PANDEM-2
PANDEMIC PREPAREDNESS AND RESPONSE

List and description of selected data sources and analytical tools to monitor pandemics. FAIR-compliant metadata profiles. Data sources specification document and agreed database structure for WP3

Deliverable D2.2

08 July 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 883285

The material presented and views expressed here are the responsibility of the author(s) only. The EU Commission takes no responsibility for any use made of the information set out.

PANDEM-2

List and description of selected data sources and analytical tools to monitor pandemics. FAIR-compliant metadata profiles. Data sources specification document and agreed database structure for WP3.

Deliverable No: 2.2

Document date: 08 July 2022

Document version: 2.0

Full Name	Short Name	Beneficiary Number	Role
NATIONAL UNIVERSITY OF IRELAND GALWAY	NUIG	1	Coordinator
FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V	FRAUNHOFER	2	Beneficiary
UNIVERSITE CATHOLIQUE DE LOUVAIN	UCL	3	Beneficiary
PINTAIL LTD	PT	4	Beneficiary
FOLKHALSOMYNDIGHETEN	FOHM	5	Beneficiary
RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEU	RIVM	6	Beneficiary
CARR COMMUNICATIONS LIMITED	CARR	7	Beneficiary
TERVEYDEN JA HYVINVOINNIN LAITOS	THL	8	Beneficiary
INSTITUTUL NATIONAL DE SANATATE PUBLICA	NIPH	9	Beneficiary
ROBERT KOCH-INSTITUT	RKI	10	Beneficiary
STICHTING KATHOLIEKE UNIVERSITEIT	RUNMC	11	Beneficiary
CLARISOFT TECHNOLOGIES ROM SRL	CLARISOFT	12	Beneficiary
OSTERREICHISCHES ROTES KREUZ	ORK	13	Beneficiary
EPICONCEPT	EPIC	14	Beneficiary
INSTITUTO NACIONAL DE EMERGÊNCIA MÉDICA	INEM	15	Beneficiary
TRILATERAL RESEARCH LTD	TRI	16	Beneficiary
ISTITUTO PER L'INTERSCAMBIO SCIENTIFICO	ISI	17	Beneficiary
ASSOCIAZIONE DELLA CROCE ROSSA ITALIANA	ITRC	18	Beneficiary
INSTITUTO NACIONAL DE SAUDE DR. RICARDO JORGE	INSA	19	Beneficiary

Version History table

Version number (date)	Details
1.0 (29.10.2021)	Initial submission to EC
2.0 (08.07.2022)	Updated deliverable based on recommendations from review report for Reporting Period 1 -Clarified the data sources integration process in the Approach section and added Appendix 6.8 Final list of sources. In this last section the strategy implemented to prioritize the data to integrate is explained. -Better structured Sections "3. Approach" and "4. Results". Added a diagram at the beginning of the "Approach" section explicitly indicating tasks, results and dependencies between them. The titles of the sections of the document and appendix have been updated to reflect this diagram. -Added a new section in Section 3. Approach: ''End users data requests". - Added the conclusions of the data survey in the section "4. Results - Data families' priorities" and in Appendix 6.3.2. Also, the outputs of the survey are cited in Appendix 6.8 Final list of sources. - Updated Appendix 6.5 on FAIR principles. - Included real examples of data integration for ECDC reported cases, number of variants detected and Influenza.net participatory surveillance in Appendix 6.6 ''PANDEM-2 Data source Integration" going from the initial source files, the list of variables and the DLS definitions until the processed and standardized dataset. Also, in sub section "6.6.2 Database schema for integrating into PANDEM-2 database" the folder and file structure where JSON files are stored is displayed.

Table of Contents

1 Executive Summary

This D2.2 deliverable describes an initial and a refined list of data sources needed to meet pandemic managers data needs including those specific sources that are independent tasks in WP2 (see Appendix 6.4). Furthermore, we have detailed in a specific section (Appendix 6.7) open available data sources identified as useful for the PANDEM-2 IT solution. This list may still evolve during project development because of data availability or user needs.

PANDEM-2 needs to integrate data from different domains and sources into a coherent model (developed in task 3.1). To achieve this goal, a clear variable metadata profile is provided (see Appendix 6.5) following the FAIR principles (data made findable, accessible, interoperable and reusable).

To make data input into PANDEM-2 as simple as possible for end users, while keeping track of the semantic of all data flowing through the system, we have designed a 'data labelling schema' (DLS) to meet the agreed database structure for WP3 (D3.1).

2 Introduction & Background

This deliverable 2.2 (D2.2) is an output from WP2 task 2.1 (Mapping and selection of data sources and analytic tools) which consists of identifying data sources and tools for pandemic management and evaluating their timeliness, accessibility, and usefulness in order to select the more relevant ones according to the expert end-user partners in the consortium. The output from this deliverable represents the basis for the deliverable 2.6 (D2.6), ''Software to extract pandemic indicators from selected data sources'' where the list of variables collected from the different data sources will be included. D2.2 is also related with other tasks in WP2 such as task 2.4 (Indicators for pandemic monitoring) where public health agencies and first responders are identifying key indicators for pandemic management and EPIC will adapt or develop new tools to extract the required information. Besides providing an initial and a refined list of data sources, including open access data sources and data uploaded by end users (restricted or sharable), this deliverable includes the variable profile (FAIR metadata profile) that the EPIC and PANDEM-2 technical team is using to include the variables and indicators in the PANDEM-2 database through the software under development. Consequently, D2.2. substantially contributes to the general goal of WP2: "to identify, capture, normalise and aggregate surveillance data from multiple sources for situational awareness".

Finally, this document specifies the structure, attributes, users and purpose of the selected data sources and aligns with CLAR on the database structure in which all data sources will be inputted for WP3 (Situational awareness) and WP4 (Pandemic planning). Gaining access to structured (Go.Data, TESSy, Influenzanet, national agencies, LIMS data…) and unstructured (social media) surveillance data and aggregating it into a common data model and database is the key purpose of WP2. Once this data will be available (or synthetic data has been generated), the dashboard and planning functions can be tested.

Therefore, this deliverable identifies the data sources, how to integrate them in the PANDEM-2 database and describes the "variable profile" for all the information that will be gathered in WP2. Information and sources will be selected according the WP2 overall objectives:

(1) the identification, description and selection of analytic tools and surveillance systems based on traditional (practitioner, hospital, laboratory) and non-traditional data sources (e.g., school/work absenteeism, environmental data, airline flight path data, social media, pharmacy sales, etc.) adapted to monitor pandemics (e.g., real-time data).

(2) the strengthening of participatory surveillance systems.

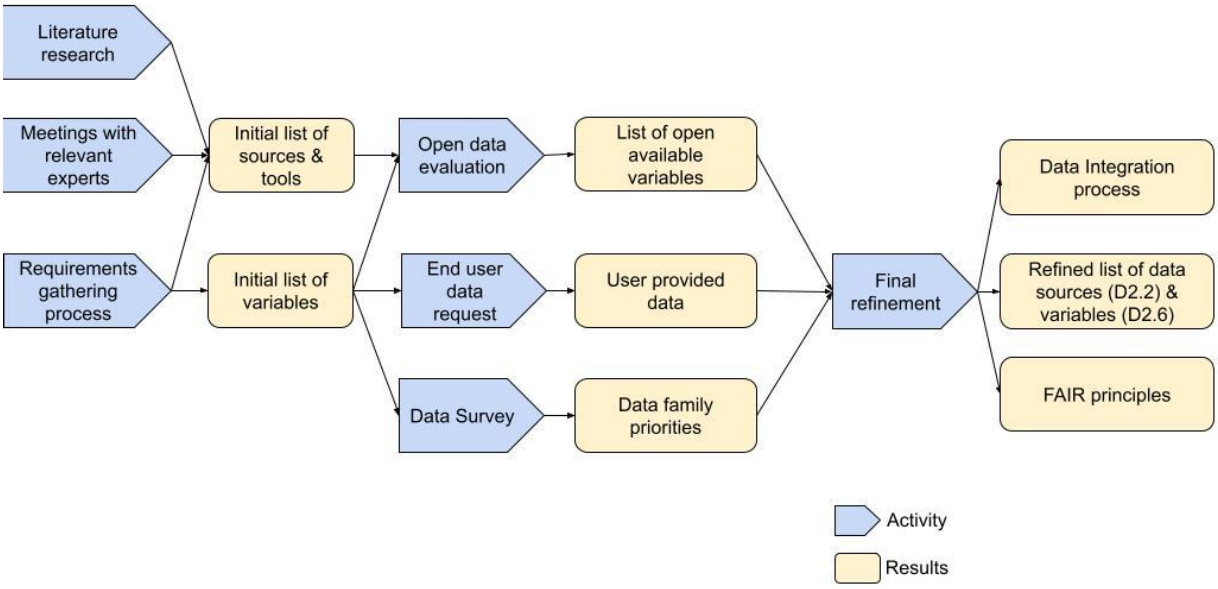
(3) the design and development of a novel text mining system to support community/participatory surveillance.

(4) the integration of the identified data sources (traditional and non-traditional; structured and unstructured) into a database and adaptation/development of analytic tools to generate key indicators for pandemics monitoring to be used in WP3.

(5) the identification and mapping of laboratory data for pandemic detection and monitoring including Next Generation Sequencing (NGS) data and integration with other data sources.

3 Approach

The process followed by the PANDEM-2 team for defining a refined list of variables and data sources was designed to cover current and future needs for pandemic management and take in consideration data availability at public and national level. The process was not linear but rather an incremental and parallel process. The diagram here below schematizes this approach by stating tasks, outputs and dependencies.


Literature research
Meetings with relevant experts
Requirements gathering process
Initial list of sources & tools
Initial list of variables
Open data evaluation
End user data request
Data Survey
List of open available variables
User provided data
Data family priorities
Final refinement
Data Integration process
Refined list of data sources (D2.2) & variables (D2.6)
FAIR principles
Activity
Results

In order to gather necessary input for obtaining an initial list of sources and tools, three parallel activities were driven. a) Literature research and b) Meetings with relevant experts and c) a requirement gathering process including pandemic managers within the consortium. This last step also produced an initial listing of necessary variables for pandemic management.

The initial list of variables based on our consortium partners requirements was used as an input for three parallel steps c.1) A data survey among responders within the consortium to establish data availability and importance c.2) A template containing all initial variables was designed and provided to users to collect available data within the institution and c.3) An open data evaluation process was driven in order to obtain the list of openly available variables with possibility of automated data extraction.

All these elements were analysed to produce three final outputs a) A coherent data integration process, b) A refined list of sources (the final list of variables will be available on D2.6), and c) to analyse if the data included in PANDEM-2 will follow the FAIR principles.

3.1. The literature research

Between February and May 2021, a detailed search of potential pandemic related data sources and tools was performed. The EPIC team conducted a literature review based on different keyword strategies to identify open data sources and tools in the scientific literature (Appendix 6.1). However, we found that most of these sources were not present or accessible in medical browsers such as PubMed. This highlighted that data sources and tools created during the COVID-19 pandemic were done based on demand and not published as a scientific manuscript.

3.2. Meetings with Experts

Therefore, almost all the resources and tools were identified through weekly/biweekly meetings or contacts with experts from relevant organizations such ECDC or WHO, partners in the consortium (in particular end-users) or from other H2020 consortiums or through general internet searches. The results were shared in different meetings held during the same period with all the partners in the consortium and with ECDC, WHO, and colleagues from other H2020 projects, which provide feedback including providing some additional potential data sources and tools that were included in the final list (Appendix 6.4, contains the summary list of the main sources and tools identified).

3.3. Requirement gathering process

During the first 6 months of the project, the PANDEM-2 consortium organised plenary meetings and workshops group to gather feedback from different kinds of end-users (public health agencies, first responders and others such as researchers or microbiologists) about data needs and dashboard requirements for pandemic management. After the requirements were classified by CLAR, EPIC listed, grouped, and further defined those related with task 2.1 (Mapping and selection of data sources and analytic tools) and 2.4 (Indicators for pandemic monitoring). This list and classification can be seen in Appendix 6.2.

3.4. Data availability and priority survey

As a final step, we requested feedback to this list from all the partners by filing a "data availability and priority survey" and by clarifying and suggesting additional items or changes when needed (see Appendix 6.3). In the survey, we also gathered data about timeliness (e.g., daily, weekly or monthly basis) and the available geographical level for each item (e.g., municipality, region, or national).

3.5. End user data request

A template excel file was created based on the initial list of variables in order to allow end users to collect data from their own institutions. The objective of this task was to verify the feasibility of collecting data from different countries into a single database and to evaluate difficulties in data collection. In order to ensure data privacy, the template was designed to include only aggregated data. Four end-users were able to fill the provided template at least partially (RIVM, RKI, THL and NIPH) and will be included into the final software.

3.6. Open data evaluation

The initial list of variables and sources went through a technical evaluation in order to determine the feasibility of automatic data extraction and evaluate the most suitable sources for when more than one was identified.

3.7. Final refinement

All the collected inputs were analysed and a data collection strategy was defined by each data family and indicator. This information will be useful to select the most valuable data to include in the data gathering process and to test the software and the first PANDEM-2 prototype to be installed by our end users. Then, with all this information, we detailed the list of sources and data priorities to elaborate the deliverable D2.2 which represents the first step to identify the final list of variables and indicators to be extracted through the software (D2.6). Furthermore, the variables and indicators identified by partners from other tasks or WP (NGS, Social data mining, Influenza net, GO.DATA ...) are being added to the final list. Finally, we detail how these data sources could integrate information in PANDEM-2 and the profile variables the technical team will follow when creating them. During this step we also defined a standardised approach for data integration and how the data will accomplish the FAIR principles.

4 Results

4.1. Initial list of variables (Appendix 6.2)

Based on the consortium partners requirements and using the proposed methodology “Data labelling Schema (DLS)” we have identified 115 items or variables according to the end user requirements (Appendix 6.2). The list of requirements was classified in 14 data families.

4.2. Data families' priorities (Appendix 6.3)

For each item on the list of initial variables, we have asked our partners for feedback on the priority and availability of these data through a survey (Appendix 6.3). Public health agencies and first responders in the PANDEM-2 consortium reported that they had available data (publicly or restricted) for most of the data related to requirements that they classified as being a high priority. Some of the data classified as being a high priority but not available for some partners seem available for others, which indicates the potential to solve data needs, if data protection rules allow this in each case. We have planned to include those data needs classified as high priority in the final list of PANDEM-2 variables (D2.6), even in case data was not available for some end users. Those data classified as high priority and currently available by end users will be the first data feeding the PANDEM-2 database.

Based on survey results, families were classified into the following groups.

- Important and available (Cases, Deaths, vaccination, patient, tests, Lab)

Important variables for pandemic response and most of them are publicly reported. Often the user requirements request a level of granularity which is not publicly available, but the data exists and could potentially be (or is) shared between member states.

- Important with limited availability (NGS, Contact tracing, Population study)

Variables on this group are also important for pandemic response but in most cases, they are only available locally and not in the public domain. This data is mainly collected at individual level and contains personal information. Nevertheless, we estimate that the main issue for data sharing is the lack of standard aggregated indicators that could be easily shared.

- Lack of data or low priority: First response (including staff), transport, referentials, measures (including flights), emergency calls, resources

Most variables in this group are not systematically collected or not important for pandemic response. Even though some of these are important for end users it would not be realistic to expect data sharing in the short term.

- Not in requirements: Weather, seroprevalence

Variables on this group were not mentioned during the requirement gathering process so they were considered as low priority.

4.3. Initial list of sources and tools (Appendix 6.4)

Most of the sources and tools for pandemic management we found in our literature search (Appendix 6.1) are well known by those working in epidemics, such are the case of influenzanet or the R Epidemics Consortium (RECON) tools. Besides, our findings in the literature review confirmed the necessity of using data coming from different data sources, from traditional - including syndromic surveillance - and non-traditional surveillance (or digital surveillance), for pandemic management and including using "external data" (flights) or data not originally created with epidemiological or surveillance purposes. Moreover, it highlighted the need to integrate social media or mass media data and indicators which have been proven useful for monitoring epidemics or pandemics such as the case of the effective reproductive number. But most of the data sources and tools in our final list (Appendix 6.4) were compiled through meetings within the PANDEM-2 consortium (in particular the end-users) or other key partners (e.g., WHO, ECDC, MOOD project). We have identified that most of the data needs or “core PANDEM-2 data'' to meet our end users’ requirements could only be uploaded in a PANDEM-2 installation by the end-users themselves, which means that they are a “key specific data source”.

In addition, we have included as data sources those identified and relevant not only for our end users but also for modelling or pandemic communications such as flight, or animal or human alerts platforms. Similarly, data sources in any task within WP2 such as Lab data (including NGS-metagenomics see task 2.5), participatory surveillance data (Task 2.2 influenzanet), or data coming from social media (Task 2.3) were also integrated as potential useful data sources for pandemic management.

4.4. FAIR principles (Appendix 6.5)

The PANDEM-2 database is a data curation effort including many heterogeneous data sources into a coherent model and as such it will be made available in an open data platform in order to fulfil FAIR principles as described on Appendix 6.5.

Moreover, we have described the profile and characteristics needed for the data or variables, following the FAIR principles (Appendix 6.5), which will be finally included in the PANDEM-2 database and how all these data sources will be integrated in the database (Appendix 6.6) which has been defined in the PANDEM-2 data schema (D3.1).

4.5. Data sources integration process (Appendix 6.6)

A formalism has been proposed in Appendix 6.6 named “Data Labelling Schema” (DLS) in order to support integration of all necessary data. This schema is generalistic enough to provide a generic approach for integrating any identified variables, so we expect it to support the necessary flexibility for future pandemics data needs.

4.6. List of Open available variables (Appendix 6.7)

After open data technical evaluation a detailed list of variables for those open data sources identified so far (Appendix 6.7) is provided which represent the first variables feeding the PANDEM-2 database and that in a few months, together with end-user data uploaded in their own PANDEM-2 installations (restricted or not) will allow the use of the PANDEM-2 tools for pandemic management to be tested within the PANDEM-2 project in WP6.

5 Impact & Conclusion

This deliverable compiles the most relevant data sources for pandemic management including open-source data which can support public health agencies and first responders in their responsibilities or activities related with current and future epidemics and pandemics. The document highlights the relevance to continue working on data normalisation and following FAIR principles that are the guarantee to ensure data sharing across countries. Data sharing as has been seen in current COVID-19 pandemic, is paramount to allow prevention and control of cross-border epidemics and public health threats. We also described resource data, from variables related to contact tracing to variables related to hospital or primary care, which has been described as a limiting factor in providing a high-quality response to avoid severe disease cases or deaths.

Within the PANDEM-2 project, this deliverable provides an overview of the requirements and data sources and describes how to integrate specific variables into the final PANDEM-2 database by month 18 of the project. It is an output of all the consortium teamwork trying to synthesise lessons learned from the current pandemic to improve pandemic data management and collaboration between organisations and countries when working in epidemic preparedness and response at international level, which is the only way to really tackle pandemics.

