to adults (18+), adolescents and children (<18) overall, by age groups and in specific target groups, such as healthcare workers (HCWs) and residents in long-term care facilities (LTCFs). Doses are also reported by the different vaccines. The downloadable data files contain the data on the COVID-19 vaccine rollout mentioned above and each row contains the corresponding data for a certain week and country. The file is updated every Thursday. Data are subject to retrospective corrections; corrected datasets are released as soon as updated national data is processed.

Going more in-depth, the downloadable data files also contain information about hospitalization and Intensive Care Unit (ICU) admission rates and current occupancy for COVID-19 by date and country. Each row contains the corresponding data for a certain date (day or week) and per country. The file is updated weekly. Data can be used in line with ECDC's copyright policy. The figures displayed about hospitalization, ICU admission rates, and current occupancy are based on several data sources. The main source is case-based data submitted by Member States to TESSy. However, ECDC compiles data from

[1] The "Policy on data submission, access, and use of data within TESSy" is available from the ECDC's website

public online sources when unavailable, especially for current occupancy. The data displayed have been automatically or manually retrieved ('web-scraped') daily from national/official public online sources from EU/EEA countries. Scraped data are unavailable for all variables and/or countries due to content variability on national websites.

2.2 European Food Safety Authority (EFSA)

2.2.1 General description of EFSA

EFSA was established in 2002 (Regulation (EC) No 178/2002) to reinforce the system of scientific and technical support to the EU and the Member states in the field of food safety. EFSA serves as an impartial source of scientific advice to risk managers and to communicate risks associated with the food chain. It provides the scientific basis for laws and regulations to protect European consumers from food-related risks – from farm to fork. The core of EFSA activities is to collect, appraise and integrate scientific evidence to provide scientific advice to risk managers, jointly produced by independent experts and EFSA staff. EFSA also communicate about risks in the food chain. The EFSA activities are carried out with Member States partners, through the involvement of individual experts and competent organisations.

2.2.2 European One Health zoonoses report and The European Union Summary Report on Antimicrobial Resistance in Zoonotic and indicator bacteria from humans, animals, and food

The production of the European Union One Health summary Report on zoonoses and food-borne outbreaks , as well as the European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food, is underpinned by Directive 2003/99/EC laying down the EU system for monitoring and reporting of information on zoonoses, which obligates the MSs to collect data on zoonoses, zoonotic agents, AMR and food-borne outbreaks. EFSA is assigned the tasks of examining the data collected and preparing the European Union Summary Report in collaboration with the European Centre for Disease Prevention and Control (ECDC). For the reporting of the annual data, EFSA provides the Data Collection Framework (DCF) that allows data providers to transmit data under extensible markup language (XML) format through a web service. This DCF specifically aims at guiding the reporting of information/data under the framework of Directive 2003/99/EC, Regulation (EU) 1375/2015, Regulation (EU) 2017/625, Commission Implementing Regulation (EU) 2019/627 and Commission Implementing Decision 2013/652/EC. The data collected by EFSA are freely available in the Knowledge Junction, an open repository curated by EFSA for the exchange of evidence and supporting materials used in food and feed safety risk assessments (https://zenodo.org/communities/efsa-kj/about/).

2.3 World Organization for Animal Health (WOAH)

2.3.1 General description of WOAH

WOAH is the global authority on animal health founded in 1924 as the Office International des Epizooties (OIE), in May 2003 was adopted the common name World Organization for Animal Health. It's an intergovernmental organization, which focus on transparently disseminating information on animal diseases. Headquartered in Paris, the organization maintains permanent relations with over 70 international and regional organizations and has Regional and Sub-regional Offices worldwide.

Since 1990 has adopted a strategic planning cycle for its five-year work program. WOAH's Seventh Strategic Plan, adopted by the Member Countries who met during the 88th General Session of the World Assembly of Delegates, covers the period from 2021 to 2025. The plan leverages the organization's experience and expertise, with the support of its network of Reference Centers, to foster the necessary changes and provide leadership in global animal health governance so that Veterinary Services are better equipped to anticipate and respond to new expectations.

International Standards are set to support the safe trade of animals and animal products and improve the prevention and control of animal diseases. Additionally, are collated data to improve knowledge of animal health situations worldwide.

2.3.2 World Animal Health Information System (WAHIS)

WAHIS is the global animal health reference database of the World Organization for Animal Health (WOAH). WAHIS data reflect the validated information since 2005 reported by the Veterinary Services from Member and non-Member Countries and Territories on terrestrial and aquatic Listed diseases in domestic animals and wildlife, as well as on emerging diseases and zoonoses.

WAHIS includes interactive mapping tools and dashboards to support data consultation, visualization and extraction of officially validated animal health data.

Analytics dashboards enable users to consult, visualize and extract officially validated animal health information. The dashboard "Reports" is based on real-time information on exceptional animal disease events for listed and emerging diseases collected via the early warning system and the monthly reporting of listed diseases to the monitoring system.

The content of these dashboards is based on the data contained in the official reports (immediate notifications and follow-up reports, six-monthly reports and annual reports) submitted by the relevant Veterinary Services through WOAH-WAHIS. For visualization purposes, provided data has been aggregated in a comprehensive way. If you want to consult the detailed information, please go to the specific "Reports" section. This dashboard is refreshed every 1-2 hours.

2.4 Food and Agriculture Organization (FAO)

2.4.1 General description of FAO

FAO is a specialized agency of the United Nations that leads international efforts to defeat hunger. In Quebec City, Canada, the first session of the newly created United Nations establishes the FAO as a specialized UN agency. Washington D.C. is designated as a temporary FAO headquarters.

The mission is to achieve food security for all and make sure that people have regular access to enough high-quality food to lead active, healthy lives. With 195 members - 194 countries and the European Union, FAO works in over 130 countries worldwide.

FAO promotes the exchange of scientific research and technical knowledge related to all aspects of food and agriculture. Through a series of knowledge programs, FAO helps to increase the accessibility and visibility of research products in its Member Countries and to make this information available, accessible and usable worldwide.

2.4.2 EMPRES Global Animal Disease Information System (Empres-i)

In 2004 FAO’s Emergency Prevention System (EMPRES) designed and developed a web-based secure information system to support country-level veterinary services by facilitating regional and global disease information: EMPRES Global Animal Disease Information System (EMPRES-i). EMPRES-i is a global reference database for animal diseases including zoonosis (FAO to prevent threats to the food chain).