The results of this deliverable directly feed into the subsequent D2.6 deliverable "Software to extract pandemic indicators from selected data sources".

6 Appendix: DATA SOURCES for the PANDEM-2 PROJECT

PROJECT DETAILS

Acronym:	PANDEM-2
Title:	Pandemic preparedness and response
Project coordinator:	NUIG (Máire Connolly)
Programme:	H2020-SU-SEC-2019
Topic:	SU-DRS05-2019
Type:	Research and Innovation Action (RIA)
Grant agreement #	883285
Start:	01 February 2021
Duration:	24 months
Website:	https://pandem-2.eu/

List of Beneficiaries:

Full Name	**Short Name**	**Beneficiary Number**	**Role**
NATIONAL UNIVERSITY OF IRELAND GALWAY	NUIG	1	Coordinator
FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V	FRAUNHOFER	2	Beneficiary
UNIVERSITE CATHOLIQUE DE LOUVAIN	UCL	3	Beneficiary
PINTAIL LTD	PT	4	Beneficiary
FOLKHALSOMYNDIGHETEN	FOHM	5	Beneficiary
RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEU	RIVM	6	Beneficiary
CARR COMMUNICATIONS LIMITED	CARR	7	Beneficiary
TERVEYDEN JA HYVINVOINNIN LAITOS	THL	8	Beneficiary
INSTITUTUL NATIONAL DE SANATATE PUBLICA	NIPH	9	Beneficiary
ROBERT KOCH-INSTITUT	RKI	10	Beneficiary
STICHTING KATHOLIEKE UNIVERSITEIT	RUNMC	11	Beneficiary
CLARISOFT TECHNOLOGIES ROM SRL	CLARISOFT	12	Beneficiary
OSTERREICHISCHES ROTES KREUZ	ORK	13	Beneficiary
EPICONCEPT	EPIC	14	Beneficiary
INSTITUTO NACIONAL DE EMERGÊNCIA MÉDICA	INEM	15	Beneficiary
TRILATERAL RESEARCH LTD	TRI	16	Beneficiary
ISTITUTO PER L'INTERSCAMBIO SCIENTIFICO	ISI	17	Beneficiary
ASSOCIAZIONE DELLA CROCE ROSSA ITALIANA	ITRC	18	Beneficiary
INSTITUTO NACIONAL DE SAUDE DR. RICARDO JORGE	INSA	19	Beneficiary

6.1 Literature review of data sources and tools for pandemic management

These are three strategies we followed to find relevant data sources and tools for pandemic management:

6.1.1. Search terms strategy 1

(for last 10 years)

(((pandemics [MeSH Terms]) OR (epidemics [MeSH Terms]) OR (disease outbreaks [MeSH Terms])) AND ((monitoring [Title/Abstract]) OR (preparedness [Title/Abstract]) OR (management [Title/Abstract]) OR (response [Title/Abstract]) OR (Epidemiological Monitoring [MeSH Terms]) OR (emergency preparedness [MeSH Terms]) OR (early detection of disease [MeSH Terms]) OR (Occupational health [MeSH Terms]) OR (One Health [MeSH Terms]) OR (Environmental Monitoring [MeSH Terms]) OR (animal surveillance [Title/Abstract]) AND ((Infectious Disease Medicine [MeSH Terms] OR infectious disease [Title/Abstract] OR communicable disease [MeSH Terms]) OR (emerging communicable disease [MeSH Terms])))

Number of results (by February 8): 1,804 results

6.1.2. Search terms strategy 2

(for last 10 years)

(((Health Information Systems [MeSH Terms]) OR (Data Collection [MeSH Terms] OR data sources [Title/Abstract] OR Software Tools[MeSH Terms])) AND ((pandemics [MeSH Terms]) OR (epidemics [MeSH Terms]) OR (disease outbreaks [MeSH Terms])) AND ((monitoring [Title/Abstract]) OR (preparedness [Title/Abstract]) OR (management [Title/Abstract]) OR (response [Title/Abstract]) OR (Epidemiological Monitoring [MeSH Terms]) OR (emergency preparedness [MeSH Terms]) OR (early detection of disease [MeSH Terms]) OR (Occupational health [MeSH Terms]) OR (One Health [MeSH Terms]) OR (Environmental Monitoring [MeSH Terms]) OR (animal surveillance [Title/Abstract])) AND ((Infectious Disease Medicine [MeSH Terms] OR infectious disease [Title/Abstract] OR communicable disease [MeSH Terms]) OR (emerging communicable disease [MeSH Terms])))

Number of results (by February 8): 619 results

6.1.3. Qualitative selection

Two further qualitative selections of tools and research articles, first of 181 results and a final one of 49 (by February 8)

1. ICARES: a real-time automated detection tool for clusters of infectious diseases in the Netherlands. https://pubmed.ncbi.nlm.nih.gov/28279150/
2. The Spatiotemporal Epidemiological Modeler (STEM) Project. http://www.eclipse.org/stem/
3. OutbreakTools: a new platform for disease outbreak analysis using the R software. https://pubmed.ncbi.nlm.nih.gov/24928667/

4. EpiBasket: how e-commerce tools can improve epidemiological preparedness. Xing W, Hejblum G, Valleron AJ. Emerg Health Threats J. 2013 Oct 31;6:19748. doi: 10.3402/ehtj.v6i0.19748.
5. A system for automated outbreak detection of communicable diseases in Germany. Salmon M, Schumacher D, Burmann H, Frank C, Claus H, Höhle M. Euro Surveill. 2016;21(13). doi: 10.2807/1560-7917.ES.2016.21.13.30180.
6. Developing open source, self-contained disease surveillance software applications for use in resource-limited settings. Campbell TC, Hodanics CJ, Babin SM, Poku AM, Wojcik RA, Skora JF, Coberly JS, Mistry ZS, Lewis SH.
7. The EpiCanvas infectious disease weather map: an interactive visual exploration of temporal and spatial correlations. Gesteland PH, Livnat Y, Galli N, Samore MH, Gundlapalli AV.
8. Epinome: a visual-analytics workbench for epidemiology data. Livnat Y, Rhyne TM, Samore MH.
9. Covidom, a Telesurveillance Solution for Home Monitoring Patients With COVID-19. Yordanov Y, Dechartres A, Lescure X, Apra C, Villie P, Marchand-Arvier J, Debuc E, Dinh A, Jourdain P; AP-HP / Universities / Inserm COVID-19 Research Collaboration.
10. Dynamics of Interorganizational Public Health Emergency Management Networks: Following the 2015 MERS Response in South Korea. Kim K, Jung K.
11. Improving Animal Disease Detection Through an Enhanced Passive Surveillance Platform. Thompson CW, Holmstrom L, Biggers K, Wall J, Beckham T, Coats M, Korslund J, Colby MM. Health Secur. 2016 Jul-Aug;14(4):264-71. doi: 10.1089/hs.2016.0016. Epub 2016 Jul 15.
12. Performance of Digital Contact Tracing Tools for COVID-19 Response in Singapore: Cross-Sectional Study. Huang Z, Guo H, Lee YM, Ho EC, Ang H, Chow A.
13. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, Koppeschaar C, Rehn M, Smallenburg R, Turbelin C, Van Noort S, Vespignani A.
14. Citizen-Centered Mobile Health Apps Collecting Individual-Level Spatial Data for Infectious Disease Management: Scoping Review. Wirth FN, Johns M, Meurers T, Prasser F.
15. Douglas JV, Bianco S, Edlund S, Engelhardt T, Filter M, Günther T, Hu KM, Nixon EJ, Sevilla NL, Swaid A, Kaufman JH. STEM: An Open Source Tool for Disease Modeling. Health Secur. 2019 Jul/Aug;17(4):291-306. doi: 10.1089/hs.2019.0018. PMID: 31433284; PMCID: PMC6708268.
16. Álvarez-Díaz DA, Laiton-Donato K, Franco-Muñoz C, Mercado-Reyes M. SARS- CoV-2 sequencing: The technological initiative to strengthen early warning systems for public health emergencies in Latin America and the Caribbean. Biomedica. 2020 Oct 30;40(Supl. 2):188-197. English, Spanish. doi: 10.7705/biomedica.5841. PMID: 33152203; PMCID: PMC7676827.
17. Fähnrich C, Denecke K, Adeoye OO, Benzler J, Claus H, Kirchner G, Mall S, Richter R, Schapranow MP, Schwarz N, Tom-Aba D, Uflacker M, Poggensee G, Krause G. Surveillance and Outbreak Response Management System (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa. Euro Surveill. 2015 Mar 26;20(12):21071. doi: 10.2807/1560-7917.es2015.20.12.21071. PMID: 2584649.
18. Jourdain F, Roiz D, Perrin Y, Grucker K, Simard F, Paupy C. Facteurs entomologiques d'émergence des arboviroses [Entomological factors of arboviruses emergences].

Transfus Clin Biol. 2015 Aug;22(3):101-6. French. doi: 10.1016/j.tracli.2015.06.001. Epub 2015 Jun 30. PMID: 26141429.

19. Nouvellet P, Cori A, Garske T, Blake IM, Dorigatti I, Hinsley W, Jombart T, Mills HL, Nedjati-Gilani G, Van Kerkhove MD, Fraser C, Donnelly CA, Ferguson NM, Riley S. A simple approach to measure transmissibility and forecast incidence. Epidemics. 2018 Mar;22:29-35. doi: 10.1016/j.epidem.2017.02.012. Epub 2017 Feb 24. PMID: 28351674; PMCID: PMC5871640.
20. Bertolini G, Nattino G, Langer M, Tavola M, Crespi D, Mondini M, Rossi C, Previtali C, Marshall J, Poole D; GiViTI. The role of the intensive care unit in real-time surveillance of emerging pandemics: the Italian GiViTI experience. Epidemiol Infect. 2016 Jan;144(2):408-12. doi: 10.1017/S0950268815001399. Epub 2015 Jun 29. PMID: 26119282.
21. Torres G, Ciaravino V, Ascaso S, Flores V, Romero L, Simón F. Syndromic surveillance system based on near real-time cattle mortality monitoring. Prev Vet Med. 2015 May 1;119(3-4):216-21. doi: 10.1016/j.prevetmed.2015.03.003. Epub 2015 Mar 17. PMID: 25827083.
22. Velasco E. Disease detection, epidemiology and outbreak response: the digital future of public health practice. Life Sci Soc Policy. 2018 Apr. 1;14(1):7. doi: 10.1186/s40504-018-0071-4. PMID: 29607463; PMCID: PMC5879035.
23. Bennouar S, Bachir Cherif A, Kessira A, Hamel H, Boudahdir A, Bouamra A, Bennouar D, Abdi S. Usefulness of biological markers in the early prediction of coronavirus disease-2019 severity. Scand J Clin Lab Invest. 2020 Dec;80(8):611-618. doi: 10.1080/00365513.2020.1821396. Epub 2020 Sep 18. PMID: 32945705.
24. Spreco A, Timpka T. Algorithms for detecting and predicting influenza outbreaks: meta narrative review of prospective evaluations. BMJ Open. 2016 May 6;6(5):e010683. doi: 10.1136/bmjopen-2015-010683. PMID: 27154479; PMCID: PMC4861093.
25. Hong YR, Lawrence J, Williams D Jr, Mainous III A. Population-Level Interest and Telehealth Capacity of US Hospitals in Response to COVID-19: Cross-Sectional Analysis of Google Search and National Hospital Survey Data. JMIR Public Health Surveill. 2020 Apr 7;6(2):e18961. doi: 10.2196/18961. PMID: 32250963; PMCID: PMC7141249.
26. Alwashmi MF. The Use of Digital Health in the Detection and Management of COVID-19. Int J Environ Res Public Health. 2020 Apr 23;17(8):2906. doi: 10.3390/ijerph17082906. PMID: 32340107; PMCID: PMC7215737.
27. Ho HJ, Zhang ZX, Huang Z, Aung AH, Lim WY, Chow A. Use of a Real-Time Locating System for Contact Tracing of Health Care Workers During the COVID-19 Pandemic at an Infectious Disease Center in Singapore: Validation Study. J Med Internet Res. 2020 May 26;22(5):e19437. doi: 10.2196/19437. PMID: 32412416; PMCID: PMC7252199.
28. Desai AN, Kraemer MUG, Bhatia S, Cori A, Nouvellet P, Herringer M, Cohn EL, Carrion M, Brownstein JS, Madoff LC, Lassmann B. Real-time Epidemic Forecasting: Challenges and Opportunities. Health Secur. 2019 Jul/Aug;17(4):268-275. doi: 10.1089/hs.2019.0022. PMID: 31433279; PMCID: PMC6708259.
29. Hoti K, Jakupi A, Hetemi D, Raka D, Hughes J, Desselle S. Provision of community pharmacy services during COVID-19 pandemic: a cross sectional study of community pharmacists' experiences with preventative measures and sources of information. Int J Clin Pharm. 2020 Aug;42(4):1197-1206. doi: 10.1007/s11096-020-01078-1. Epub 2020 Jun 11. PMID: 32524513; PMCID: PMC7286815.
30. Ridenhour B, Kowalik JM, Shay DK. Unraveling R0: considerations for public health applications. Am J Public Health. 2014 Feb;104(2):e32-41. doi:

10.2105/AJPH.2013.301704. Epub 2013 Dec 12. PMID: 24328646; PMCID: PMC3935673.

31. Joshi A, Sparks R, Karimi S, Yan SJ, Chughtai AA, Paris C, MacIntyre CR. Automated monitoring of tweets for early detection of the 2014 Ebola epidemic. PLoS One. 2020 Mar 17;15(3):e0230322. doi: 10.1371/journal.pone.0230322. PMID: 32182277; PMCID: PMC7077840.
32. Abat C, Chaudet H, Rolain JM, Colson P, Raoult D. Traditional and syndromic surveillance of infectious diseases and pathogens. Int J Infect Dis. 2016 Jul;48:22-8. doi: 10.1016/j.ijid.2016.04.021. Epub 2016 Apr 30. PMID: 27143522; PMCID: PMC7110877.
33. Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, Mehta S, Solomon S, Labrique A, Kishore N, Buckee CO, Wesolowski A. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nat Commun. 2020 Sep 30;11(1):4961. doi: 10.1038/s41467-020-18190-5. PMID: 32999287; PMCID: PMC7528106.
34. Bourhis Y, Gottwald T, van den Bosch F. Translating surveillance data into incidence estimates. Philos Trans R Soc Lond B Biol Sci. 2019 Jul 8;374(1776):20180262. doi: 10.1098/rstb.2018.0262. PMID: 31104599; PMCID: PMC6558556.
35. Cori A, Donnelly CA, Dorigatti I, Ferguson NM, Fraser C, Garske T, Jombart T, Nedjati-Gilani G, Nouvellet P, Riley S, Van Kerkhove MD, Mills HL, Blake IM. Key data for outbreak evaluation: building on the Ebola experience. Philos Trans R Soc Lond B Biol Sci. 2017 May 26;372(1721):20160371. doi: 10.1098/rstb.2016.0371. PMID: 28396480; PMCID: PMC5394647.
36. Kim K, Jung K. Dynamics of Interorganizational Public Health Emergency Management Networks: Following the 2015 MERS Response in South Korea. Asia Pac J Public Health. 2018 Apr;30(3):207-216. doi: 10.1177/1010539518762847. Epub 2018 Mar 21. PMID: 29561166.
37. Dai Y, Wang J. Identifying the outbreak signal of COVID-19 before the response of the traditional disease monitoring system. PLoS Negl Trop Dis. 2020 Oct 1;14(10):e0008758. doi: 10.1371/journal.pntd.0008758. PMID: 33001985; PMCID: PMC7553315.
38. Schellpfeffer N, Collins A, Brousseau DC, Martin ET, Hashikawa A. Web-Based Surveillance of Illness in Childcare Centers. Health Secur. 2017 Sep/Oct;15(5):463-472. doi: 10.1089/hs.2016.0124. Epub 2017 Sep 22. PMID: 28937791; PMCID: PMC6913116.
39. Jian SW, Chen CM, Lee CY, Liu DP. Real-Time Surveillance of Infectious Diseases: Taiwan's Experience. Health Secur. 2017 Mar/Apr;15(2):144-153. doi: 10.1089/hs.2016.0107. PMID: 28418738; PMCID: PMC5404256
40. Zens M, Brammertz A, Herpich J, Südkamp N, Hinterseer M. App-Based Tracking of Self-Reported COVID-19 Symptoms: Analysis of Questionnaire Data. J Med Internet Res. 2020 Sep 9;22(9):e21956. doi: 10.2196/21956. PMID: 32791493; PMCID: PMC7480999.
41. Thompson CW, Holmstrom L, Biggers K, Wall J, Beckham T, Coats M, Korslund J, Colby MM. Improving Animal Disease Detection Through an Enhanced Passive Surveillance Platform. Health Secur. 2016 Jul-Aug;14(4):264-71. doi: 10.1089/hs.2016.0016. Epub 2016 Jul 15. PMID: 27419928.
42. Muscatello DJ, Chughtai AA, Heywood A, Gardner LM, Heslop DJ, MacIntyre CR. Translation of Real-Time Infectious Disease Modeling into Routine Public Health Practice. Emerg Infect Dis. 2017 May;23(5):e161720. doi: 10.3201/eid2305.161720. PMID: 28418309; PMCID: PMC5403034.

43. Drury G, Jolliffe S, Mukhopadhyay TK. Process mapping of vaccines:Understanding the limitations in current response to emerging epidemic threats. Vaccine. 2019 Apr 17;37(17):2415-2421. doi: 10.1016/j.vaccine.2019.01.050. Epub 2019 Mar 22. PMID: 30910404; PMCID: PMC7173310.
44. Li X, Liu C, Mao Z, Xiao M, Wang L, Qi S, Zhou F. Predictive values of neutrophil-to-lymphocyte ratio on disease severity and mortality in COVID-19 patients: a systematic review and meta-analysis. Crit Care. 2020 Nov 16;24(1):647. doi: 10.1186/s13054-020-03374-8. PMID: 33198786; PMCID: PMC7667659.
45. Khan K, McNabb SJ, Memish ZA, Eckhardt R, Hu W, Kossowsky D, Sears J, Arino J, Johansson A, Barbeschi M, McCloskey B, Henry B, Cetron M, Brownstein JS. Infectious disease surveillance and modelling across geographic frontiers and scientific specialties. Lancet Infect Dis. 2012 Mar;12(3):222-30. doi: 10.1016/S1473-3099(11)70313-9. Epub 2012 Jan 16. PMID: 22252149.
46. Sha D, Miao X, Lan H, Stewart K, Ruan S, Tian Y, Tian Y, Yang C. Spatiotemporal analysis of medical resource deficiencies in the U.S. under COVID-19 pandemic. PLoS One. 2020 Oct 14;15(10):e0240348. doi: 10.1371/journal.pone.0240348. PMID: 33052956; PMCID: PMC7556467.
47. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, Koppeschaar C, Rehn M, Smallenburg R, Turbelin C, Van Noort S, Vespignani A. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clin Microbiol Infect. 2014 Jan;20(1):17-21. doi: 10.1111/1469-0691.12477. PMID: 24350723; PMCID: PMC7128292.
48. Morton MJ, Hsu EB, Shah SH, Hsieh YH, Kirsch TD. Pandemic influenza and major disease outbreak preparedness in US emergency departments: A selected survey of emergency health professionals. Am J Disaster Med. 2020 Fall;14(4):269-277. doi: 10.5055/ajdm.2019.0340. PMID: 32803746.
49. Kool JL, Paterson B, Pavlin BI, Durrheim D, Musto J, Kolbe A. Pacific-wide simplified syndromic surveillance for early warning of outbreaks. Glob Public Health. 2012;7(7):670-81. doi: 10.1080/17441692.2012.699536. Epub 2012 Jul 24. PMID: 22823595; PMCID: PMC3419547.