Empres-I facilitates the organization's access to national, regional, and global disease data and information. EMPRES-i also integrates data from other databases, i.e. livestock density or environmental layers from FAO databases, e.g, the Global Livestock Production and Health Atlas, GLiPHA (user-friendly, highly interactive electronic atlas using the Key Indicator Data System (KIDS) and from other systems. EMPRES-i provides up-to-date information on global animal disease distribution and current threats at national, regional and global level. 'Disease Events' can be presented on a map and further analyzed by choosing from a variety of optional layers.

2.5 EU – Animal Disease Information System (ADIS)

2.5.1 General description of ADIS

The EU Animal Diseases Information System (ADIS) is designed to register and document the evolution of important infectious animal diseases as provided by the Animal Health Law. ADIS provides uniform conditions for implementing Union notification and reporting as provided by Commission Implementing Regulation (EU) 2020/2002.

It is a disease management tool that ensures immediate notification of alert messages and detailed information about outbreaks of the most relevant animal diseases in the countries connected to the application. This permits immediate access to information about contagious animal disease outbreaks and ensures implementation of early warning, which enables a prompt response for controlling the epidemiological situation. This directly impacts the trade of live animals and their products both for the internal market and for international trade with third countries.

While ADIS is a system not directly related to food safety, it impacts public health in relation to all zoonotic diseases within its scope.

ADIS has been developed in close collaboration with the World Organization for Animal Health (WOAH) to facilitate data exchange between ADIS and WAHIS (World Animal Health Information System). This feature is partially implemented; a further release of ADIS will allow for a two-way exchange of information with the submission of outbreaks to WOAH -WAHIS.

The operational objective of the system is to ensure the rapid exchange of Union Notifications containing information between the competent authorities responsible for animal health in each EU country and the Commission on outbreaks of selected contagious animal diseases.

The system allows the coordination and monitoring of outbreaks of contagious animal diseases and enables EU countries and Commission services to take immediate measures to prevent the spread of the diseases in question. ADIS will also implement the management of Union reporting data (still under development).

The EU countries and the other countries connected to the application are responsible for supplying ADIS with the necessary information. A weekly (every Friday) e-mail message is sent to all the ADIS members summarizing all outbreaks entered into the system. The designated competent authorities in an EU country enter information on outbreaks into the ADIS. This information is automatically sent to all ADIS users and the Commission.

The Commission correlates data and sends back the information to the veterinary offices of the EU countries every week.

2.6 European Medicines Agency (EMA)

2.6.1 General description of EMA

The centralized marketing authorization procedure for human and veterinary medicines is based on two pieces of EU legislation which lay down the rules for the authorization of medicines and their placing on the EU market. These are Regulation (EC) No 726/2004 (as amended), which enabled the establishment of EMA, and Regulation (EU) No 2019/6.

Founded in 1995, the European Medicines Agency (EMA) works in the European Union (EU) and globally to protect public and animal health by assessing medicines to rigorous scientific standards and providing partners and stakeholders with independent, science-based information. The mission of EMA is to foster scientific excellence in evaluating and supervising medicines for the benefit of public and animal health in the European Union (EU). To fulfill its mission, EMA works closely with national competent authorities in a regulatory network. The Agency also implements policies and procedures to ensure it works independently, openly, and transparently and upholds the highest standards in its scientific recommendations.

This privacy statement describes how the European Medicines Agency (EMA) collects and uses personal information about you in accordance with Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data.

2.6.2 The European Surveillance of Veterinary Antimicrobial Consumption (ESVAC) project

The European Medicines Agency (EMA) started this project in September 2009, following a request from the European Commission to develop a harmonized approach for collecting and reporting data on the use of antimicrobial agents in animals from EU and European Economic Area (EEA) Member States The ESVAC project collects information on how antimicrobial medicines are used in animals across the European Union (EU). Voluntary participation in the ESVAC project has increased from 9 to 31 reporting countries since 2010. The data collected are made available through the interactive ESVAC database (https://esvacbi.ema.europa.eu/analytics/saw.dll?PortalPages) that allows users to access a summary of the specific ESVAC data they are interested in, including data for a specific country or sales of a particular antimicrobial class.

2.6.3 European database of suspected adverse drug reaction reports (EUDRA Vigilance)

According to Article 24(1) of Regulation (EC) No 726/2004, the EudraVigilance database shall contain information on suspected adverse reactions in human beings arising from the use of the medicinal product within the terms of the marketing authorization as well as from uses outside the terms of the marketing authorization and on those occurring in the course of post-authorization studies with the medicinal product or associated with occupational exposure. Stakeholders being granted access to EudraVigilance data according to European Medicines Agency's (EMA) EudraVigilance access policy can be grouped in Medicines regulatory authorities in EEA Member States; Healthcare Professionals and the Public; Marketing Authorization Holders; Academia; WHO – Uppsala Monitoring Centre; Medicines regulatory authorities in third countries. The policy provides as much information as possible while meeting data protection obligations. The 2010 pharmacovigilance legislation i.e. Regulation (EC) No 726/2004, Directive 2001/83/EC, and the Commission Implementing Regulation (EU) No 520/2012 introduced significant changes in the way adverse reactions are to be reported to and accessed in EudraVigilance.

There are two modules for the collection of data within the EudraVigilance database:

-The EudraVigilance Clinical Trial Module (EVCTM), designed to receive reports by sponsors on suspected unexpected serious adverse reactions (SUSARs) that occur in the frame of interventional studies;

-The EudraVigilance Post-Authorization Module (EVPM), designed to receive spontaneous reports from: Healthcare professionals and patients; post-authorization studies (non-interventional); worldwide scientific literature; individual use (compassionate use, Law 648/96, named patient use).

All reports submitted to EudraVigilance are also transferred to the World Health Organisation database (WHO) VigiBase and are therefore available to all international competent authorities in the field of pharmacovigilance in a short time period after submission.

2.7 The Program for Monitoring Emerging Diseases (ProMED)

2.7.1 General description of PROMED and data collected

The Program for Monitoring Emerging Diseases (ProMED) is a program of the International Society of Infectious Diseases (ISID). ProMED was launched in 1994 as an Internet service to identify unusual health events related to emerging and re-emerging infectious diseases and toxins affecting humans, animals and plants. ProMED is the largest publicly-available system conducting global reporting of infectious disease outbreaks. It is an essential source of information used daily by international public health leaders, government officials, physicians, veterinarians, researchers, private companies, journalists and the general public, providing timely reporting of important emerging pathogens and their vectors using a One Health approach. Reports are produced and commentary provided by a multidisciplinary global team of subject matter expert (SME) Moderators in a variety of fields including virology, parasitology, epidemiology, entomology, veterinary and plant diseases. ProMED reports data on outbreaks globally, 24 hours a day, 7 days a week, averaging eight outbreak reports per day.

ProMED is based on innovative and informal disease surveillance, which allows it to disseminate information faster than traditional surveillance systems. By relying on local media, professional networks and on-the-ground experts. These reports are simultaneously sent out to subscribers via email and posted to ProMED’s website, which receives thousands of visits each day. Reports are also disseminated by social media, RSS feeds and smartphone apps.

ProMED is based on innovative and informal disease surveillance, which allows it to disseminate information faster than traditional surveillance systems. By relying on local media, professional networks and on-the-ground experts. These reports are simultaneously sent out to subscribers

via email and posted to ProMED’s website, which receives thousands of visits each day. Reports are also disseminated by social media, RSS feeds and smartphone apps.

ProMED comprises one global network and eight regional networks operating in multiple languages including French, Spanish, Portuguese, Russian and Arabic. Subscribers are able to choose which lists they wish to receive, whether in real time or as a digest, and additionally posts focused only on plants, zoonotic and animal diseases or posts that relate only to disease occurrence (without commentary and discussion). The specialist moderators (experts in specific subject matter or in regional diseases) are located in 37 countries and are constantly scanning for, reviewing and posting information to the network. In order to maintain a level of consistency, as well as highlight the most important events of the day, ProMED staff carefully consider each post and aim for quality of reporting and accuracy over quantity of posts. ProMED information flowchart it’s composed by 3 main steps: receipt of information; internal review andverification and dissemination (Carrion and Madoff, 2017).

2.8 National Databases

The WP2 collected information within MOOD's partners Countries on the availability of national data-set for West Nile, Influenza A, and Tickborne Encephalitis. The information collection aimed to detect national data, identify ownership, and explore the shareability of such data. The survey also aimed to identify national-level data from different domains, particularly animal disease surveillance, environment and vector surveillance, and human surveillance.
A summary of the result of the survey is provided in Table 1:

Table 1: Summary of the survey among MOOD partner Countries on the availability of national data for West Nile, Influenza and Tickborne Encephalitis.

Country	Disease	Animal data	Entomological data	Human data
Belgium	Influenza			YES
	Influenza A in poultry	YES		
	Tickborne Encephalitis	YES		YES
	West Nile			YES
Finland	Influenza	YES		YES
	Influenza A in other species	YES		
	influenza A in pigs	YES		
	Influenza A in poultry	YES		
	Tickborne Encephalitis			YES
	West Nile	YES		YES
France	Tickborne Encephalitis	YES	YES	YES
Italy	Influenza			YES
	Influenza A in poultry	YES		
	West Nile	YES	YES	YES
Serbia	Tickborne Encephalitis			YES
	West Nile	YES		YES
Spain	Influenza			YES
	Tickborne Encephalitis			YES
	West Nile	YES		YES

The survey's information and results are available in the MOOD platform and data repository (AL FRESCO).

2.9 Data from the literature

To develop the MOOD tools and services, the project uses a number of emerging pathogens and related diseases which can serve as models for related infections, and important information for each model disease is based on a literature review and expert input. The goal is to collect and continuously update the information on each model disease throughout the project duration. The result of the literature review, for some of the diseases, have been discussed and integrated with the experts.

A multi-step approach has been followed in addressing the search:

1. description of the required contents for each pathogen and disease profile;
2. use of appropriate bibliographic databases;
3. definition of a common standard protocol, customised for each of the different pathogens and related diseases.
4. extrapolation of relevant thematic and quantitative information including data needed to parametrise the models developed in WP4;
5. identification of co-variates which are drivers for each model pathogen and disease (to be listed in Milestone 16 as the results of the profiling become available)
6. identification of keywords needed to collect unofficial epidemiological data through web-scraping and text mining (Subtask 2.2.3).

The approach implemented was the scoping review, as described by Tricco et al. (Tricco et al., 2018).

As for the other activities that involve data collection, these and the data retrieved from the published sources has been managed according to the general MOOD policy for data management (see D 7.4 Data management plan), and made openly available within the consortium, to allow further exploitation of datasets with reuse value, to feed scientific models and knowledge, and to bring scientific evidence to support decision and policy making.

According to the definition of scoping review, the method and the purpose of this activity in the framework of the WP2, that is more to “map the evidence” than to synthetize it, we conducted a number of searches and scoping reviews to cover all the pathogens identified as “models” in the MOOD project.

2.9.2 Details of the review approach developed in MOOD

To obtain the

- list of experts relevant to each model disease,
- the identification of critical unknowns and controversial scientific issues;
- the relevant thematic and quantitative information, including the co-variates for each model pathogen and disease
- keywords needed to collect un-official epidemiological data through web-scraping and text mining
- the Task 2.1 objective (profiling of prototype pathogens and diseases) states that each “profile” will summarize the available knowledge about several specific topics for each disease or pathogen:
- the biological, ecological, and molecular features of the causative agent;

- the natural history of disease in humans and vectors, including symptoms, morbidity and mortality;
- availability of preventive, therapeutic, and control measures, including licensed or pipelined vaccines;
- the epidemiological situation, past and current trends at different spatial scales;
- the sociological and demographical dimensions affecting susceptibility and exposure, including gender;
- diagnostic procedures and notification systems used at local, national and European scales;
- the infrastructure capacity to identify pathogens for each member state;
- the estimated influence of environmental change on the disease future trends.

To collect the information from the available literature, an initial general searching strategy was developed, based on list of key questions to be answered. This approach has been standardized in order to ensure quality and repeatability of the process following the reference provided by Tricco A. *et al*. on how to extend PRISMA approach to scoping reviews (Tricco et al., 2018).

The Key questions were standardized to ensure a common approach for each review, one for each pathogen identified in the project which include:

- influenza A (all virus types) for airborne pathogens;
- tick-borne encephalitis and Lyme borreliosis as models of endemic pathogens transmitted by endemic vectors;
- West Nile and Usutu viruses as examples of exotic pathogens transmitted by endemic vectors;
- Chikungunya, Dengue and Zika viruses as models of exotic pathogens transmitted by invasive mosquito species;
- Tularemia and Leptospirosis as models of neglected endemic pathogens with multiple transmission routes and reservoirs;
- unknown pathogens (disease X) as a challenge for any epidemic intelligence system.

The basic key questions used for a first exploration of the databases were defined as follow:

Q1: What is the infectivity, pathogenicity, growth kinetics (?); transmission route; host spectrum; environmental survival, persistence of [the given pathogen/disease]?