6.1.4. Conclusions on the literature research:

Between February and May 2021, a detailed search of potential pandemic related data sources and tools was performed. A literature review based on different keyword strategies to identify open data sources and tools in the scientific literature. However, we found that most of these sources were not present or accessible in medical browsers such as PubMed. This highlighted that data sources and tools created during the COVID-19 pandemic were done based on demand and not published as a scientific manuscript.

Therefore, almost all the resources and tools were identified through weekly/biweekly meetings or contacts with experts from relevant organizations such ECDC or WHO, partners in the consortium (in particular end-users) or from other H2020 consortiums or through general internet searches. The results were shared in different meetings held during the same period with all the partners in the consortium and with ECDC, WHO, and colleagues from other H2020 projects, which provide feedback including providing some additional potential data sources and tools that were included in the final list (Appendix 6.4, contains the summary list of the main sources and tools identified).

6.2 Initial list of variables

As described in deliverable D3.1 (database schema and set-up report), the list of requirements was divided in 14 data families and each item on the list was classified as either an observation, a characteristic, an indicator, a resource, a document or a referential. We formalize these definitions here below:

- **Observation**: Any variable or item which reflects something that was measured or observed and reported by a data source. These variables are mainly numeric and identify concepts like the number of people or resources but could also be a statement (tweet) published by a person.
- **Indicator**: Indicators are a particular case of observations used to monitor specific questions and, in some cases, are characterized as calculations which have a reference methodology and formula associated but they could be directly reported by users.
- **Characteristic**: any variable or item which describes or categorises an observation. It is always related to variables of the type of observation. It could be by example the country, the date or the variant of observed cases.
- **Referential**: Referential are a particular case of characteristics, they are characterized by the fact of having a unique code that can map to other characteristics. An example of referential are municipalities. Each municipality has unique code and links to a country.
- **Resources**: At a particular case of referential that concerns resources needed to deal with pandemics. They play a particularly important role in PANDEM-2 for resource modelling features (e.g., nurses' beds or material resources). They are associated with observations like "number of available resources"
- **Document**: refer to any document, generic term. It is yet to be defined if PANDEM-2 documents will be stored on the PANDEM-2 database or on a dedicated location.

The following list is the result of the dedicated analysis performed of user requirements and was described in deliverable D3.1. Each of these items make sense from a functional point of view but in some cases, further refinement is necessary. Ultimately each source will need to map to one of these elements. More details about these items including details like data type, relationship and how to map them to data sources are provided in section 6.6.1 particularly in subsection C1.

The refinement of variables will be performed iteratively when integrating data sources. At that moment user requirements will be compared with data availability on existing sources (public or restricted) and the level of detail will be adjusted. The importance of this list is to define the functional scope of data that has been identified by users to ensure it can be easily integrated into the PANDEM-2 database.

The current definition of variables is:

Data Family	Variables/items	Variable/item Type
01. Cases	Confirmed cases	Observation
01. Cases	Suspected cases (possible, probable and unclassified)	Observation
01. Cases	Cases per variant	Characteristic
01. Cases	Cases per (age, sex, comorbidities)	Characteristic
01. Cases	Incidence rates (last week, two weeks, month, and other not known potential time period)	Indicator
01. Cases	Incidence rates (age, sex, comorbidities, variant)	Indicator
01. Cases	Rt number	Indicator
01. Cases	PPE Protective equipment (stock, type & need)	Resources
01. Cases	Outbreak Id (if associated to known outbreak)	Observation
01. Cases	N° of patients per severity level	Characteristic
01. Cases	N° and proportion of imported cases	Observation
02. Deaths	N° of deaths by X	Observation
02. Deaths	N° of deaths	Observation
02. Deaths	Mortality rates (age, sex, comorbidities)	Indicator
02. Deaths	Mortality rates (with X) (last week, two weeks, month and other not known potential time period)	Indicator
02. Deaths	Mortality rates with X (age, sex, comorbidities, variant)	Indicator
03. Patients	N° of infected patients per bed type (clinic care, ICU, ventilator)	Observation
03. Patient	N° of non-infected patients per bed type	Observation
03. Patient	Length of stay	Observation
03. Patient	Patient current status (recovered, death, still in care)	Observation
03. Patient	N° of patients per facility type (hospital / LTHF / primary care and other not known potential values)	Characteristic
03. Patient	Treatment received	Characteristic
03. Patient	Patients potential risk factors (age, gender, comorbidities, municipality of residence)	Characteristic

03. Patient	Hospital staff	Resources
03. Patient	Patient resources: Beds, ventilators, oxygen, medicines, disinfection materials, ICU supplies...	Resources
03. Patient	Hospital staff type e.g., intensive care, emergency care, etc...	Resources
03. Patient	Bed type (ICU, regular clinical wards, emergency)	Resources
03. Patient	Primary care staff	Resources
03. Patient	Beds/Room occupancy and types (e.g., isolation)	Resources
03. Patient	Vaccination Status	Observation
04.Tests	Number of tests performed by type	Observation
04.Tests	Link between test results and epidemiological surveys	Observation
04.Tests	Test type details (brand and characteristics)	Characteristic
04.Tests	Test results (positive, negative, unknown, pending)	Characteristic
04.Tests	Positivity rate	Indicator
04.Tests	Test resources (staff, supplies)	Resources
05. Vaccination	Doses injected	Observation
05. Vaccination	people that have received at least one dose	Observation
05. Vaccination	People fully vaccinated	Observation
05. Vaccination	Doses scheduled and target population	Observation
05. Vaccination	Doses injected by age group, risk group, and brand/type	Characteristic
05. Vaccination	Doses by vendor, batch	Characteristic
05. Vaccination	Doses injected by occupation (HCW and other essential professionals and other not known potential values...)	Characteristic
05. Vaccination	Doses injected in high-risk individuals - potential risk factors (immunosuppressed, comorbidities, pregnant women, elderly and other not known potential factors)	Characteristic
05. Vaccination	Vaccination Side effects AEFI observed and severity	Characteristic
05. Vaccination	Vaccination progress (proportion of vaccinated, overall, by age and risk group)	Indicator
05. Vaccination	Vaccination resources (Staff, centres, supplies)	Resources
06. Contact tracing	N° of index cases studied	Observation

06. Contact tracing	Nº of contacts, secondary and tertiary cases per index case	Indicator
06. Contact tracing	N° of clusters found (and cluster type - definition)	Characteristic
06. Contact tracing	Confirmed cases that had travel during infectious period	Observation
06. Contact tracing	Types of contact	Characteristic
06. Contact tracing	Contact tracing details at individual level: travel, contacts, date of isolation, date of quarantine, transmission chains and other not known potential variables...	Observation
06. Contact tracing	Cluster identification and characterisation	Characteristic
06. Contact tracing	Notification delay (onset of symptoms - notification date)	Indicator
06. Contact tracing	Contact tracers (staff working in contact tracing)	Resources
07. Lab	Number of tests performed (overall and by individual)	Observation
07. Lab	Speed of spread of variants (proportion among overall cases)	Observation
07. Lab	seroprevalence (and test type)	Indicator
07. Lab	Seaway water virus presence (and levels)	Observation
07. Lab	Mutations/Sequences-spread and distributions	Characteristic
07. Lab	link lab data with cases/patient data	Characteristic
07. Lab	link with aggregated epidemiological data	Characteristic
07. Lab	Sensibility & specificity of test methods	Document
08. Emergency calls	Monitoring number of emergency calls (overall and by syndrome)	Observation
08. Emergency calls	Comparison current situation with peacetime symptoms, notifications and diagnostic rates.	Observation
08. Emergency calls	Severity of victims (at call and scene)	Characteristic
08. Emergency calls	N° of calls from people declared as confirm case	Observation
08. Emergency calls	Monitoring of symptoms from emergency calls	Characteristic
09. First response	Ongoing emergencies (types)	Observation
09. First response	Visits to general practitioner (GP) with compatible symptoms (disease X)	Observation
09. First response	Details/type of protocol applied	Characteristic
09. First response	Public health Staff (surveillance, prevention and control activities and other not known potential activities...)	Resources

09. First response	Emergency Staff	Resources
10. Transport	Patient transportation type (for suspicious or confirmed cases)	Observation
10. Transport	Current ambulance activity	Observation
10. Transport	Number of patients transferred	Observation
10. Transport	Transport statistics (duration, times)	Characteristic
10. Transport	Transport resources (ambulances)	Resources
10. Transport	Ambulances / type	Resources
10. Transport	patient transport threshold	Resources
10. Transport	patient transport protocols	Document
11. Measures	N° of people entering to the country (by country of origin, and city-airport of entrance)	Observation
11. Measures	Mitigation measures and policies	Observation
11. Measures	Prevention or control measure details: - type (e.g., lockdown) - start - end - place (place or geographical location)	Characteristic
11. Measures	Border rules/laws	Document
12. Population study	Adherence to prevention and control measures	Observation
12. Population study	Are people understanding public health communication	Observation
12. Population study	Alerts & Early warning signals	Indicator
12. Population study	Social media custom analysis	Indicator
12. Population study	Vaccination acceptance willingness	Indicator
12. Population study	Level of trust in the Government and institutions	Indicator
12. Population study	Measure social impact (psychological, lifestyle)	Indicator
12. Population study	Indirect impact on health (other notifiable disease, disruption of services, indirect deaths and morbidity...)	Indicator
12. Population study	People beliefs and opinions on pandemic	Indicator
12. Population study	Most consulted public information sites	Observation
12. Population study	People information needs	Document
13. Referentials	Denominators for potential risk factors or individuals at risk	Referential

13. Referentials	pathogen specific referential epidemiological parameters (Host, vector, latency, contagiousness, Serial interval, Susceptibility...)	Referential
13. Referentials	Symptoms & signs by pathogen	Referential
13. Referentials	Care procedures (for new diseases)	Referential
13. Referentials	Variants (VOI, VOC): variant of interest (VOI), variant of concern (VOC) according to WHO and ECDC indications	Referential
13. Referentials	denominators and maps for different Geographic location (local to international)	Referential
13. Referentials	Population denominators (age, sex, country of birth, vaccination status, comorbidities, age group ...)	Referential
13. Referentials	Social determinants by area or case (country of birth, wealth, studies, occupation...)	Referential
13. Referentials	Care providers by area	Referential
13. Referentials	User shared guidelines	Document
13. Referentials	Places of infection	Document
13. Referentials	supplies for potential or confirmed effective medication	Document
14. Metadata	Variable definitions (calculation method, description)	Referential
14. Metadata	Source contact	Referential
14. Metadata	Data owner	Referential
14. Metadata	Data providers for dashboards	Referential
14. Metadata	Dashboard profile e.g., emergency	Referential
14. Metadata	GDPR compliance	Referential

6.3 Data families priorities

We have asked the PANDEM-2 consortium for feedback on the list of requirements above (Appendix 6.2) in order to clarify data availability and priority for each of the requirements identified. From those partners we had feedback, they had available data (publicly or restricted) for most of the data related with requirements that they classified as being a high priority data need. In the survey, we also gathered data about timeliness (e.g., daily, weekly or monthly basis) and the available geographical level for each item (e.g., municipality, region, or national). This data will be useful in the data gathering process and to test the software and first PANDEM-2 prototype to be installed by our end users.

Nine partners answered the data survey: FOHM, INEM, ISI, NIPH, ORK, RIVM, RKI, RUNMC, and THL. Some of the data classified as being a high priority but not available for some partners seems available for others such is the case of the following items which first responders answered that are not available for them, but seem to be available for PH agencies or at hospital level (RUNMC) (to be further analysed as to if it can be shareable):

Confirmed cases, Cases stratified per age, sex, and comorbidities, Hospital staff type e.g., intensive care, emergency care, etc..., Bed type (ICU, regular clinical wards, emergency), Primary care staff, Beds/Room occupancy and types (e.g., isolation), Comparison current situation with peacetime symptoms, notifications and diagnostic rates, User shared guidelines, PPE Protective equipment (stock, type & need).

Data was aggregated among all answers assigning scores per data availability and priority:

› Data availability score (public and at institution)
"Patient Level" => 4, "Municipality" => 3, "Regional" =>2 "National"=> 1, other => 0

› Data availability GAP = at institution availability score - public availability score

› Data priority score (dashboard, forecasting, resource planning)
"High" => 2, "Medium" => 2, "Low" =>1, other => 0

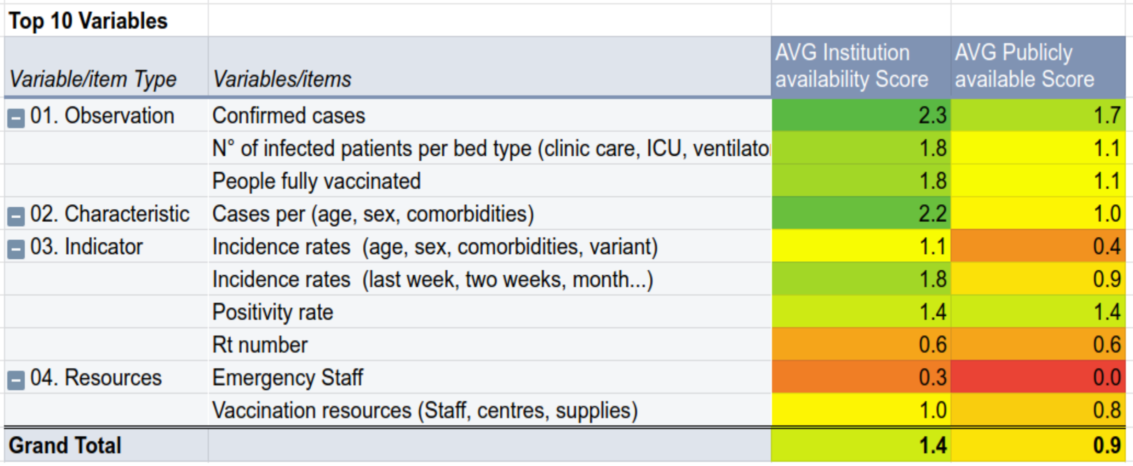
› Total priority = dashboard priority + forecasting priority + resource planning priority

6.3.1 Lessons learned on data survey

In this section we present the main feedback obtained from the Data survey in 10 summarised tables/charts. These lessons were used to identify the most important variables and the level availability of these indicators.

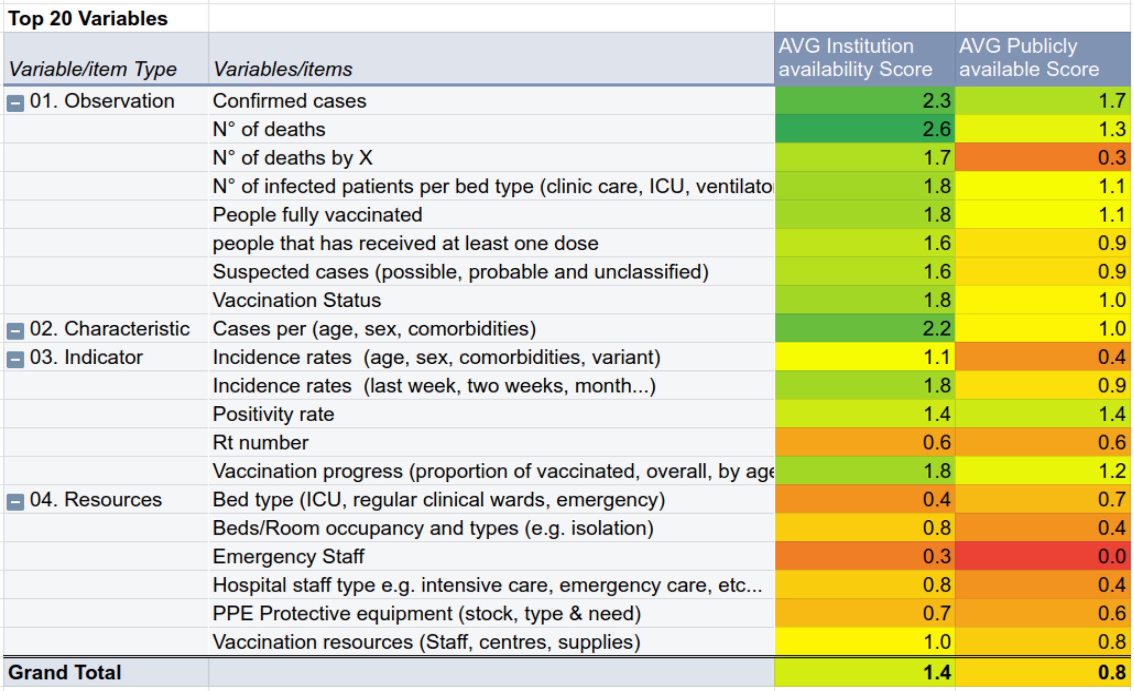
■ The most important variables as per end users

› The most important 10 variables



Top 10 Variables			
Variable/item Type	*Variables/items*	AVG Institution availability Score	AVG Publicly available Score
⊟ 01. Observation	Confirmed cases	2.3	1.7
	N° of infected patients per bed type (clinic care, ICU, ventilato	1.8	1.1
	People fully vaccinated	1.8	1.1
⊟ 02. Characteristic	Cases per (age, sex, comorbidities)	2.2	1.0
⊟ 03. Indicator	Incidence rates (age, sex, comorbidities, variant)	1.1	0.4
	Incidence rates (last week, two weeks, month...)	1.8	0.9
	Positivity rate	1.4	1.4
	Rt number	0.6	0.6
⊟ 04. Resources	Emergency Staff	0.3	0.0
	Vaccination resources (Staff, centres, supplies)	1.0	0.8
Grand Total		**1.4**	**0.9**