Q2: What are the effects of the water temperature, air temperature, soil temperature, humidity, precipitation, wind speed, wind direction on [the given pathogen/disease]?

Q3: What are the infection features, including: contact transmission rate, infection recovery rate, competence, morbidity, mortality rate, lethality of [the given pathogen/disease]?

Q4: What are the clinical features, trends, prevalence, incidence, seasonality, diagnostics, surveillance, immunity rate of [the given pathogen/disease]?

Q4a: What are the risk factors for infection, severe infection, sequelae and death of [the given pathogen/disease]?

Q5: What are the social features including: education, behavioral risks, ethnicity (race), gender, profession, population, religion, age, income, community correlates, risk perception of [the given pathogen/disease]?

Q6: What are the preventive measures, therapy, control measures, detection, diagnosis, vaccination efficacy of [the given pathogen/disease]?

2.9.3 Literature search strategy:

According to the questions defined above, a list of keywords (both general and specific for each pathogen/disease) was developed. The Keywords used in the literature search were then checked and extracted from the MeSH database and Embase vocabulary, then integrated with text words founds in relevant papers.

The draft search strategy (in particular the keywords regarding each disease) was submitted to the WP2 experts for review and revision.

The list of databases included in the literature search included: MEDLINE, EMBASE, SCISEARCH, BIOSIS, HCAPLUS, SCOPUS.

Inclusion and exclusion criteria are defined follows:

Study design:

included Primary research (i.e. studies generating new data) or data collections or systematic reviews (surveys, case-control studies, cohort studies, descriptive studies, systematic review and experimental in vitro and in vivo studies);

excluded case reports (if considered relevant by the experts, case report can be included for specific pathogens/diseases), Modelling estimates only, No denominator or No identified reference population

Pathogens/diseases:

included those listed in MOOD;

excluded those not included in MOOD

- Language of the full text:
  a. included full-text document in English OR other EU languages;
  b. excluded Chinese, Japanese, Russian and Arabic
- Time frame: from 2010 to date (from 1990 to date for environmental and climatic variables)
- Publication type:
- included primary research (i.e. studies generating new data) or data collection or systematic reviews;
- excluded patents, editorials and letters
- NO geographical restrictions
- Additional exclusions: the unavailability of full reference text, duplication of data, Low quality or quality not assessable of the study.

Once the literature search has been completed and all the duplicate records removed, the process for selecting the studies according to the defined criteria for inclusion and exclusion is performed independently in an un blinded standardized manner by at least 2 reviewers; disagreements between reviewers are resolved by consensus.

To ensure interactivity among the experts during the selection of the studies and the overall traceability of the process, this part of the activity has been done with the aid of the RAYYAN software (https://rayyan.ai/ - Mourad Ouzzani et al. Systematic Reviews (2016) DOI: 10.1186/s13643-016-0384-4.).

A data extraction sheet based on the Cochrane Consumers and Communication Review Group's data extraction template was developed to allow the extraction of relevant data from the selected studies and validated. The referent people for each disease made the selection of the relevant articles and the data extraction. The resulting tables, with the list of articles and the important disease covariates classified in:

1. Human covariates: demographic, social and economic covariates including gender if appropriate;
2. Animal covariates: list of important animal species and their roles in the dynamic of the diseases;
3. Vector covariates: list of important vectors (vertebrates and invertebrates);
4. Environmental covariates: list of important environmental covariates including temperature, seasonality, humidity and land cover.

For all the covariates, the unit of measurement and the quantitative correlation with the disease was provided in the table (table 1 provides an example of data extracted).

This information was shared among the project partners.

In conclusion, the harmonized search strategy implemented in MOOD is ensuring an extended coverage of the topics needed for the MOOD project, including general information and specific data. The already concluded and the advanced searches are good examples of the kind of information that can be provided to the other WPs, like updated distribution maps, a list of main animal hosts and reservoirs with frequencies of detection, that can also be used as a list of susceptible species, list of vectors and covariates, both environmental and social. All these data are linked to geographical information to locate them. The literature search will be completed and updated along the project.

A literature search for antimicrobial resistance data in animals was conducted following a slightly different approach. The sources of information identified were:

1) government reports on veterinary antimicrobials (predominantly sales) at country-level, collected through established surveillance systems,

2) scientific articles reporting estimates of veterinary AMU at country-level (predominantly imports),

3) scientific articles reporting on-farm AMU from surveys within countries.

Articles containing data on antimicrobial sales, consumption, usage, or imports were considered a proxy for usage. Four groups of animals were included: cattle, sheep, chicken, and pigs, which total 91.1% of animal biomass raised for food globally.

The literature search was conducted in PubMed.

Another source of information was the ESVAC report.

2.10 Data from other sources

Platform for Automated Extraction of Animal Disease Information from the Web (PADI-web).

The epidemic intelligence (EI) concept corresponds to a formalized surveillance process that encompasses all activities related to the early identification of potential health hazards that may represent a risk to health, and their verification, assessment and investigation. It relies on two main channels of information: indicator-based surveillance (IBS) and event-based surveillance (EBS). Indicator-based surveillance is defined as ‘the systematic collection, monitoring, analysis and interpretation of structured data (i.e. indicators)' It corresponds to conventional surveillance of formal sources and is based on established case definitions. Event-based surveillance is defined by the WHO as ‘the organized collection, monitoring, assessment and interpretation of mainly unstructured ad hoc information regarding health events or risks, which may represent an acute risk to human [or animal] health’. The definitions and concepts from both ECDC and WHO were elaborated for public health. However, they have been successfully transferred to other domains, such as plant health and both terrestrial and aquatic animal health. Both EBS and IBS can be formally represented as consecutive steps, corresponding to the flow of epidemiological information from its detection to its communication to adapted to relevant authorities (e.g., public health national networks, ministries of health, international organizations) or a larger network (e.g., end-users of EBS systems).

PADI-web (Platform for Automated extraction of animal Disease Information from the web) is a biosurveillance system dedicated to monitoring online news sources for the detection of emerging animal infectious diseases. PADI-web has collected more than 380,000 news articles since 2016. Compared to other existing biosurveillance tools, PADI-web focuses specifically on animal health and has a fully automated pipeline based on machine learning methods. The new functionalities of PADI-web is based on the integration of: 1) a new fine-grained classification system, 2) automatic methods to extract terms and named entities with text mining approaches, 3) semantic resources for indexing keywords and 4) a notification system for end users. Compared to other biosurveillance tools, PADI-web, which is integrated in the French Platform for Animal Health Surveillance (ESA Platform), offers strong coverage of the animal sector, a multilingual approach, an automated information extraction module and a notification tool configurable according to end-user needs.

Sources of genomic data .

Following databases from which genomic data can be retrieved:

Genbank/EMBL

The GenBank sequence database (https://www.ncbi.nlm.nih.gov/genbank/) and the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) are members of the tri-partide International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. These sequence databases maintain open access, annotated collections of all publicly available nucleotide sequences and their protein translations. There are several ways to search and retrieve data from GenBank, including searching GenBank for sequence identifiers and annotations with Entrez Nucleotide, searching and aligning GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently and search, linking, and downloading sequences programatically using NCBI e-utilities.

GISAID

the GISAID EpiFlu database provides a collection of influenza sequences containing associated metadata, both clinical and epidemiological. GISAID EpiFlu database is a resource that stores information about Influenza virus. It includes data-sharing platform through which sequence data are recommended for inclusion in seasonal and pre-pandemic vaccines. These data are available for research scientists, public and animal health officials and the pharmaceutical industry. GISAID is also the primary genome resource for the novel coronavirus responsible for COVID-19.

The Influenza Research Database (IRD) and ViPR: IRD and ViPR are freely available, NIAID-funded resources that support the research of viral pathogens in the NIAID Category A-C Priority Pathogen lists and those causing (re)emerging infectious diseases. IRD and ViPR integrate data from external sources (GenBank, UniProt, Immune Epitope Database, Protein Data Bank, etc.), direct submissions, and internal curation and analysis pipelines, and provide a suite of bioinformatics analysis and visualization tools to expedite virology research.

OpenFlu database:

The OpenFlu database (OpenFluDB) is part of a collaborative effort to share observations on the evolution of Influenza virus in both animals and humans. It contains genomic and protein sequences as well as epidemiological data from more than 25,000 isolates. The isolate annotations include virus type, subtype and lineage, host, geographical location, experimentally tested antiviral resistance. Protein sequences are automatically derived from nucleotide sequences. From these, putative enhanced pathogenicity and human adaptation propensity are computed. Each virus isolate can be associated with the laboratories that collected, sequenced and submitted it.

Table 2: Summary of the important data sources

Data Source	**Availability**	**Type of data**	**Update frequency**	**Notes**
ECDC - TESSy	Available upon request - restricted	Human cases single records	yearly	
ECDC–diseases data	Dashboard publicly available	Human cases aggregated	yearly	
ECDC-vectors data	Publicly available	Presence maps	yearly	
ECDC-other diseases	Publicly available as report	Human cases aggregated	Not regularly	
EFSA zoonoses data	Available upon registration	Human, animal food data	yearly	Only data on foodborne pathogens and few others
EFSA Antimicrobial resistance data	Available upon registration	Human, animal food data single cases outbreaks	yearly	
WOAH-WAHIS	Publicly Available	Animal single cases, outbreaks	Weekly	Connecting to ADIS
FAO-EMPRES	Publicly Available	Animal single cases, outbreaks	Weekly	
EU-ADIS	Pubicly available	Animal outbreaks	Weekly	Aggregated data and maps - connecting to WAHIS
PROMED	Publicly available	Human, animal, plant diseases/outbreaks	Daily	Alert available upon registration

3.4 Data extracted from Public scientific papers

Data extraction sheet is based on the Cochrane Consumers and Communication Review Group's. The data extraction template was developed to allow the extraction of relevant data from the selected studies. It was pilot-tested on 30 randomly-selected included studies and refined accordingly. The data extraction template is organized in sheets: Ref-Sheet which contains general information of selected articles and information about the study and Cov-Sheets instead contain specific information on extracted covariates (Human, Environmental, Animal and Vector). (See Annex 1).4. Data acquisition and analyses

4.1 Open access and restricted data

Most of the data acquired are from public sources so can be redistributed without restriction. Data sources must always be cited where appropriate to ensure due acknowledgement and academic courtesy.

Covariates and Diseases data

Access and redistribution rights for all covariate datasets held by project partners is assessed when they are first acquired – which may have been many years before the project. The great majority are public domain – and may therefore be used without restriction for research purposes. A few, which are clearly flagged, require the data owners' permission for redistribution. Access and redistribution permissions for newly acquired data will be determined at the time of acquisition. New data sets produced by processing of combining datasets already held by partners will inherit the contributory dataset permissions. Conventional disease data are provided by MSs as a mandatory surveillance activity. The basic information to be collected for each notified case, and the data flow from the local level to the EU level, is established by law. The official reports are publicly available without restriction of use. Agreement with ECDC or other national institutions for the use of data collected during the surveillance activities, but not published in reports (details about age, gender, location), can be negotiated with the ECDC. Data from text mining are also derived from public sources and so are not subject to restriction. An exception to this is Twitter data, which needs to be used in compliance with Twitter's developer agreement and policies. Pathogen genome data from publicly accessible databases are generally public domain. For specific databases such as GISAID, specific terms of use apply (https://www.gisaid.org/registration/terms-of-use/), which for example involve acknowledgement and/or co-authorship in publications and best efforts to collaborate with representatives of the originating laboratory responsible for obtaining the specimen(s). AMR point prevalence survey data are taken from public domain sources which are already open source and so redistribution is also not restricted.

4.2 Procedures for data acquisition

Covariates and Diseases data

The standard ways of sourcing data were used to find new datasets - web searches, enquiries to professional networks, and trawling online archives, all of which are routine MOOD activities. Data pointers and author links from simultaneous literature sources will be used.

For some covariates proxy were used.

The annual number of cases, annual incidence and spatial distribution of cases (at NUTS1, but more details are available from the national reports) are the types of data most readily accessible. Surveillance data collected at national level and aggregated at EU level are published in annual reports and also publicly available at institutional websites (e.g. https://www.ecdc.europa.eu/en/surveillance-atlas-infectious-diseases). National surveillance reports published on institutional websites have also been explored to complement official EU surveillance data provided by ECDC, particularly for diseases not included in the EU surveillance but notified by some MSs only.

Public Health Institutions partners of the project have been requested to provide information on national more fine-grained data and details on how to access them. The national data have been used for TBE and West Nile publications. However, a regular supply of national data to the MOOD platform is very difficult if sustainability is an issue.

The need for specific agreements to be regularly renewed between MOOD (that should become a legal entity to sign these agreements) and the national data owner and the data security and protection issues make these data sources very difficult to include.

Generic and specific text mining approaches have been developed and used for monitoring specific diseases. To date, most of the testing has been done on Avian Influenza. The challenging issue is to implement algorithms adapted for two types of surveillance: 1) Disease specific: for monitoring known diseases. This surveillance is based on disease names (and variations) as keywords; and 2) Non-specific: syndromic monitoring for an unspecified disease X with specific keywords like symptoms.