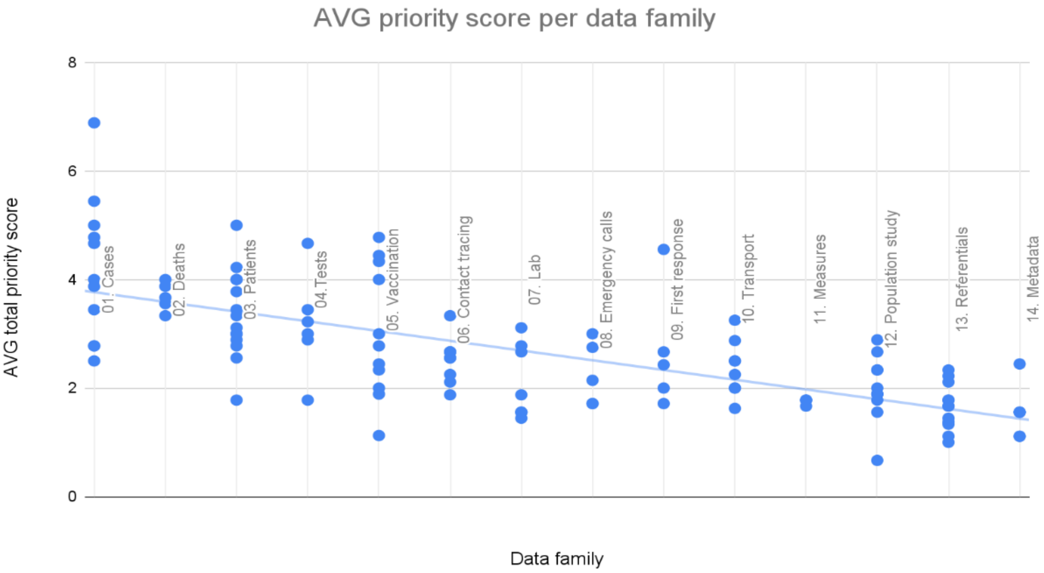
› The most important 20 variables



Top 20 Variables			
Variable/item Type	*Variables/items*	AVG Institution availability Score	AVG Publicly available Score
⊟ 01. Observation	Confirmed cases	2.3	1.7
	N° of deaths	2.6	1.3
	N° of deaths by X	1.7	0.3
	N° of infected patients per bed type (clinic care, ICU, ventilato	1.8	1.1
	People fully vaccinated	1.8	1.1
	people that has received at least one dose	1.6	0.9
	Suspected cases (possible, probable and unclassified)	1.6	0.9
	Vaccination Status	1.8	1.0
⊟ 02. Characteristic	Cases per (age, sex, comorbidities)	2.2	1.0
⊟ 03. Indicator	Incidence rates (age, sex, comorbidities, variant)	1.1	0.4
	Incidence rates (last week, two weeks, month...)	1.8	0.9
	Positivity rate	1.4	1.4
	Rt number	0.6	0.6
	Vaccination progress (proportion of vaccinated, overall, by age	1.8	1.2
⊟ 04. Resources	Bed type (ICU, regular clinical wards, emergency)	0.4	0.7
	Beds/Room occupancy and types (e.g. isolation)	0.8	0.4
	Emergency Staff	0.3	0.0
	Hospital staff type e.g. intensive care, emergency care, etc...	0.8	0.4
	PPE Protective equipment (stock, type & need)	0.7	0.6
	Vaccination resources (Staff, centres, supplies)	1.0	0.8
Grand Total		**1.4**	**0.8**

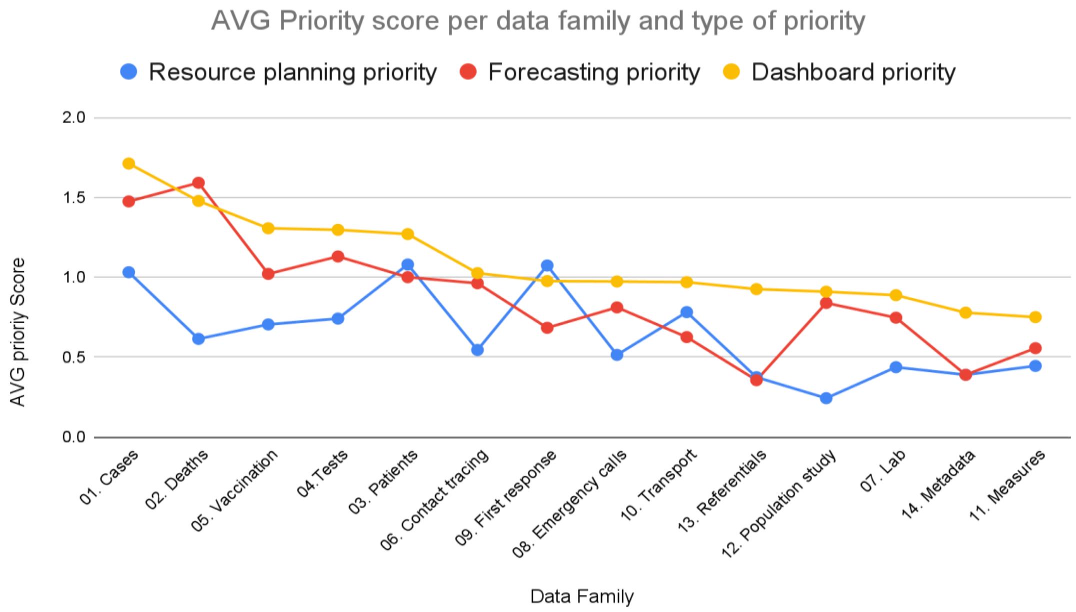
■ The most important data families

› The most relevant data families are: Cases, deaths, patients, tests and vaccination.



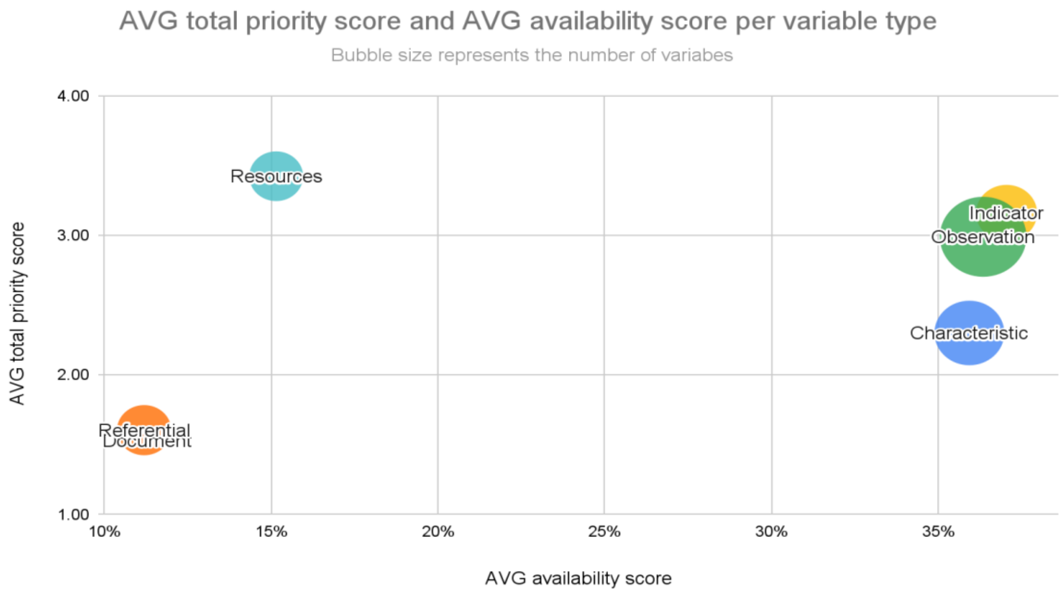
- The most important PANDEM-2 features

› The most important priority are dashboards, with strong differences between data families.



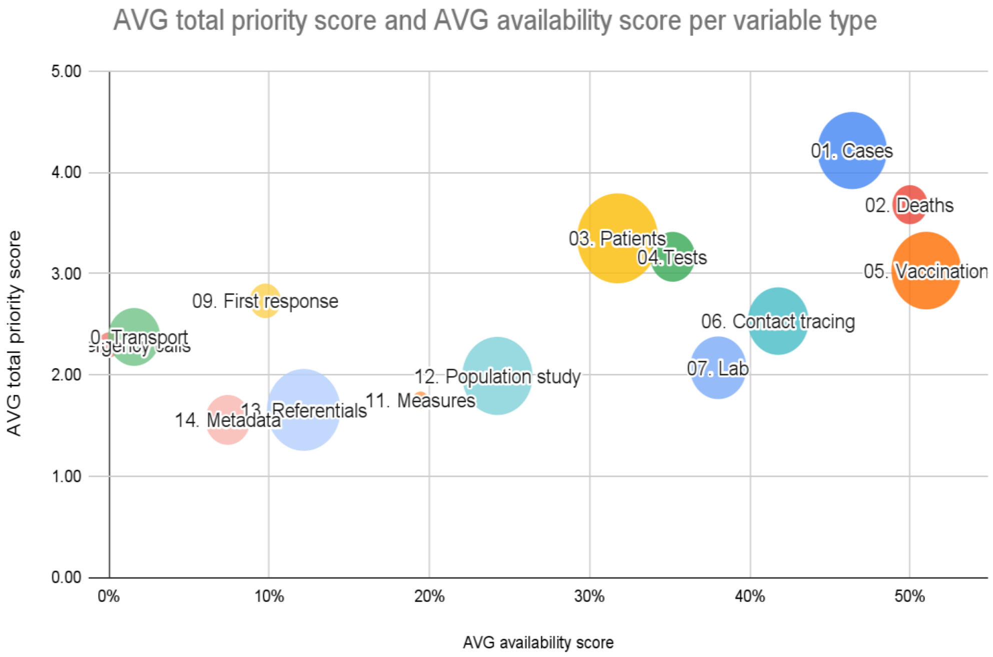
- Some data is not available

› There is a lack of knowledge about resource availability and yet they are a big priority.



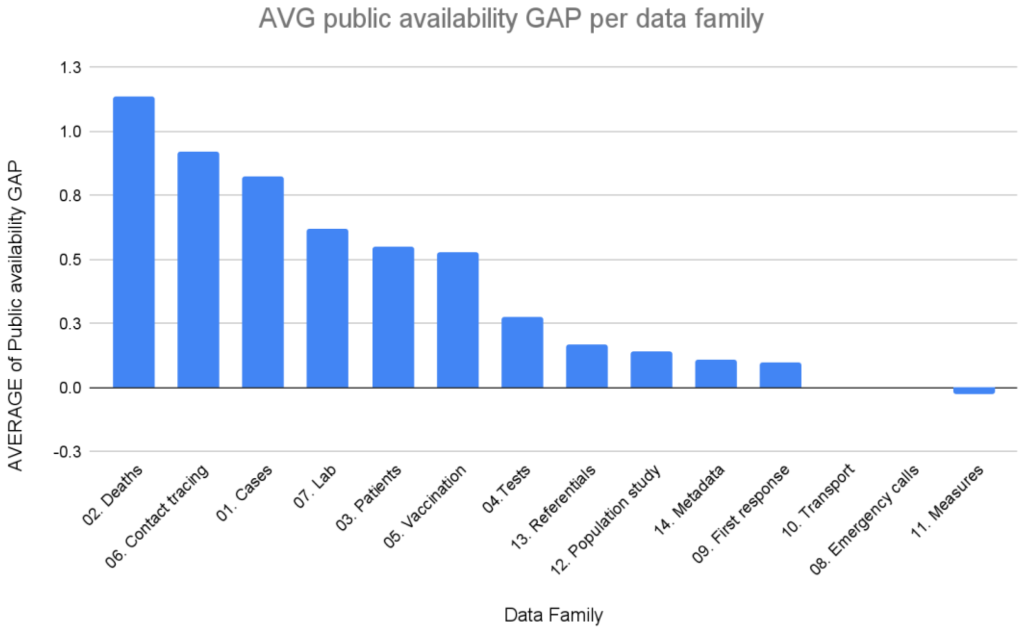
- Data exists for the most important data families

› The most important data families are also the most available.



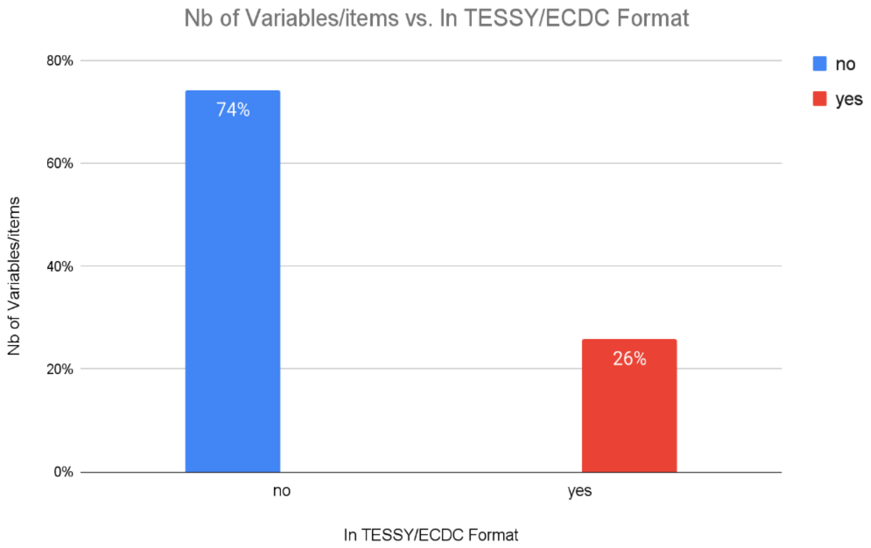
- The main gaps between availability and importance

› The most important GAP in public data availability are contact tracing and deaths



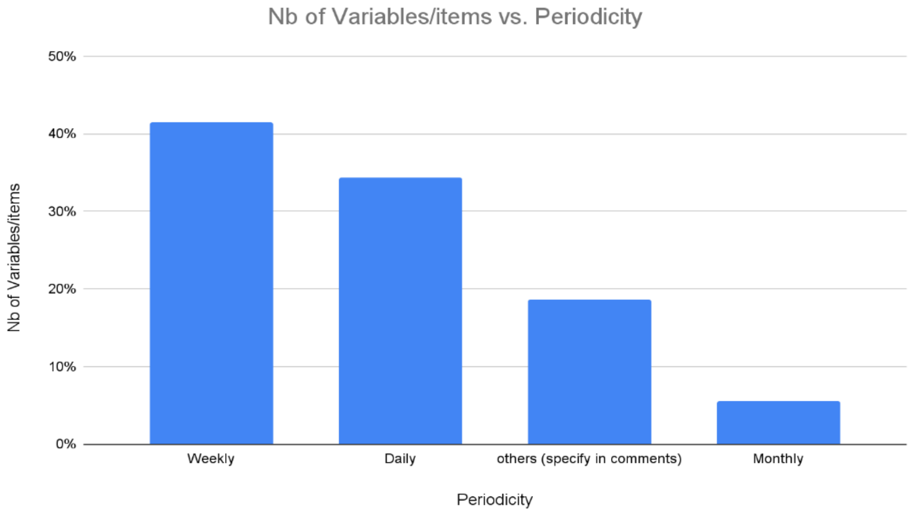
- Is existing data available in TESSY format

› 74% of data is not on TESSY format (probably underestimated)



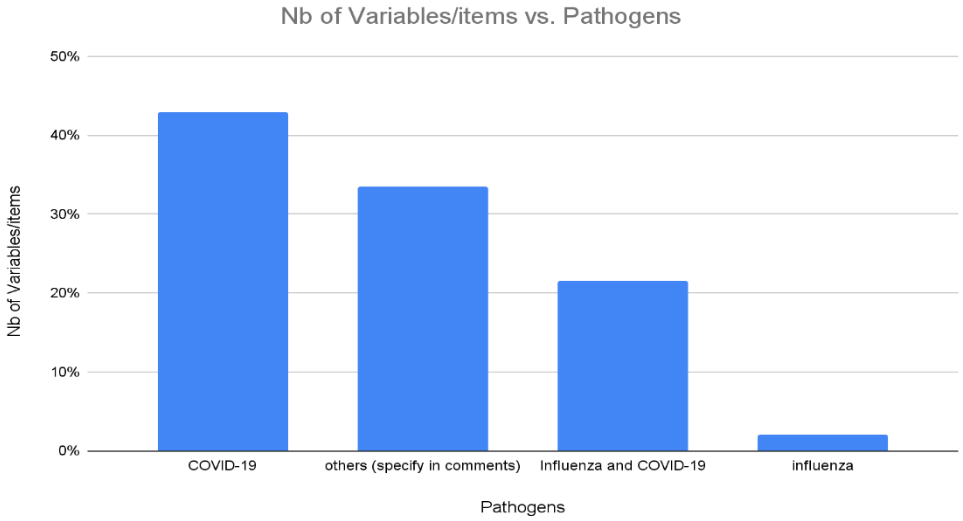
- The preferred reporting period

› Data collection is mainly done on a weekly and daily basis



- Estimating the part of variable only available for COVID-19

› Most of the data is only available for COVID (with big nuances)



6.3.2 Data survey conclusions

The conclusions of the data survey show us that the priority and availability of variables is very heterogeneous even within data families. Many variables are publicly domain e.g. cases, deaths, tests or vaccination but others are hard to estimate even for public health agencies. Human & material resource availability is an example of this since these figures are hard to collect and can be adapted on the field depending on the responder's organisation.

Despite this variability, some clear trends in terms of priority and availability were observed for different data families on the survey allowing us to classify data families on the following groups

- Important and available (Cases, Deaths, vaccination, patient, tests, Lab)

Important variables for pandemic response and most of them are publicly reported. Often the user requirements request a level of granularity which is not publicly available but the data exists and could potentially be (or is) shared between member states.

- Important with limited availability (NGS, Contact tracing, Population study)

Variables on this group are also important for pandemic response but in most cases they are only available locally and not public domain. This data is mainly collected at individual level and contains personal information. Nevertheless we estimate that the main issue for data sharing is the lack of standard aggregated indicators that could be easily shared.

- Lack of data or low priority: First response (including staff), transport, referentials, measures (including flights), emergency calls, resources

Most variables in this group are not systematically collected or not important for pandemic response. Even though some of these are important for end users it would not be realistic to expect data sharing in the short term.

- Not in requirements: Weather, seroprevalence

Variables on this group were not mentioned during the requirement gathering process so they were considered as low priority.

The groups listed here above were used to define the final refinement of data sources shown in Appendix 6.8 Final list of sources.

6.4 List of initial data sources identified for PANDEM-2

The list below indicates the identified data sources during the initial assessment in the PANDEM-2 project. These are the data sources needed to meet end-user requirements and which we expect will be needed to collect the final list of indicators, to compute and input them into the software in D2.6. We use the terminology "provided by end-users" meaning that data will be obtained or generated by the partners in the PANDEM-2. We will use a data collection template to collect aggregated data from users in the consortium. All imported data will follow the same annotation and processing pipeline described in section 6.6. This applies both for data obtained from public sources and data uploaded by end users in their local installations of PANDEM-2.

Data being publicly available covers many of the data requirements but often at an insufficient level of geographical detail or with some missing variables. To mitigate this risk if data is not available for end users then synthetic data will be used.