Disease-specific and syndromic surveillance started for the COVID-19. Using different lexical variations (e.g. coronavirus, covid-19, covid19, ncov, 2019-ncov, n-cov2019, ncov2019, etc.) specific keywords and strategies have being used as a case study for both types of automatic surveillance systems.

Media data were collected through multiple sources 1) PADI-Web (targeted at news articles from Google News); 2) Scientific publications; 3) Social media data (Twitter).

The initial vocabulary used for data collection has been enriched by mining a scientific publication corpus.

Genomic data have been downloaded from publicly accessible databases. The following critical resources have been identified: Genbanl/EMBL; ISAID; The Influenza Research Database (IRD) and ViPR; The OpenFlu database.

Point prevalence surveys (PPS) reporting AMR rates in food producing animals were identified through a literature search in Pubmed, Scopus, and Web of Knowledge. Relevant PPS data as illustrated above were collected in an Excel spreadsheet.

A set of common principles have been established regarding the data characteristics such as resolution, coverage, or metadata requirements, to be considered for the different types of data. These principles are detailed in the following sections and summarised as follows:

- MOOD should aim for consistency in the data formats across all MOOD work-packages, including file formats and spatial coordinate reference systems;
- MOOD should promote the use of a consistent spatial extent for the different variables;
- MOOD should support data at a minimal set of resolutions, appropriate for different types of analysis;
- MOOD should support the handling of both spatial and spatio-temporal data, although this should be kept as simple as possible;
- MOOD should support the storage of data together with minimal descriptive metadata.

Achieving these general objectives involves processing the data from different sources to build standardized products. This processing involves using standard GIS operations (e.g., spatial data upscaling, downscaling, interpolation, density estimation, etc.), as well as the use of tailored procedures for geo-referencing unstructured and tabular data.

The project should rely on a common gazetteer, namely GeoNames1, to identify the geographic coordinates for place names, needed to assign them to administrative areas or raster image locations (e.g., when processing data originally available in tabular form).

Geographic data types and formats

The project provides both vector and raster data in the analytical tools under development.

In the Module called “Data & covariates access” it is possible the visualization and download of relevant standardized covariates relative to the MOOD model diseases and, more generally, to infectious disease emergence in support of risk assessment and modeling.

All data will be accompanied by description and metadata using standardized schemes based on ISO/TS 19115 and 19139. A more precise definition of the metadata elements that should be considered is yet to be defined, but should include the following minimal information for source/provenance attribution.

- Contact information, both for external users (i.e., a general contact address for the project) and internally within MOOD (i.e., contacts for the individual responsible for processing or generating the data).
- Technical specifications, including spatial resolution, collection date, coordinate reference system, etc.
- Description for the processes used in preparing the data, including references to accompanying documentation.
- Indexing key-words, ideally defined according to a normalized schema. This is particularly important, given that semantic interoperability is also an issue to consider. Data products regarding each of the two general categories (e.g., data on covariates, and data on diseases/vectors) should ideally be associated to standardized names.

Spatial resolution, extent, and standardized boundaries

Within the MOOD project, preference should be given to using raster representations with a reasonably high spatial resolution. In some cases, this can involve using spatial downscaling methods to convert the original data into the project suggested format and resolution.

Regarding raster data (stored in the GeoTIFF format), most variables should ideally be represented at two separate spatial resolutions, namely as a coarse-level grid of approximately 5km per cell, and as a thin-grained resolution of 1km per cell. We acknowledge that some variables may only be represented at one of these resolutions, with preference given to high-resolution data whenever possible.

The use of two distinct administrative level resolutions (i.e., aggregation levels) is recommended, corresponding to coarse-level data using country level divisions of the territory, and thin-grained data using sub-national administrative divisions. The administrative divisions used by the ECDC VectorNet project[2] have been identified as a potential project standard, as they have a number of practical advantages:

- The ECDC VectorNet divisions mostly correspond to either NUTS III regions for European countries, but this scheme also considers divisions for other relevant territories that do not use the NUTS scheme (e.g., based on the Global Administrative Unit Layers (GAUL));
- The ECDC VectorNet divisions aimed for a balanced size, e.g. mixing NUTS III and small divisions for different countries depending on the size of the regions.

As much as possible, the MOOD project recommends using a standard geographical spatial extent which includes Europe and neighbouring countries. A suitable candidate is, once again the ECDC VectorNet extent which runs from Iceland to Turkey and Finland to Morocco.

Specifications required for data requests.

The data specifications will be similar for both dependent and independent variables and comprise a range of descriptors, based on the principles set out above, for which required specifications should be defined for every category. The list below should be seen as a guideline for the descriptors needed to define data requested for modeling, although this still requires further details (e.g., the project should ideally decide on the specific data sources for each type of variable, it should decide on an appropriate update periodicity, etc.)

A. Dependent variable specifications

Disease or vector name, according to a standardized taxonomy.

Units and denominators: e.g. incidence, prevalence, cases, deaths, values per unit time or population number, presence/absence. This may include details of sample effort or method.

Genomic data: collection of pathogen sequences containing associated metadata, both clinical and epidemiological.

Data on AMR pathogens: collection of epidemiological data and associated covariates for the selected model pathogens.

B. Geographical specifications: all data supplied should be georeferenced, and the relevant parameters should be defined with basis on the aforementioned recommendations

Extent – the bounding coordinates or country/administrative unit lists defining the area of Interest, by default corresponding to the extent considered for the ECDC VectorNet project.

Spatial resolution: Covariate, disease or vector data will be supplied as raster imagery will be supplied at either 1km or 5km resolution, as appropriate. The default is 1km and, in some cases, one can consider separate files for each resolution.

Mapping units: by default, mapping units are based on the Nomenclature of Territorial Units NUTS (where available) or Global Administrative Unit Layers (GAUL), as made in the ECDC VectorNet project. Place names, in connection to these units, to support tabular file formats, should be consistent with one of the major gazetteers (e.g., GeoNames (https://www.geonames.org).

A. Temporal Specifications

Date: Start and end date of the observations contained in the dataset
Temporal resolution (Frequency): for example sample interval – weekly, annual, monthly.

B. Data file formats: different data types may have different formats

Geographic polygon file format: *GeoPackage*
Raster Image: *GeoTIFF*
Accompanying tabular data format: csv, Excel compatible

The geographical data, including administrative boundaries and borders represented and utilized in the MOOD tools and modules are in line with the European regulations and based on the information provided by the Eurostat the statistical authority of the EU. In case of the use of geographical data from other sources, information concerning the important features and comparability/harmonization issues will be provided. MOOD will maintain the datasets along the project life, and will establish conditions and rules for the use beyond the end of the project. MOOD accepts no responsibility or liability whatsoever with regard to the information on its geographical datasets.

4.3 important discussion points from the stakeholders and end users meeting, Helsinki 28-29 June 2023

Demo & interaction with potential end-users:
The tools that the MOOD consortium is developing have been presented to MOOD partners, stakeholders, and a selection of potential end users.
Three tools were presented:

1. AVIA Gis Covariates module: an online platform that displays environmental and vector covariates.
2. PADI-Web: a web scraping tool for epidemiology surveillance. It is an EBS system for animal diseases.
3. Epid Data Explorer tool: a visualization tool that enables simplified exploration of epidemic data and allows for easy comparisons of indicators across different geographical areas or periods.

They all work with specific data collected according to the abovementioned criteria and approaches.
From the interactive workshops with stakeholders and end users the important discussion points raised concerning the data and their use were:

1. Importance of the proposed tools for the professional activity

2. Need of additional information about the data uploaded in the tools (spatio-temporal span, harmonization level, integration, summary of participants (institution, gender, etc)
3. Need of additional data (environmental, vector/hosts, and social data. In particular mobility data)

The meeting included also the presentation to stakeholders and end users of three case studies; on Avian influenza, Dengue and West Nile/TBE.

From the interaction between MOOD partners and stakeholders in the case studies, availability of updated data on the diseases, the environmental and social covariates of importance and the vectors/vertebrate hosts was pointed out in order to have tools and information useful for risk assessment.

Specific comment on Mobility data:

Human mobility pertains to how people move across space and plays a crucial role in the spatiotemporal transmission dynamics of infectious diseases. During the COVID-19 pandemic many studies have applied mobility data to explore spatiotemporal trends over time, investigate associations with other variables, and predict or simulate the spread of COVID-19. The multi-source human mobility data contain rich, multi-faceted spatiotemporal information on human mobility patterns. Three primary sources of human mobility datasets can be considered: (1) public transit systems; (2) mobile network operators; and (3) mobile phone applications.

Among the overall mobility data, social activity data reflect human social activity behavior with access to different places of interest, such as workplaces, residential areas, public transit, health care facilities, schools, shopping centers, and recreational and sports facilities. Most of such social activity data are GPS-derived metrics of foot traffic or POI access frequency sourced from large information technology (IT) companies with mapping services (e.g. Google and Apple). One of the mobility datasets most commonly used is Google Mobility Reports that provides the percentage change of place visit frequencies in six types of locations (workplaces, residential, parks, grocery and pharmacy, retail and recreation, and transit stations) compared to a pre-pandemic baseline value of mobility from 3 January to 6 February 2020.

The Google's Community Mobility Reports go through a robust anonymization process that employs differential privacy techniques to ensure that personal data, including an individual's location, movement, or contacts, cannot be derived from the metrics, while providing researchers and public health authorities with valuable insights to help inform official decision making.

The privacy of human mobility measures varies across different types of datasets. They are more likely to be released as aggregated-level metrics based on a large volume of anonymous location data. Such aggregated data indicate the overall patterns or changes of human mobility in a particular spatial unit with less concern about identifiable user-based information. Governments, professional associations and organizations, data providers, and researchers have made joint efforts to improve the stringency and implementation of regulations in addition to ensuring the ethics clearance in the process of data sharing and manipulation. The quality of mobility data depends on data types, and this makes it challenging to assess data quality without careful comparison studies. Human mobility data used in COVID-19 research vary greatly in spatial and temporal data coverage. For social activity data, Apple and Google mobility data are available globally at the state or city level in some countries. Conventional data sources, such as public transit system, have started to provide rich population flows data before the COVID-19 pandemic.

To help researchers and governments worldwide with the response to COVID-19, technology companies and research institutions have made human mobility datasets publicly available after pandemic. There are several ways to publish or share these datasets: (1) online data dashboards via the official websites of data providers; (2) GitHub (https://github.com/); (3) Harvard Dataverse (https://dataverse.harvard.edu/); (4) user applications (Hu et al. 2021DOI: 10.1080/17538947.2021.1952324).

4.4 General considerations of data acquisition

In today's information age, the challenge is not the lack of data but rather how to identify the most relevant data for meaningful results and combine data from various sources that might not be standardized or interoperable to enable analysis. Epidemiologists need to determine whether existing data can be analyzed to inform the decision-makers, whether additional data must be collected, and how to do so most efficiently and expeditiously.

Multiple factors must be considered when identifying relevant data sources. These include objectives and scope of the final analysis, whether requisite data exist and can be accessed, to what extent data from different sources can be practically combined, methods for and feasibility of primary data collection, and resources (e.g., staff, funding) available.

Although engaging stakeholders and data owners such as public health agencies and international health institutions early in the data analysis process is time-consuming, including them is essential. Discussing the purpose of the analysis and the data collection processes will prove invaluable in the long run when collaborators are needed during data collection, analysis, and communication with stakeholders. After evaluating whether existing data can address the study objectives, the epidemiologist must determine whether additional data must be collected and, if so, what and how.

Changes in technology also challenge data collection. The expanded use of computer technology makes many new data sources possible. It is incumbent upon epidemiologists to adapt to these changes.

Novel data streams have the potential to reshape epidemic preparedness and response. Still, a pressing need remains to address data sharing, sustainability, and scalability issues that could hinder future epidemic preparedness.

The MOOD project focuses on data mining, aggregation, and advanced analysis to aid early detection and risk assessment of infectious diseases. Data is one of MOOD's most valuable resources, and the timely and sustainable availability of the most appropriate and relevant data is critical to the entire MOOD framework.

In case of an epidemic, there is a need for fundamental insights into the epidemiological characteristics of the infection, especially if it is a new or emerging pathogen, from transmission potential to natural history. On one hand, this requires a rapid scale-up of testing and sequencing, a fast assessment of clinical impact, and open sharing of early findings. On the other hand, information for forecasting of disease dynamics, estimation of potential burden and evaluation of interventions.

The MOOD's innovative goal is to create a real-world data ecosystem that facilitates data exchange and utilization in a strongly “one health” oriented way. But how to integrate the data, whether the data is sufficiently real-time, the degree of structure and standardization of data, quality, and reliability of the data remain important challenges.