The following sources include sources identified and described in the present document, some of these sources are relevant not only directly for our end users to be displayed in the dashboard but also for modelling (e.g., flights, weather or data coming from animal or human alerts platforms). Finally, we have also included data sources developed in any task within WP2 such as Lab data (including NGS-metagenomics task 2.5), participatory surveillance data (Influenzanet task 2.2) or data coming from social media (task 2.3) which has been described as relevant for pandemic management.

a. Surveillance - traditional sources:
    a1. TESSY[1]
    a2. Influenza net[2]
    a3. European mortality monitoring activity (EuroMOMO)[3]
    a4. WHO-Flunet[4] and GISAID[5]

[1] https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy

[2] http://influenzanet.info/#page/info

[3] https://www.euromomo.eu/

[4] https://www.who.int/tools/flunet

a5. Bioscomsc[6]
a6. GO.DATA[7] – API, Contact Tracing data provided by end users' partners
a7. Lab Surveillance and monitoring: (NGS - WGS - RT - PCR) data provided by end-users and data coming from publicly available repositories e.g., ENA[8], NCBI[9] , TESSY - lab data.
a8. Surveillance data provided by end-users.

b. Surveillance - non-traditional sources:
b1. Early detection & monitoring (Human) (RSS): PROMED[10], Health Map[11], European Media Monitor-MEDSyS[12], GPHIN[13], WHO-EIOS[14].
b2. Early detection & monitoring (Animal) (RSS): EMPRES-i FAO[15], ECDC-VECTOR NET[16]
b3. Social Media: Integrate Social Media Analysis (SMA) algorithms from NUIG, Epitweetr[17] (which uses twitter API), Reddit[18] API, Signal detection[19]

c. Surveillance – System capacity
c1. Staff (provided by end-users, Eurostat[20], TESSY)
c2. Resources (provided by end-users, Eurostat, TESSY)

d. Weather & climate
d1. Weather monitoring (NOAA[21], MODIS-NASA[22])
d2. Disasters (OCHA-relief web[23])

e. Mobility
e1. Flights (public APIs, e.g., OpenSky Network[24], OpenFlights[25], Flight radar 24[26])

f. Countermeasures - Communications

[5] https://www.gisaid.org/
[6] https://biocomsc.upc.edu/en/covid-19
[7] https://www.who.int/tools/godata/about
[8] https://www.ebi.ac.uk/ena/portal/api/
[9] https://www.ncbi.nlm.nih.gov/nucleotide/
[10] https://promedmail.org/
[11] https://healthmap.org/en/
[12] https://medisys.newsbrief.eu/medisys/homeedition/en/home.html
[13] https://gphin.canada.ca/cepr/aboutgphin-rmispenbref.jsp?language=en_CA
[14] https://www.who.int/initiatives/eios
[15] https://empres-i.review.fao.org/#/
[16] https://www.ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/vector-net
[17] https://www.ecdc.europa.eu/en/publications-data/epitweetr-tool
[18] https://www.reddit.com/
[19] https://www.ecdc.europa.eu/en/publications-data/episignaldetection-tool
[20] https://ec.europa.eu/eurostat
[21] https://www.noaa.gov/weather
[22] https://modis.gsfc.nasa.gov/data/
[23] https://reliefweb.int/disasters
[24] https://opensky-network.org/
[25] https://openflights.org/
[26] https://www.flightradar24.com/40.43,-4.05/6

f1. Social media interventions (provided by end-user partners, ECDC, Epitweetr)
f2. Non-pharmaceutical interventions (provided by end-user partners, ECDC, Epitweetr)
f3. Vaccines roll-outs (provided by end-user partners, ECDC, Epitweetr)

g. Repositories
g1. Research (updated scientific publications of interest & ongoing initiatives)
g2. Training (Training material & scenarios)
g3. Policy (static content on Dashboards)

h. Denominators (Eurostat and provided by end-user partners)

6.5 FAIR principles

The data management plan (DMP, deliverable 2.1) guides the way data is managed in order to meet the project objectives, ensuring FAIR principles[27](how data is made findable, accessible, interoperable and reusable) are followed as recommended by the European commission and many other organizations. In addition, it ensures all partners are cognisant of Ethical issues and European regulations (e.g., the General Data Protection Regulation or GDPR) that need to be addressed. PANDEM-2 will produce tools for pandemic management and preparedness and response, based on existing data.

The datasets/data sources identified in WP2 will consist of data coming from publicly available information online and also data from national or international organizations. During development, data will be produced as the result of data collection, standardisation, synthetic data generation and public data provided from end users,

The scientific interest of the PANDEM-2 database needs to be seen from a holistic perspective and will increase as more sources are integrated in a coherent way.

We will make the datasets, and PANDEM-2 generated metadata, discoverable for other researchers by publishing them in different data repository platforms such as health Zenodo[28], data portal - PHIRI[29] or other European initiatives. These datasets and metadata will receive a unique DOI (Digital Object Identifier), which will make the datasets detectable for others.

As mentioned in the DMP, in order to generate FAIR metadata, for each data set metadata will be created containing the following information:

[27] https://www.go-fair.org/fair-principles/
[28] Zenodo: https://zenodo.org/
[29] health data portal - PHIRI: https://www.healthinformationportal.eu/health-information-portal

- Title: name of the dataset (naming conventions will be taken from TESSY when possible)
- Unique identifier: DOI (A DOI will be provided for resources with scientific interest, such as predictions, metadata collections or synthetic data generated within PANDEM-2).
- Creators: names and addresses of the people and organisations who created or contributed to the data collection.
- Date of creation: key dates associated with the data, including project start and end date, and the period covered by the data.
- Keywords: keywords.
- Funder: the organisation that funded the research.
- Rights: the intellectual property rights held in the data.
- Language: language(s) of the intellectual content of the resource.
- Location: information on the spatial coverage of the data.
- Method: information on how the data was generated, including equipment or software
  used.
- Data availability, conditions for access to the data or for rebuilding the dataset.

In order to ensure FAIR compliance as the project evolves, we have defined the PANDEM-2 data labelling schema (DLS) (section 6.6.1). This process will be systematically applied when importing data into the PANDEM-2 database and will drive how data is stored. The PANDEM-2 DLS guarantees that a detailed and updated documentation of all present variables and indicators is always kept and can be seen by users at any time or shared with any meta data initiative.

The detailed profile of variables will always be updated as a JSON file derived from the list presented in section 6.2. The file will describe all variables in a canonical structure including description and the linked source dataset. The link with source datasets will facilitate FAIR compliance because it will explicitly indicate each variable provenance.

A theoretical output for the list of variables prepared for FAIR compliance is shown in section 6.6.1. as well as the explanation on the construction method. The full file can be seen on the link here below:
https://github.com/pandem2/pandem-source/blob/main/pandem2source/data/DLS/variables.json

6.6. Data sources integration process

6.6.1. PANDEM-2 Data Labelling Schema (DLS)

- <u>A. Hypothesis & definitions</u>

The DLS make the following assumptions which determine the variable profiles in PANDEM-2:

- All pandemic related data can be seen as a set of variables containing:
    - A **label** indicating general variable name e.g., number of cases, Incidence rate, symptom name, etc. Names are previously defined and are associated with a particular definition, the type (numeric, text), and calculation method.
    - A **value** for the observation that could be a quantity, a text or a date.
    - A **reporting user** indicating the PANDEM-2 user that is responsible for updating this data periodically
    - A **named source** indicating a particular file that is recurrently uploaded by the reporting user. It could also be a table, a website or a git repository where this data was obtained from.
    - Variables have a **class** among *observations* and *attributes*. Observations contain mainly epidemiological information like "number of cases" or "Incidence rate". Attributes provide extra information associated with a particular observation e.g., the age group of the observed cases. This class attribute relates with the classification described in section 6.2 as follows:
        - Observation
            - Observation
            - Indicator
        - Attributes
            - Characteristic
            - Referential
            - Resources
        - Out of scope
            - Document
    - Variables have a data type: integer, numeric, date, datetime, string.
- Variables are tied together in valued tuples[30] e.g. ("confirmed cases":13, "pathogen":"dengue", "age-group":"10-18", "date":"2021-12-13", source:"DE-by-MUN-es").
- Each tuple has to be self-describing. Which means that the storage location on the database depends exclusively on the variables present in the tuple. More information on this is in section 6.6.1.C1.

[30] A tuple is a finite ordered list (sequence) of elements. In DLS we assume that each element on tuples contains a value that is associated to a well known variable.

- B. Design goals
  - **Make data load easy**: For ease of adoption and integration, PANDEM-2 should be able to read existing sources with no or minimal modification. Adding a new source will not need any code change if the reference variables and acquisition channel are already supported.
  - **Data load on WP3 can be done in small iterations**
  - **Keep a trace of data**: Data shown in PANDEM-2 reports can be drill-through to tuples. This allows tracing back the origin of the data, the reporting institution (reported in the data source definition) and methodology applied (specified in the list of variables).
  - **Make abstraction of source format**: By transforming data sources into a canonical structure, the PANDEM-2 database can be designed independently of the format provided by users. User input files can have their own naming convention for column names or attributes. The importing process will translate them into a well-known and coherent format.
  - **Independent of database schema**: By transforming data sources into a canonical structure, the PANDEM-2 database design can evolve without impacting data collection procedures. The format sent to the database is a list of tuples acting as units of data that can be unequivocally identified using the defined observations and attributes. This simple and constant format for input data will allow the database to make any design choice.
  - **Define data Standardisation & validation:** At the moment of adding a new source, a technical PANDEM-2 installation administrator must define how each field would be standardized and validated. This will be done by creating the necessary mapping files (manually or using an interface). After this, the new source will be automatically acquired and integrated into the PANDEM-2 database.
    - **Standardisation**: Standardisation is the process of changing variable names and values to a PANDEM-2 format. In the case of repositories like countries PANDEM-2 will adopt a well-known standard reference like ISO-3166.2 if a data source references countries by name, and the DLS will define that a translation from labels to codes have to be performed. Many translations could exist and each one will have a unique name and need to be provided if it does not yet exist at the moment of integrating files.
    - **Validation**: Validation is automatically done in terms of predefined data type and matches found in the case of standardisation. If any error arises during this process the responsible user will be informed.
  - **Simplify FAIR compliance**: Since each input field needs to be associated with a standard and common dictionary using international standards, data imported to PANDEM-2 will be easily made FAIR compliant.
  - **Facilitate multi-tenancy**: By allowing file heterogeneity while keeping a clear trace of data origin. PANDEM-2 could easily isolate different tenants on the same database. It is yet to be defined if PANDEM-2 will support multi-tenancy, but the DLS should facilitate this.

- C. DLS implementation:

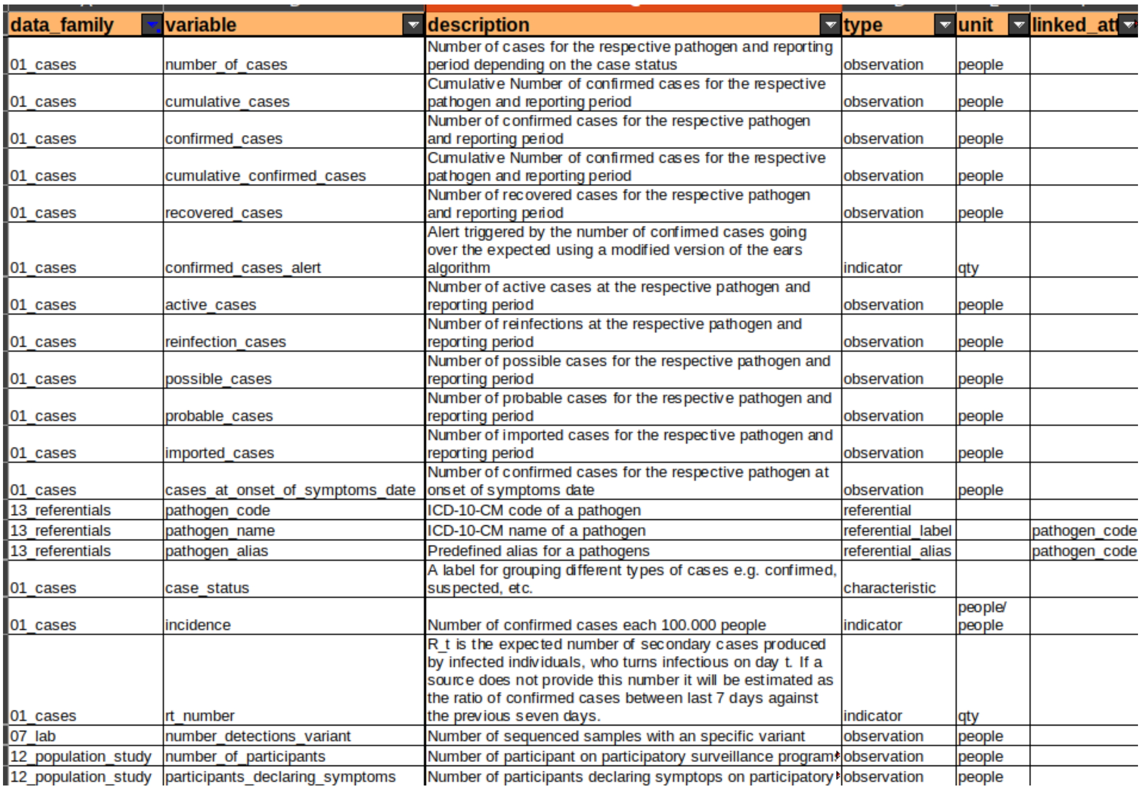
C1. DLS implementation: Reference variables

PANDEM-2 will contain an updated list of variables[31] representing all possible data that can be imported and shown in reports. For each variable the following elements will be identified:

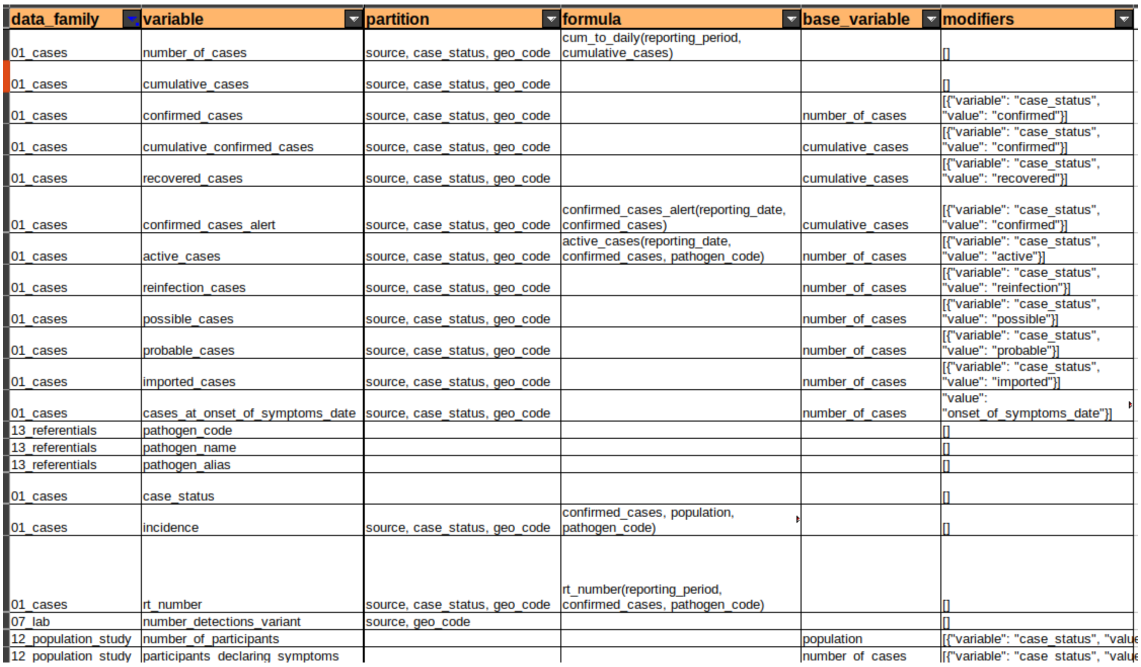
- Data family: A category used for grouping variable from a functional perspective
- Variable Name: The variable name
- Description: The functional description of the variable
- Datasets: The input datasets that are currently filling this variable
- Type: Whether the variable is an observation, an attribute an indicator or a resource
- Unit of measure: The unit of the variable e.g. People or Kilograms
- Linked Attributes: All of the expected attributes that can be linked to the variable if an observation or indicator. This information will allow the validation of user input and provide guidance for database design.
- Aliases: This element will explicitly indicate when a variable is a modification of another by changing an attribute. e.g., “number of cases” can be derived on “number of confirmed cases” and “number of suspected cases”. From a functional point of view, they are different variables, but from model design it is the same variable (Number of cases) with a different value on the attribute ‘Case Type”. Each alias is composed of
  - Alias’s name
  - Base variable
  - List of modified attributes as a list of variable and values pairs

Here below is an example of the list of variables for three data families (Cases, Variants and population study (participatory surveillance). The full file can be seen on the link here below https://github.com/pandem2/pandem-source/blob/main/pandemsource/data/list-of-variables.csv