Effective epidemic response hinges on fast and reliable data sharing. An efficient and secure data sharing enables responses based on the best evidence. It can also support rapid follow-up analysis that builds on earlier work, reducing repetition. However, there can be challenges in ensuring rapid analysis and sharing while maintaining appropriate credit and quality control.

Using publicly available data from community-based platforms for data sharing and analysis or the repositories supported by international institutions provides one route to balancing these considerations. Such initiatives can also be integrated with novel data collections from environmental data repositories and population (human, animal) landscapes in the MOOD environment.

However, there remain obstacles to collating and ethically sharing such data. To balance the needs of data sharing and privacy protection, researchers may process the data in a way that is suitable for multiple levels of access, strict security policies for individual data collection and sharing and strong procedures for anonymization and data aggregation.

As well as being shareable, data analysis also needs to be sustainable. The volume and diversity of data generated during the COVID-19 pandemic have been unparalleled. Scientists have built new tools and upgraded existing software to handle large, previously imaginable data inputs. The urgency of the pandemic has driven a proliferation of creation and improvement, but generally, few of these tools and resources have long-term funding, making sustainability difficult.

However, the COVID-19 pandemic has put into sharp focus the need to take a longer view of resource development and maintenance.

Integrated programs linking different kinds of partners, including academia, government departments, research funders, health organizations, and private sector groups, could enable efficient coordination of analysis development and clear responsibility for maintenance and implementation.

Countries will also need to look at data capacity beyond their own borders.

Synthesis of multiple data streams in a shareable, sustainable way will also improve countries’ ability to respond to old and new pathogens and future epidemics.

The MOOD’s choice to focus on international sources of standardized and publicly available data, should help the sustainability of the project framework and tools. Moreover, it reduces the risk of ethical issues and limitations in the use and integration of the data.

References

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data

and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

2. Strategy and prioritization of the data collection. Milestone 10
3. Agreement with WP3 and WP4 on data requirement. Milestone 11
4. The identification of potential sources of genomic data. Milestone 12
5. Systematic review protocol and searches strategy. Milestone 14
6. Sarah Valentin, Elena Arsevska , Julien Rabatel , Sylvain Falala, Aliz´e Mercier, Renaud Lancelot, Mathieu Roche. PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance.
7. Kim S, Rhee C, Kang SJ, Tak S (2021) A scoping review on data integration in the field of infectious diseases, 2009-2018, Int. J. One Health, 7(2): 151-157.
8. Jakob CEM, Stecher M, Fuhrmann S, Wingen-Heimann S, Heinen S, Anton G, Behnke M, Behrends U, Boeker M, Castell S, Demski H, Diefenbach M, Falgenhauer JC, Fritzenwanker M, Gastmeier P, Gerhard M, Glöckner S, Golubovic M, Gunsenheimer Bartmeyer B, Ingenerf J, Kaiser R, Körner ML, Loag W, Mchardy A, Molitor E, Nübel U, Pritsch M, Ramharter M, Rieg SR, Rupp J, Schindler D, Schwudke D, Spinner C, Stottmeier B, Vehreschild M, Willmann M, Vehreschild JJ. Needs for an Integration of Specific Data Sources and Items - First Insights of a National Survey Within the German Center for Infection Research. Stud Health Technol Inform. 2021 May 24;278:237-244. doi: 10.3233/SHTI210075. PMID: 34042900.
9. Regulation (EC) No 851/2004 of the European Parliament and of the Council of 21 April 2004 establishing a European centre for disease prevention and control
10. Commission Decision of 22 December 1999 on the communicable diseases to be progressively covered by the Community network under Decision No 1082/2013/EU of the European Parliament and of the Council of 22 October 2013 on serious cross-border threats to health
11. Commission Implementing Decision (EU) 2018/945 of 22 June 2018 on the communicable diseases and related special health issues to be covered by epidemiological surveillance as well as relevant case definitions (Text with EEA relevance) - Link to the European Commission page with all translations
12. Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety
13. Regulation (Eu) 2019/1381 of the European Parliament and of the Council of 20 June 2019 on the transparency and sustainability of the EU risk assessment in the food chain and amending Regulations (EC) No 178/2002, (EC) No 1829/2003, (EC) No 1831/2003, (EC) No 2065/2003, (EC) No 1935/2004, (EC) No 1331/2008, (EC) No 1107/2009, (EU) 2015/2283 and Directive 2001/18/EC
14. Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan — a web and mobile app for systematic reviews. Systematic Reviews (2016) 5:210, DOI: 10.1186/s13643-016-0384-4).
15. The European Surveillance System (TESSy) (europa.eu)

16. https://zenodo.org/communities/efsa-kj/?page=1&size=20
17. Home - WOAH - World Organisation for Animal Health https://wahis.oie.int/#/home
18. About FAO | Food and Agriculture Organization of the United Nations EMPRES: Global Animal Disease Information System | AIMS (fao.org)
19. Animal Disease Information System (ADIS) (europa.eu)
20. About us | European Medicines Agency (europa.eu)
21. Promed About ProMED - ProMED-mail (promedmail.org) Carrion M, Madoff LC. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. Int Health. 2017 May 1;9(3):177-183. doi: 10.1093/inthealth/ihx014. PMID: 28582558; PMCID: PMC588
22. Mulchandani R, Wang Y, Gilbert M, Van Boeckel TP. Global trends in antimicrobial use in food-producing animals: 2020 to 2030. PLOS Glob Public Health. 2023 Feb 1;3(2):e0001305. doi: 10.1371/journal.pgph.0001305. PMID: 36963007; PMCID: PMC10021213.
23. Carrion, M., Madoff, L.C., 2017. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. Int Health 9, 177–183. https://doi.org/10.1093/inthealth/ihx014
24. Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D.J., Horsley, T., Weeks, L., Hempel, S., Akl, E.A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M.G., Garritty, C., Lewin, S., Godfrey, C.M., Macdonald, M.T., Langlois, E.V., Soares-Weiser, K., Moriarty, J., Clifford, T., Tunçalp, Ö., Straus, S.E., 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 169, 467–473. https://doi.org/10.7326/M18-0850
25. Tao Hu, Siqin Wang, Bing She, Mengxi Zhang, Xiao Huang, Yunhe Cui, Jacob Khuri, Yaxin Hu, Xiaokang Fu, Xiaoyue Wang, Peixiao Wang, Xinyan Zhu, Shuming Bao, Wendy Guan & Zhenlong Li (2021) Human mobility data in the COVID-19 pandemic: characteristics, applications, and challenges, International Journal of Digital Earth, 14:9, 1126-1147, DOI: 10.1080/17538947.2021.1952324