[31] https://github.com/pandem2/pandem-source/blob/main/pandem2source/data/DLS/variables.json



data_family	variable	description	type	unit	linked_att
01_cases	number_of_cases	Number of cases for the respective pathogen and reporting period depending on the case status	observation	people	
01_cases	cumulative_cases	Cumulative Number of confirmed cases for the respective pathogen and reporting period	observation	people	
01_cases	confirmed_cases	Number of confirmed cases for the respective pathogen and reporting period	observation	people	
01_cases	cumulative_confirmed_cases	Cumulative Number of confirmed cases for the respective pathogen and reporting period	observation	people	
01_cases	recovered_cases	Number of recovered cases for the respective pathogen and reporting period	observation	people	
01_cases	confirmed_cases_alert	Alert triggered by the number of confirmed cases going over the expected using a modified version of the ears algorithm	indicator	qty	
01_cases	active_cases	Number of active cases at the respective pathogen and reporting period	observation	people	
01_cases	reinfection_cases	Number of reinfections at the respective pathogen and reporting period	observation	people	
01_cases	possible_cases	Number of possible cases for the respective pathogen and reporting period	observation	people	
01_cases	probable_cases	Number of probable cases for the respective pathogen and reporting period	observation	people	
01_cases	imported_cases	Number of imported cases for the respective pathogen and reporting period	observation	people	
01_cases	cases_at_onset_of_symptoms_date	Number of confirmed cases for the respective pathogen at onset of symptoms date	observation	people	
13_referentials	pathogen_code	ICD-10-CM code of a pathogen	referential		
13_referentials	pathogen_name	ICD-10-CM name of a pathogen	referential_label		pathogen_code
13_referentials	pathogen_alias	Predefined alias for a pathogens	referential_alias		pathogen_code
01_cases	case_status	A label for grouping different types of cases e.g. confirmed, suspected, etc.	characteristic		
01_cases	incidence	Number of confirmed cases each 100.000 people	indicator	people/ people	
01_cases	rt_number	R_t is the expected number of secondary cases produced by infected individuals, who turns infectious on day t. If a source does not provide this number it will be estimated as the ratio of confirmed cases between last 7 days against the previous seven days.	indicator	qty	
07_lab	number_detections_variant	Number of sequenced samples with an specific variant	observation	people	
12_population_study	number_of_participants	Number of participant on participatory surveillance program	observation	people	
12_population_study	participants_declaring_symptoms	Number of participants declaring symptops on participatory	observation	people	



data_family	variable	partition	formula	base_variable	modifiers
01_cases	number_of_cases	source, case_status, geo_code	cum_to_daily(reporting_period, cumulative_cases)		[]
01_cases	cumulative_cases	source, case_status, geo_code			[]
01_cases	confirmed_cases	source, case_status, geo_code		number_of_cases	[{"variable": "case_status", "value": "confirmed"}]
01_cases	cumulative_confirmed_cases	source, case_status, geo_code		cumulative_cases	[{"variable": "case_status", "value": "confirmed"}]
01_cases	recovered_cases	source, case_status, geo_code		cumulative_cases	[{"variable": "case_status", "value": "recovered"}]
01_cases	confirmed_cases_alert	source, case_status, geo_code	confirmed_cases_alert(reporting_date, confirmed_cases)	cumulative_cases	[{"variable": "case_status", "value": "confirmed"}]
01_cases	active_cases	source, case_status, geo_code	active_cases(reporting_date, confirmed_cases, pathogen_code)	number_of_cases	[{"variable": "case_status", "value": "active"}]
01_cases	reinfection_cases	source, case_status, geo_code		number_of_cases	[{"variable": "case_status", "value": "reinfection"}]
01_cases	possible_cases	source, case_status, geo_code		number_of_cases	[{"variable": "case_status", "value": "possible"}]
01_cases	probable_cases	source, case_status, geo_code		number_of_cases	[{"variable": "case_status", "value": "probable"}]
01_cases	imported_cases	source, case_status, geo_code		number_of_cases	[{"variable": "case_status", "value": "imported"}]
01_cases	cases_at_onset_of_symptoms_date	source, case_status, geo_code		number_of_cases	"value": "onset_of_symptoms_date"}]
13_referentials	pathogen_code				[]
13_referentials	pathogen_name				[]
13_referentials	pathogen_alias				[]
01_cases	case_status				[]
01_cases	incidence	source, case_status, geo_code	confirmed_cases, population, pathogen_code)		[]
01_cases	rt_number	source, case_status, geo_code	rt_number(reporting_period, confirmed_cases, pathogen_code)		[]
07_lab	number_detections_variant	source, geo_code			[]
12_population_study	number_of_participants			population	[{"variable": "case_status", "value
12_population_study	participants_declaring_symptoms			number_of_cases	[{"variable": "case_status", "value

C2. DLS implementation: Mapping sources

Mapping sources to the list of variables is one of the main goals of the DLS. The DLS provides a way to formalize the structure and content of source datasets as a contract. Each data source definition is a contract that can be understood by the PANDEM-2 data import process and act as an engagement of what data is expected to be delivered and on which exact perimeter e.g., weekly Influenza confirmed cases for Italy by City and age group. This approach will enable greater flexibility when it comes to importing data from different formats and sources, since the only action necessary to register a new source will be to write a source definition file and potentially to provide referential mappings (assuming that all variables are already known).

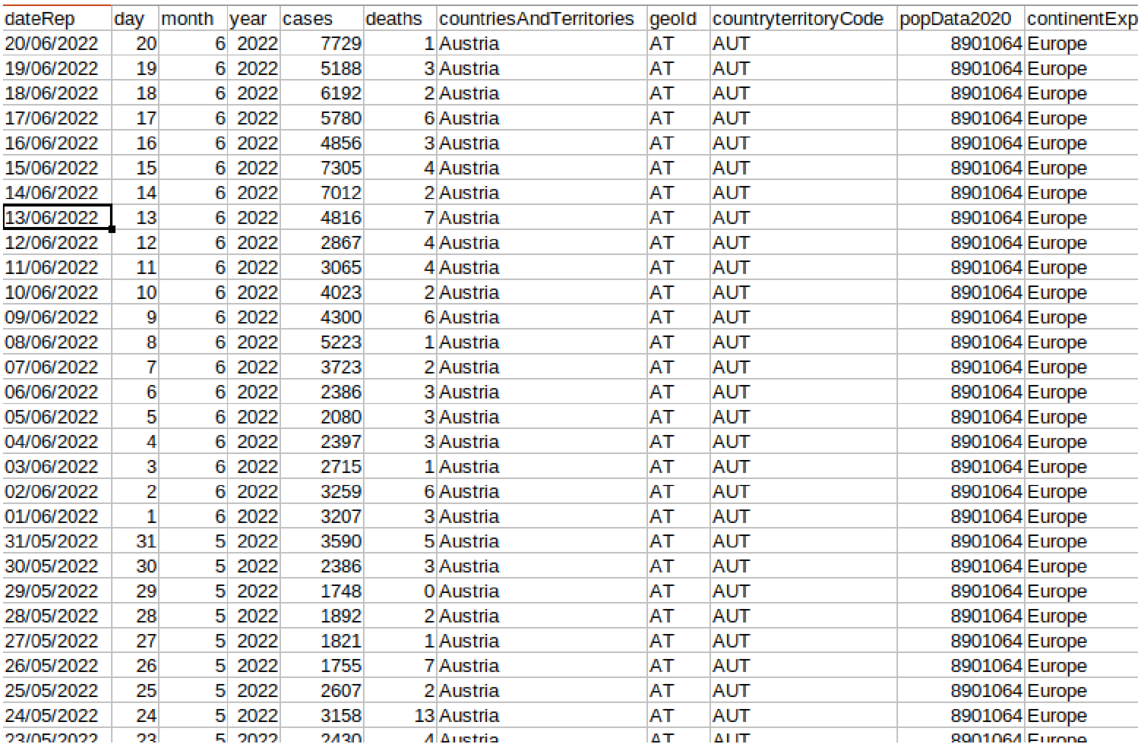
The source definition uses the following information:

- Scope: Indicates the global information of the dataset (common to all rows).
  - Source name: The name of the source which needs to be unique among sources (For open data sources this will correspond to headings in appendix 6.7)
  - Source description: A description of the source
  - Tags: A list of keywords that are associated with this dataset. These tags allow the reports to know which dataset can be merged and those which cannot. To understand how this could work let's assume the following scenario:
    - PANDEM-2 integrates data from COVID-19 confirmed daily cases from Johns Hopkins University (JHU) by country at world level, from ECDC by country at European level, and from 3 individual countries (Romania, France and Belgium) by municipality.
    - The tags are designed to provide flexibility at reporting level, the JHU Dataset could have the tags ["World"], the ECDC as ["European"] and the countries as ["EU National"]
    - Different reports could choose to display data using different tags. This allows the user (or report designer) to choose the scope of data without worrying about the underlying sources and using the same database queries.
    - Tags are expected to change over time depending on the reporting needs (by changing the source definition file).
  - Frequency: The expected frequency of the dataset refresh. It could be daily, weekly or real time.
  - Reference User: A person or institution being responsible for the correctness and timeliness of the source contents. This is the user that will receive any alert and will be responsible for fixing potential issues.
  - Reporting email: The email to contact in case of issues with the dataset
  - Globals: A list of context modifiers. They will define attributes that will apply to all data contained on the dataset e.g., the pathogen, or a particular location. The value a literal value or a reference to a particular dictionary.
  - Update scope: This indicates what existing data is expected to be replaced each time the dataset is uploaded. It can be a fixed value or a value coming from a dataset.
- Acquisition: Information on how to obtain this dataset
  - Chanel: The method of acquisition, it could be:

        - URL of a file to download (integration would be triggered automatically on file change based on ETAG or pooling)
        - Git repository with many files to integrate (integration would be triggered automatically after each commit.
        - Local folder with many files to integrate (integration would be triggered automatically when new files arrive or on file change)
        - PANDEM2-API: PANDEM-2 will also provide a REST API for receiving files on demand.
        - RSS URL providing a news feed to integrate (integration would be automatically triggered when new articles arrive
        - Custom channel: Channels requiring custom development
            - EPitweetr
            - NCBI
        - Other connectors may be developed as needed.
    - Format: Formatting options for the dataset if not already typed. Some example values here below
        - CSV:
            - decimal_sign
            - thousands_separator
            - date_format
            - encoding
            - field_separator
            - quoting_character
        - Excel
            - spreadsheet_name
- List of columns: List of expected columns on the dataset, including for each one
    - Name of the column
    - Associated variable
    - Possible attribute modifiers.
    - Standardisation instructions, like the reference repository or translation to apply into data.

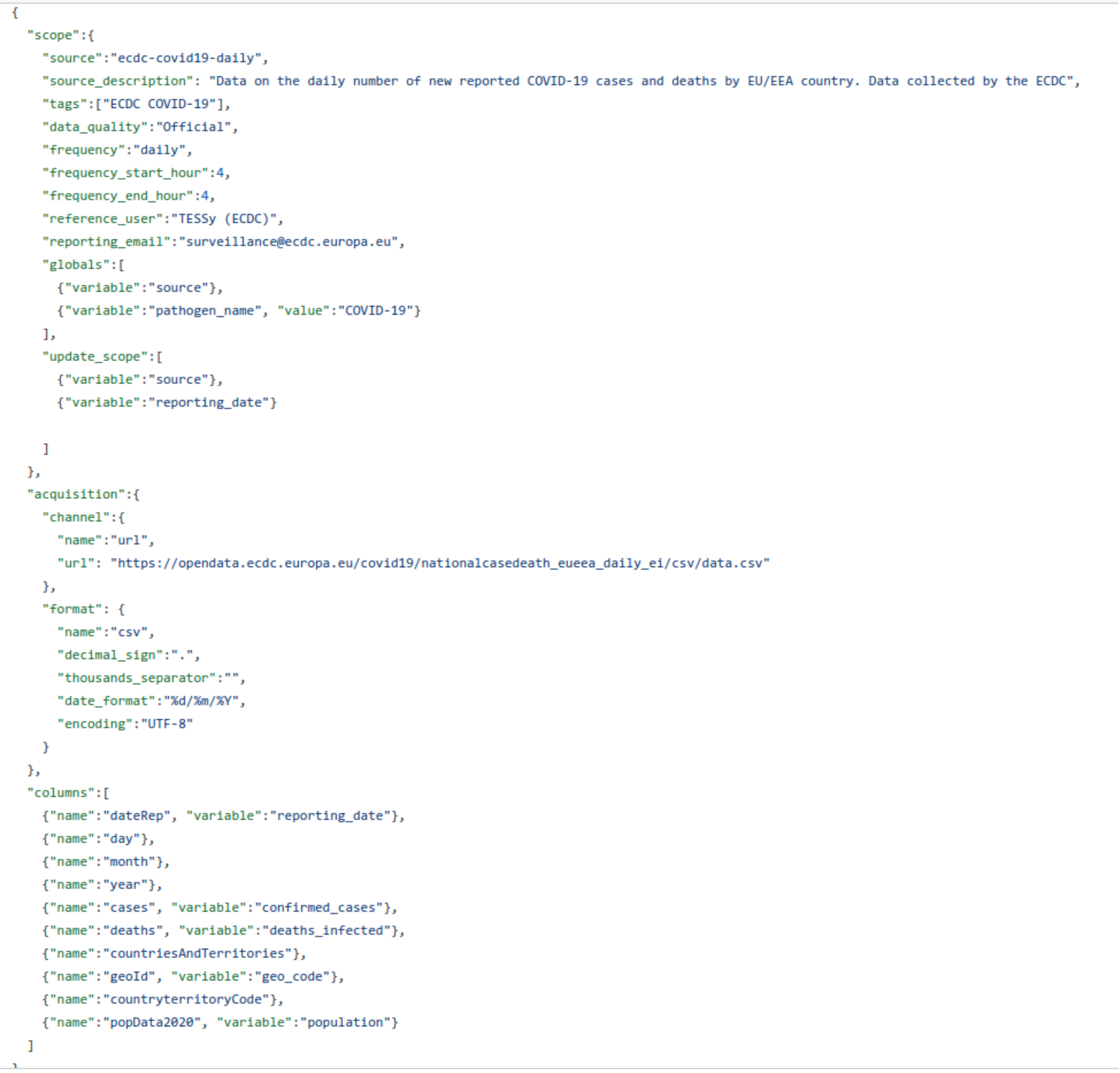
In order to exemplify the data source definition of some real sources we will show the ones developed for integrating ECDC COVID-19 cases, ECDC COVID-19 variants and ECDC influenza.net

- ECDC COVID-19 cases file structure



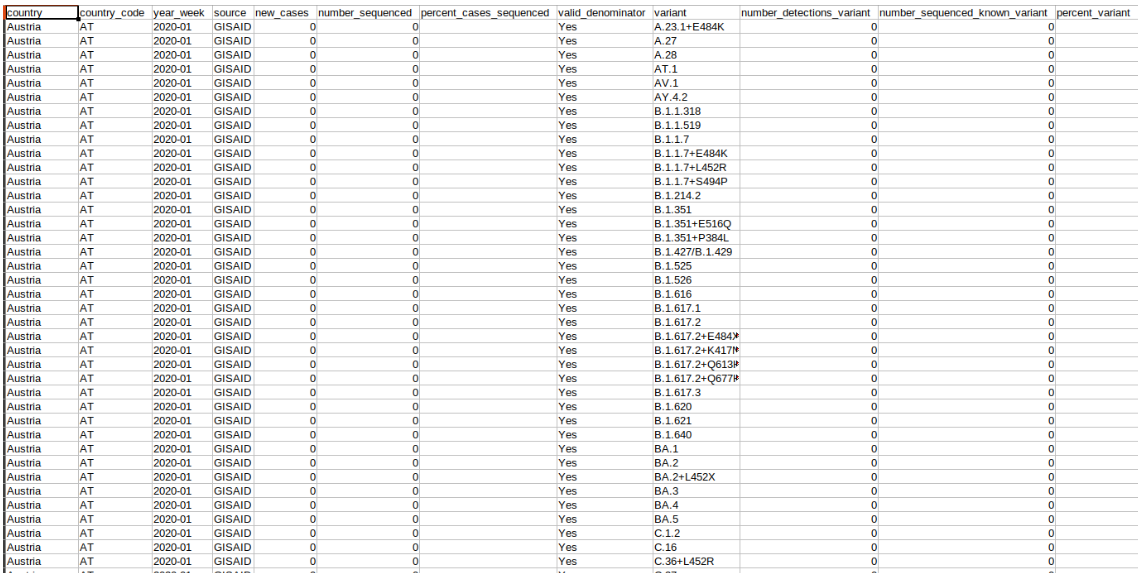
dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoId	countryterritoryCode	popData2020	continentExp
20/06/2022	20	6	2022	7729	1	Austria	AT	AUT	8901064	Europe
19/06/2022	19	6	2022	5188	3	Austria	AT	AUT	8901064	Europe
18/06/2022	18	6	2022	6192	2	Austria	AT	AUT	8901064	Europe
17/06/2022	17	6	2022	5780	6	Austria	AT	AUT	8901064	Europe
16/06/2022	16	6	2022	4856	3	Austria	AT	AUT	8901064	Europe
15/06/2022	15	6	2022	7305	4	Austria	AT	AUT	8901064	Europe
14/06/2022	14	6	2022	7012	2	Austria	AT	AUT	8901064	Europe
13/06/2022	13	6	2022	4816	7	Austria	AT	AUT	8901064	Europe
12/06/2022	12	6	2022	2867	4	Austria	AT	AUT	8901064	Europe
11/06/2022	11	6	2022	3065	4	Austria	AT	AUT	8901064	Europe
10/06/2022	10	6	2022	4023	2	Austria	AT	AUT	8901064	Europe
09/06/2022	9	6	2022	4300	6	Austria	AT	AUT	8901064	Europe
08/06/2022	8	6	2022	5223	1	Austria	AT	AUT	8901064	Europe
07/06/2022	7	6	2022	3723	2	Austria	AT	AUT	8901064	Europe
06/06/2022	6	6	2022	2386	3	Austria	AT	AUT	8901064	Europe
05/06/2022	5	6	2022	2080	3	Austria	AT	AUT	8901064	Europe
04/06/2022	4	6	2022	2397	3	Austria	AT	AUT	8901064	Europe
03/06/2022	3	6	2022	2715	1	Austria	AT	AUT	8901064	Europe
02/06/2022	2	6	2022	3259	6	Austria	AT	AUT	8901064	Europe
01/06/2022	1	6	2022	3207	3	Austria	AT	AUT	8901064	Europe
31/05/2022	31	5	2022	3590	5	Austria	AT	AUT	8901064	Europe
30/05/2022	30	5	2022	2386	3	Austria	AT	AUT	8901064	Europe
29/05/2022	29	5	2022	1748	0	Austria	AT	AUT	8901064	Europe
28/05/2022	28	5	2022	1892	2	Austria	AT	AUT	8901064	Europe
27/05/2022	27	5	2022	1821	1	Austria	AT	AUT	8901064	Europe
26/05/2022	26	5	2022	1755	7	Austria	AT	AUT	8901064	Europe
25/05/2022	25	5	2022	2607	2	Austria	AT	AUT	8901064	Europe
24/05/2022	24	5	2022	3158	13	Austria	AT	AUT	8901064	Europe
23/05/2022	23	5	2022	2430	4	Austria	AT	AUT	8901064	Europe

- ECDC COVID-19 cases DLS



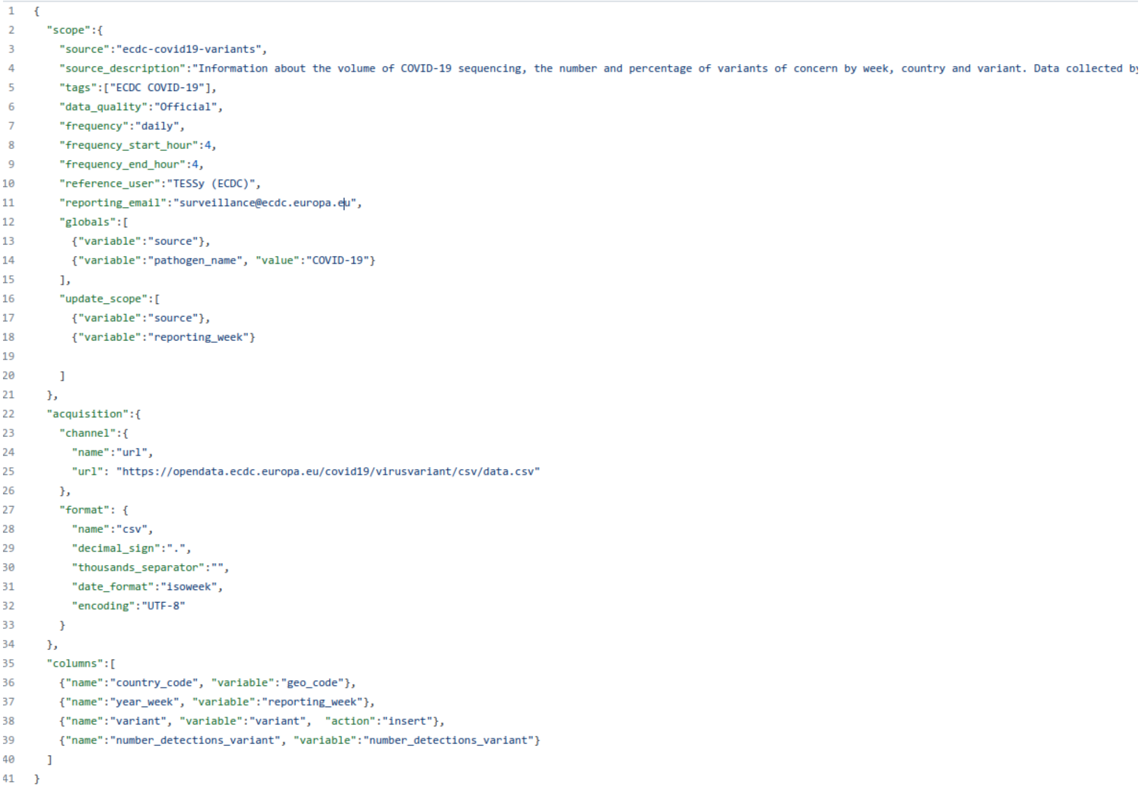
```
{
  "scope":{
    "source":"ecdc-covid19-daily",
    "source_description": "Data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country. Data collected by the ECDC",
    "tags":["ECDC COVID-19"],
    "data_quality":"Official",
    "frequency":"daily",
    "frequency_start_hour":4,
    "frequency_end_hour":4,
    "reference_user":"TESSy (ECDC)",
    "reporting_email":"surveillance@ecdc.europa.eu",
    "globals":[
      {"variable":"source"},
      {"variable":"pathogen_name", "value":"COVID-19"}
    ],
    "update_scope":[
      {"variable":"source"},
      {"variable":"reporting_date"}

    ]
  },
  "acquisition":{
    "channel":{
      "name":"url",
      "url": "https://opendata.ecdc.europa.eu/covid19/nationalcasedeath_eueea_daily_ei/csv/data.csv"
    },
    "format": {
      "name":"csv",
      "decimal_sign":".",
      "thousands_separator":"",
      "date_format":"%d/%m/%Y",
      "encoding":"UTF-8"
    }
  },
  "columns":[
    {"name":"dateRep", "variable":"reporting_date"},
    {"name":"day"},
    {"name":"month"},
    {"name":"year"},
    {"name":"cases", "variable":"confirmed_cases"},
    {"name":"deaths", "variable":"deaths_infected"},
    {"name":"countriesAndTerritories"},
    {"name":"geoId", "variable":"geo_code"},
    {"name":"countryterritoryCode"},
    {"name":"popData2020", "variable":"population"}
  ]
}
```

- ECDC COVID-19 variants



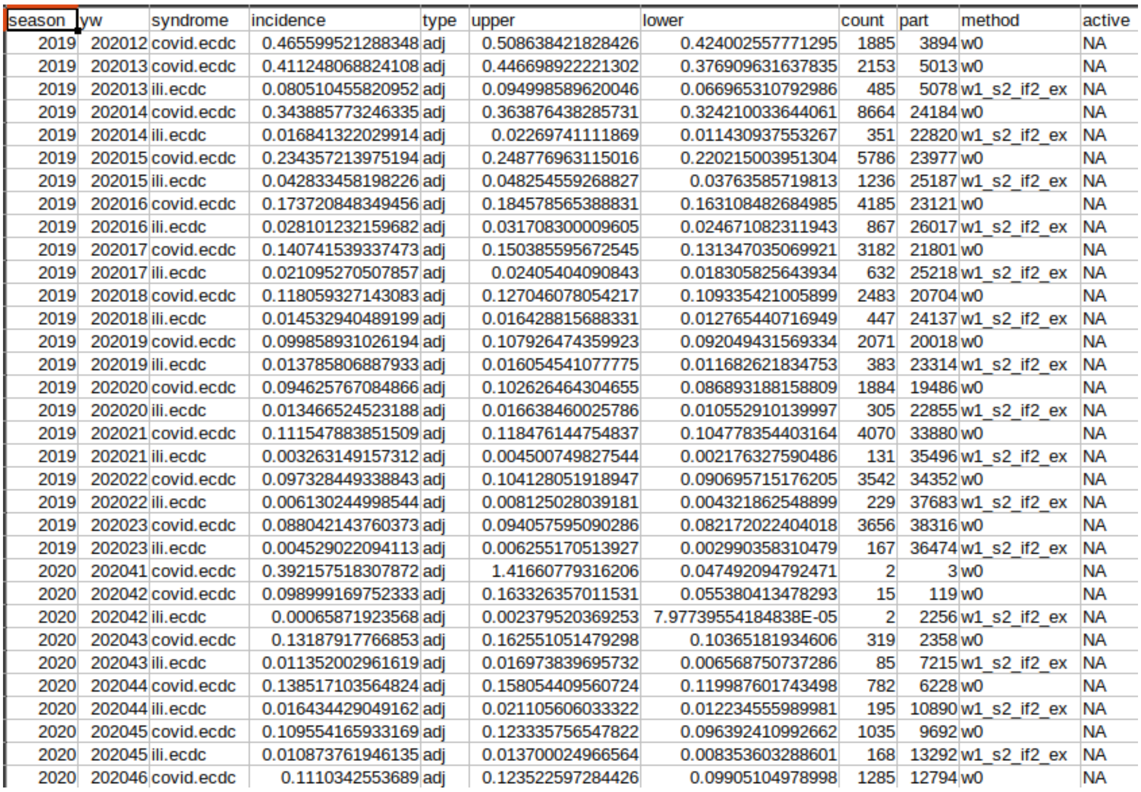
country	country_code	year_week	source	new_cases	number_sequenced	percent_cases_sequenced	valid_denominator	variant	number_detections_variant	number_sequenced_known_variant	percent_variant
Austria	AT	2020-01	GISAID	0	0		Yes	A.23.1+E484K	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	A.27	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	A.28	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	AT.1	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	AV.1	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	AY.4.2	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.1.318	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.1.519	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.1.7	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.1.7+E484K	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.1.7+L452R	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.1.7+S494P	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.214.2	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.351	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.351+E516Q	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.351+P384L	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.427/B.1.429	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.525	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.526	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.616	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.1	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.2	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.2+E484X	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.2+K417N	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.2+Q613H	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.2+Q677H	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.617.3	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.620	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.621	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	B.1.640	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	BA.1	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	BA.2	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	BA.2+L452X	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	BA.3	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	BA.4	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	BA.5	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	C.1.2	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	C.16	0	0	
Austria	AT	2020-01	GISAID	0	0		Yes	C.36+L452R	0	0	

- ECDC COVID-19 variants DLS



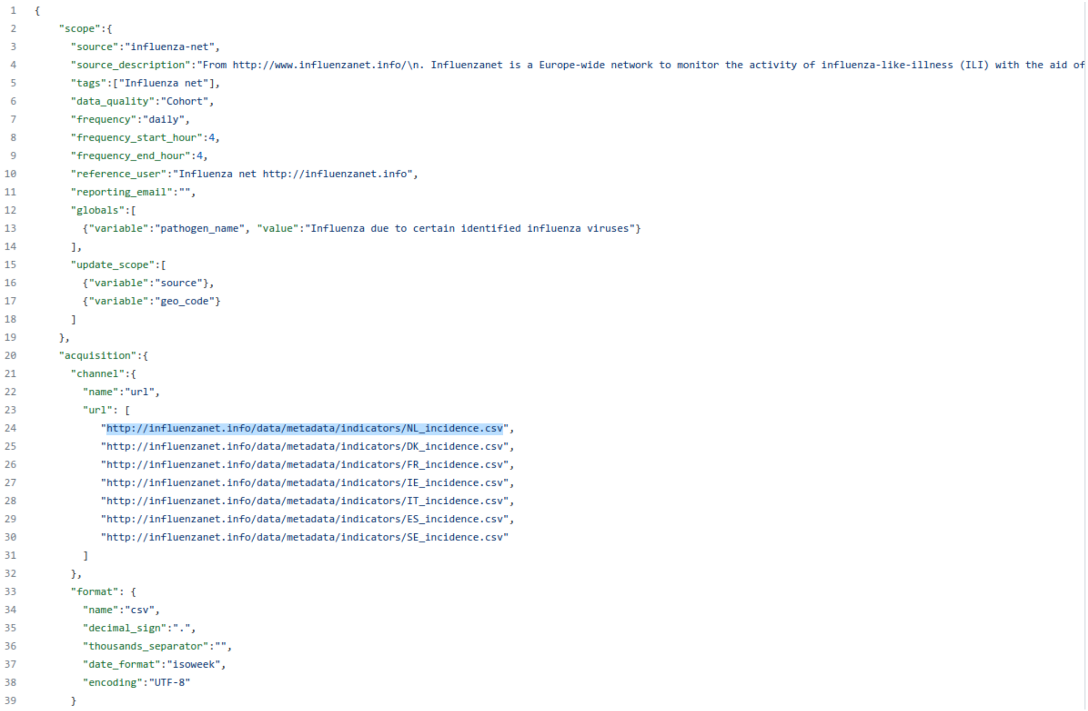
```
{
  "scope":{
    "source":"ecdc-covid19-variants",
    "source_description":"Information about the volume of COVID-19 sequencing, the number and percentage of variants of concern by week, country and variant. Data collected by
    "tags":["ECDC COVID-19"],
    "data_quality":"Official",
    "frequency":"daily",
    "frequency_start_hour":4,
    "frequency_end_hour":4,
    "reference_user":"TESSy (ECDC)",
    "reporting_email":"surveillance@ecdc.europa.eu",
    "globals":[
      {"variable":"source"},
      {"variable":"pathogen_name", "value":"COVID-19"}
    ],
    "update_scope":[
      {"variable":"source"},
      {"variable":"reporting_week"}

    ]
  },
  "acquisition":{
    "channel":{
      "name":"url",
      "url": "https://opendata.ecdc.europa.eu/covid19/virusvariant/csv/data.csv"
    },
    "format": {
      "name":"csv",
      "decimal_sign":".",
      "thousands_separator":"",
      "date_format":"isoweek",
      "encoding":"UTF-8"
    }
  },
  "columns":[
    {"name":"country_code", "variable":"geo_code"},
    {"name":"year_week", "variable":"reporting_week"},
    {"name":"variant", "variable":"variant",  "action":"insert"},
    {"name":"number_detections_variant", "variable":"number_detections_variant"}
  ]
}
```

- Influenza net participatory surveillance

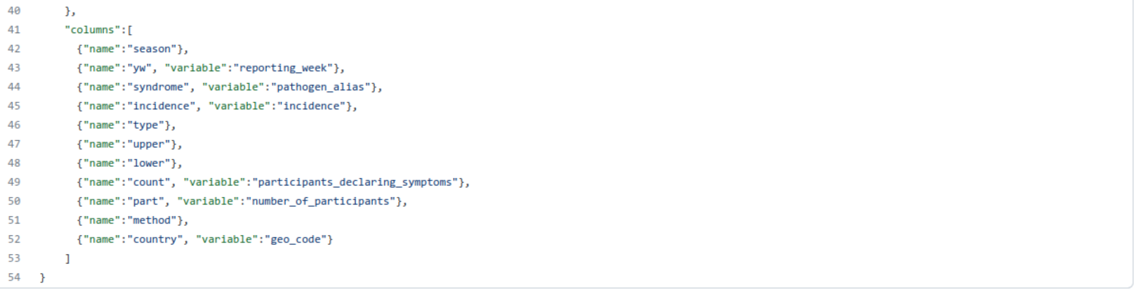


season	yw	syndrome	incidence	type	upper	lower	count	part	method	active
2019	202012	covid.ecdc	0.465599521288348	adj	0.508638421828426	0.424002557771295	1885	3894	w0	NA
2019	202013	covid.ecdc	0.411248068824108	adj	0.446698922221302	0.376909631637835	2153	5013	w0	NA
2019	202013	ili.ecdc	0.080510455820952	adj	0.094998589620046	0.066965310792986	485	5078	w1_s2_if2_ex	NA
2019	202014	covid.ecdc	0.343885773246335	adj	0.363876438285731	0.324210033644061	8664	24184	w0	NA
2019	202014	ili.ecdc	0.016841322029914	adj	0.02269741111869	0.011430937553267	351	22820	w1_s2_if2_ex	NA
2019	202015	covid.ecdc	0.234357213975194	adj	0.248776963115016	0.220215003951304	5786	23977	w0	NA
2019	202015	ili.ecdc	0.042833458198226	adj	0.048254559268827	0.03763585719813	1236	25187	w1_s2_if2_ex	NA
2019	202016	covid.ecdc	0.173720848349456	adj	0.184578565388831	0.163108482684985	4185	23121	w0	NA
2019	202016	ili.ecdc	0.028101232159682	adj	0.031708300009605	0.024671082311943	867	26017	w1_s2_if2_ex	NA
2019	202017	covid.ecdc	0.140741539337473	adj	0.150385595672545	0.131347035069921	3182	21801	w0	NA
2019	202017	ili.ecdc	0.021095270507857	adj	0.02405404090843	0.018305825643934	632	25218	w1_s2_if2_ex	NA
2019	202018	covid.ecdc	0.118059327143083	adj	0.127046078054217	0.109335421005899	2483	20704	w0	NA
2019	202018	ili.ecdc	0.014532940489199	adj	0.016428815688331	0.012765440716949	447	24137	w1_s2_if2_ex	NA
2019	202019	covid.ecdc	0.099858931026194	adj	0.107926474359923	0.092049431569334	2071	20018	w0	NA
2019	202019	ili.ecdc	0.013785806887933	adj	0.016054541077775	0.011682621834753	383	23314	w1_s2_if2_ex	NA
2019	202020	covid.ecdc	0.094625767084866	adj	0.102626464304655	0.086893188158809	1884	19486	w0	NA
2019	202020	ili.ecdc	0.013466524523188	adj	0.016638460025786	0.010552910139997	305	22855	w1_s2_if2_ex	NA
2019	202021	covid.ecdc	0.111547883851509	adj	0.118476144754837	0.104778354403164	4070	33880	w0	NA
2019	202021	ili.ecdc	0.003263149157312	adj	0.004500749827544	0.002176327590486	131	35496	w1_s2_if2_ex	NA
2019	202022	covid.ecdc	0.097328449338843	adj	0.104128051918947	0.090695715176205	3542	34352	w0	NA
2019	202022	ili.ecdc	0.006130244998544	adj	0.008125028039181	0.004321862548899	229	37683	w1_s2_if2_ex	NA
2019	202023	covid.ecdc	0.088042143760373	adj	0.094057595090286	0.082172022404018	3656	38316	w0	NA
2019	202023	ili.ecdc	0.004529022094113	adj	0.006255170513927	0.002990358310479	167	36474	w1_s2_if2_ex	NA
2020	202041	covid.ecdc	0.392157518307872	adj	1.41660779316206	0.047492094792471	2	3	w0	NA
2020	202042	covid.ecdc	0.098999169752333	adj	0.163326357011531	0.055380413478293	15	119	w0	NA
2020	202042	ili.ecdc	0.00065871923568	adj	0.002379520369253	7.97739554184838E-05	2	2256	w1_s2_if2_ex	NA
2020	202043	covid.ecdc	0.13187917766853	adj	0.162551051479298	0.10365181934606	319	2358	w0	NA
2020	202043	ili.ecdc	0.011352002961619	adj	0.016973839695732	0.006568750737286	85	7215	w1_s2_if2_ex	NA
2020	202044	covid.ecdc	0.138517103564824	adj	0.158054409560724	0.119987601743498	782	6228	w0	NA
2020	202044	ili.ecdc	0.016434429049162	adj	0.021105606033322	0.012234555989981	195	10890	w1_s2_if2_ex	NA
2020	202045	covid.ecdc	0.109554165933169	adj	0.123335756547822	0.096392410992662	1035	9692	w0	NA
2020	202045	ili.ecdc	0.010873761946135	adj	0.013700024966564	0.008353603288601	168	13292	w1_s2_if2_ex	NA
2020	202046	covid.ecdc	0.1110342553689	adj	0.123522597284426	0.09905104978998	1285	12794	w0	NA

- Influenza net participatory surveillance DLS

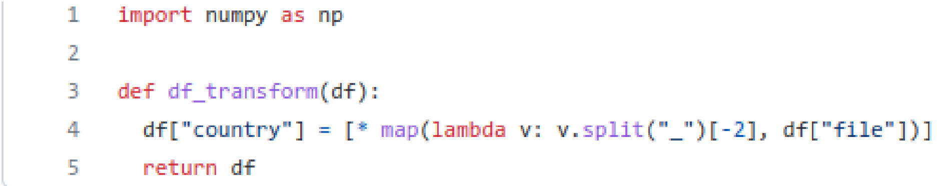


```
{
    "scope":{
      "source":"influenza-net",
      "source_description":"From http://www.influenzanet.info/\n. Influenzanet is a Europe-wide network to monitor the activity of influenza-like-illness (ILI) with the aid of
      "tags":["Influenza net"],
      "data_quality":"Cohort",
      "frequency":"daily",
      "frequency_start_hour":4,
      "frequency_end_hour":4,
      "reference_user":"Influenza net http://influenzanet.info",
      "reporting_email":"",
      "globals":[
        {"variable":"pathogen_name", "value":"Influenza due to certain identified influenza viruses"}
      ],
      "update_scope":[
        {"variable":"source"},
        {"variable":"geo_code"}
      ]
    },
    "acquisition":{
      "channel":{
        "name":"url",
        "url": [
           "http://influenzanet.info/data/metadata/indicators/NL_incidence.csv",
           "http://influenzanet.info/data/metadata/indicators/DK_incidence.csv",
           "http://influenzanet.info/data/metadata/indicators/FR_incidence.csv",
           "http://influenzanet.info/data/metadata/indicators/IE_incidence.csv",
           "http://influenzanet.info/data/metadata/indicators/IT_incidence.csv",
           "http://influenzanet.info/data/metadata/indicators/ES_incidence.csv",
           "http://influenzanet.info/data/metadata/indicators/SE_incidence.csv"
        ]
      },
      "format": {
        "name":"csv",
        "decimal_sign":".",
        "thousands_separator":"",
        "date_format":"isoweek",
        "encoding":"UTF-8"
      }
```



```
    },
    "columns":[
      {"name":"season"},
      {"name":"yw", "variable":"reporting_week"},
      {"name":"syndrome", "variable":"pathogen_alias"},
      {"name":"incidence", "variable":"incidence"},
      {"name":"type"},
      {"name":"upper"},
      {"name":"lower"},
      {"name":"count", "variable":"participants_declaring_symptoms"},
      {"name":"part", "variable":"number_of_participants"},
      {"name":"method"},
      {"name":"country", "variable":"geo_code"}
    ]
}
```

In this last particular case, the country of the file needs to be extracted from the name of the files. This kind of extractions or other custom modifications to perform on the obtained data frame can be done on a custom python script that needs to be placed on the scripts folder containing the source filename and the structure shown here below.



```python
import numpy as np

def df_transform(df):
  df["country"] = [* map(lambda v: v.split("_")[-2], df["file"])]
  return df
```

C3. DLS implementation: Integrating datasets

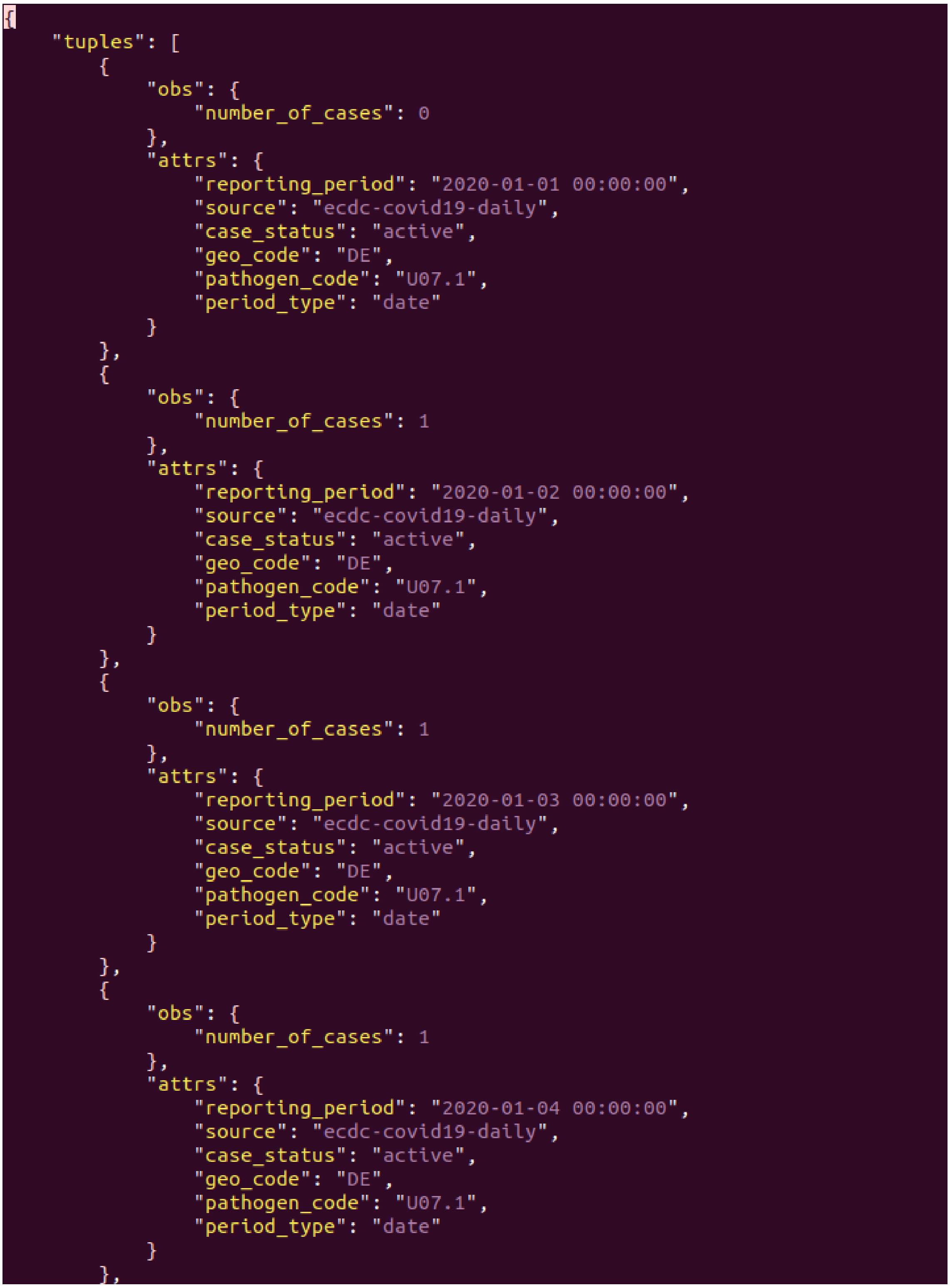
The reference variables and the source mapping will allow an automatic standardisation of each new dataset that is read by the system. The processed input will always contain the same structure independently of the data source or destination table. This will allow a robust and seamless data integration independently of the implementation of the data model. A processed input dataset will contain the following structure:

- Scope: General information of the read dataset
  - Source: Name of the source associated with this file (this will link with the data source definition including all necessary information for interpreting the file).
  - File name: Name of the source file containing the dataset.
  - Sent on: The timestamp when the file was sent to the system
  - Sent by: The user sending the file
  - Update scope: The variables and values that are to be replaced on the database if exists.
- Tuples: Data contained in the file
  - A list of tuples containing the observations, attributes and indicators in the file.
  - Each tuple can only contain attributes present in the reference variable list
  - Each tuple value is guaranteed to be on the type indicated in the reference variable list
  - Any attribute associated with a measure or indicator has to be on the list of linked attributes described in the reference variable list.
  - Any code associated to a referential is guaranteed to be a valid reference

As it can be seen, the produced output is independent of the source. Therefore importing data into the database will only require knowing, a) The reference list of variables, b) The processed output.

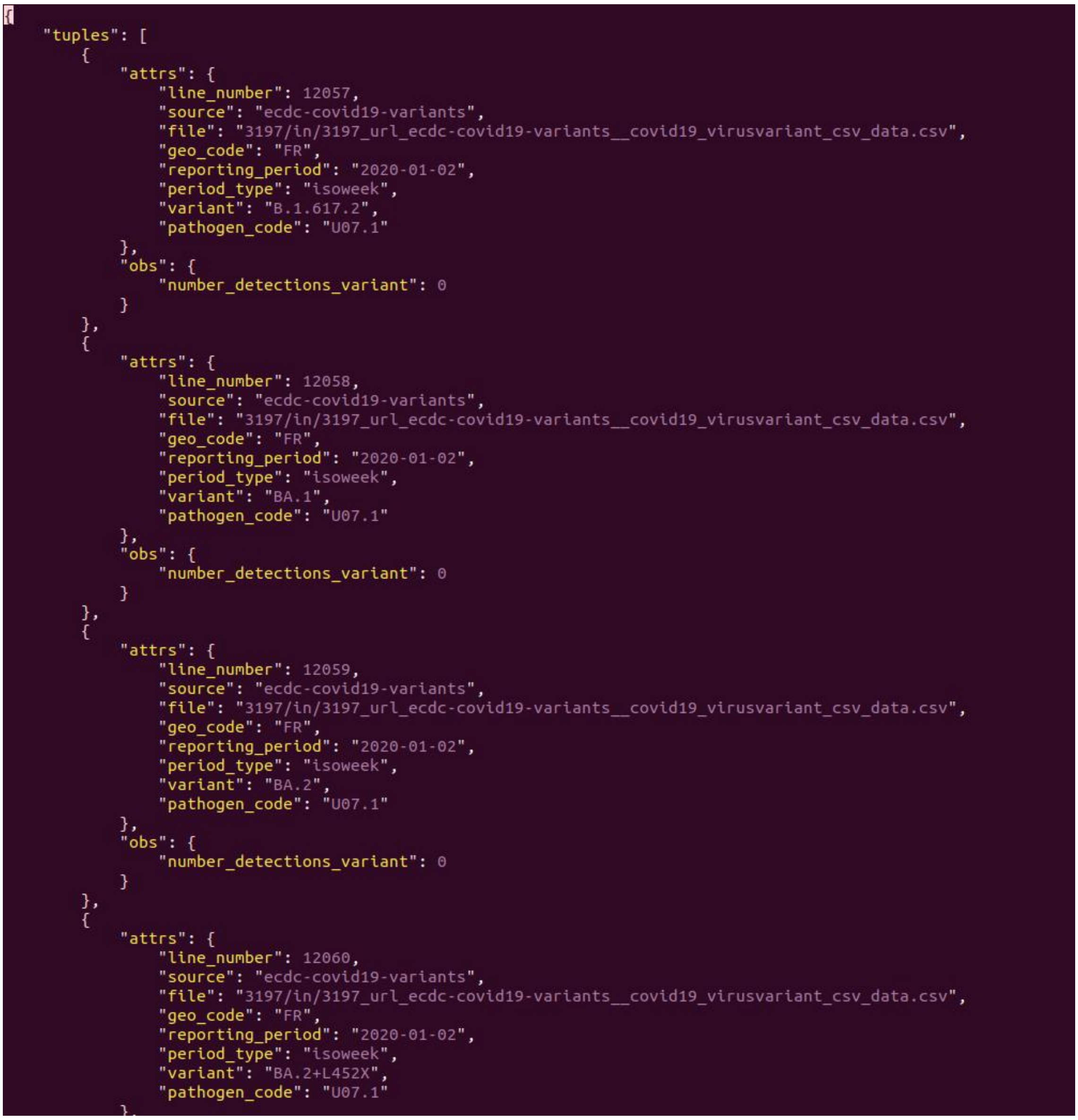
The following image shows an example of the processed files from ECDC cases, variants and Influenza net participatory surveillance

- ECDC processed cases:



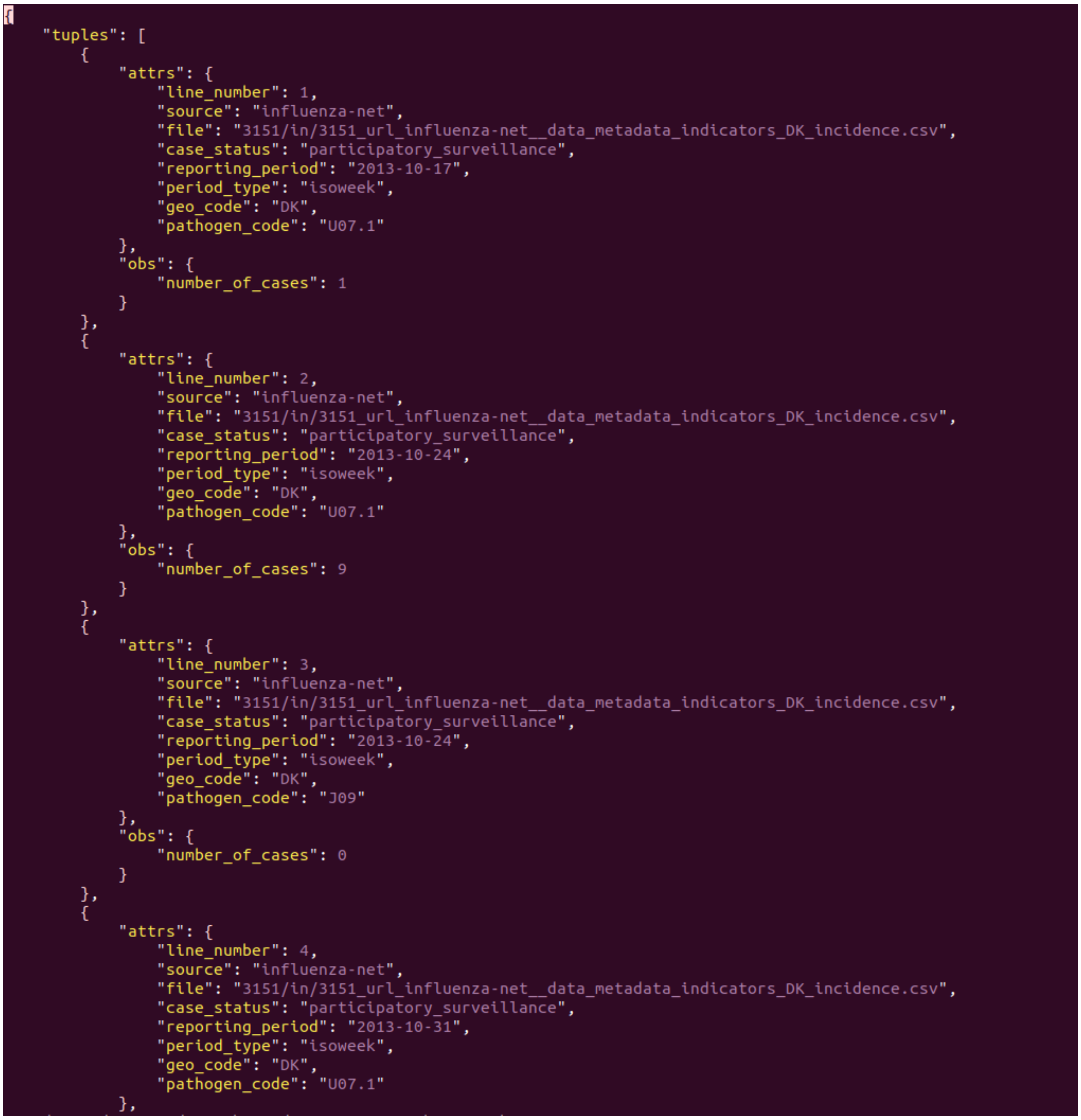
```
{
    "tuples": [
        {
            "obs": {
                "number_of_cases": 0
            },
            "attrs": {
                "reporting_period": "2020-01-01 00:00:00",
                "source": "ecdc-covid19-daily",
                "case_status": "active",
                "geo_code": "DE",
                "pathogen_code": "U07.1",
                "period_type": "date"
            }
        },
        {
            "obs": {
                "number_of_cases": 1
            },
            "attrs": {
                "reporting_period": "2020-01-02 00:00:00",
                "source": "ecdc-covid19-daily",
                "case_status": "active",
                "geo_code": "DE",
                "pathogen_code": "U07.1",
                "period_type": "date"
            }
        },
        {
            "obs": {
                "number_of_cases": 1
            },
            "attrs": {
                "reporting_period": "2020-01-03 00:00:00",
                "source": "ecdc-covid19-daily",
                "case_status": "active",
                "geo_code": "DE",
                "pathogen_code": "U07.1",
                "period_type": "date"
            }
        },
        {
            "obs": {
                "number_of_cases": 1
            },
            "attrs": {
                "reporting_period": "2020-01-04 00:00:00",
                "source": "ecdc-covid19-daily",
                "case_status": "active",
                "geo_code": "DE",
                "pathogen_code": "U07.1",
                "period_type": "date"
            }
        },
```

- ECDC processed variants



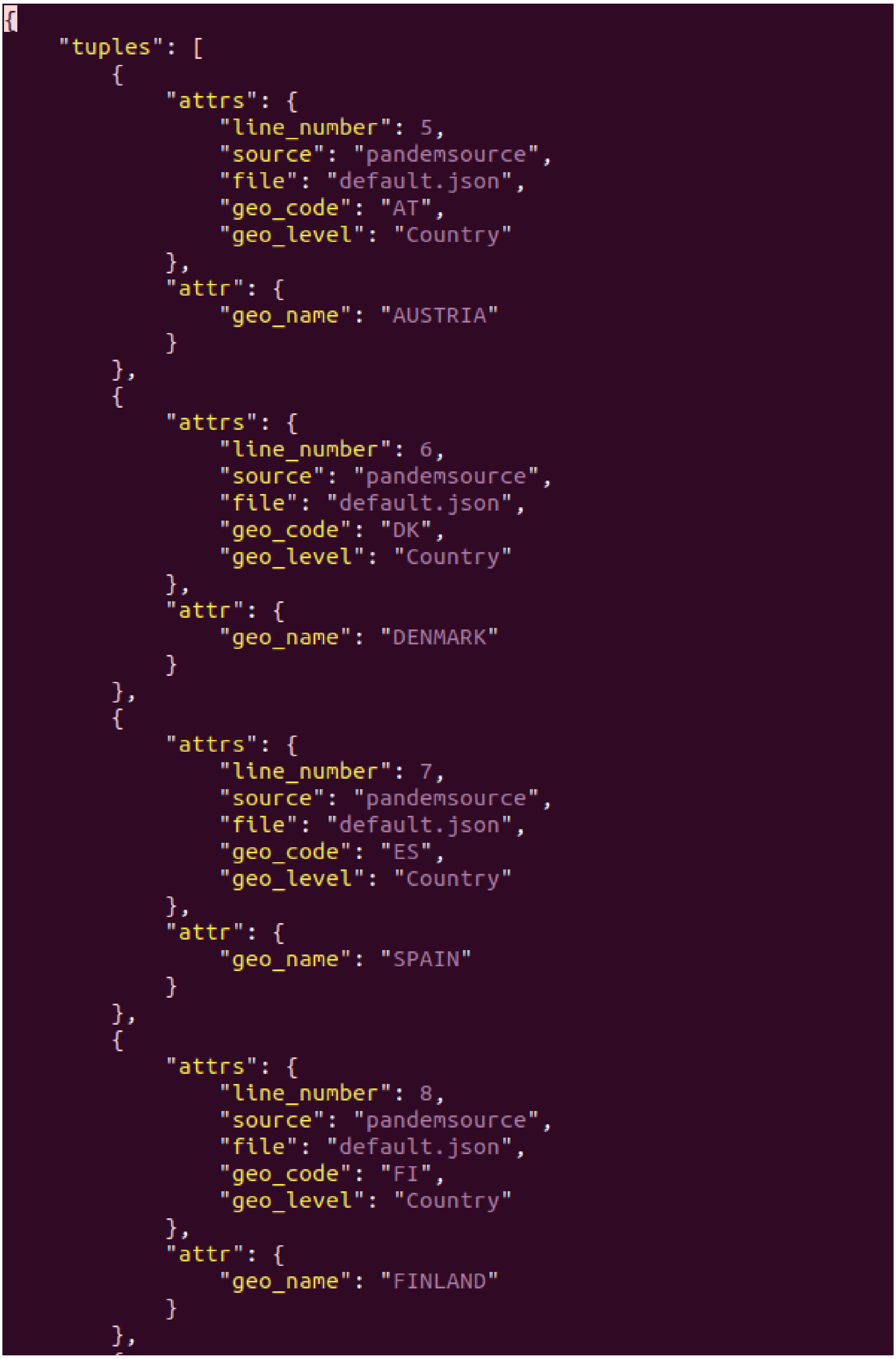
```
{
    "tuples": [
        {
            "attrs": {
                "line_number": 12057,
                "source": "ecdc-covid19-variants",
                "file": "3197/in/3197_url_ecdc-covid19-variants__covid19_virusvariant_csv_data.csv",
                "geo_code": "FR",
                "reporting_period": "2020-01-02",
                "period_type": "isoweek",
                "variant": "B.1.617.2",
                "pathogen_code": "U07.1"
            },
            "obs": {
                "number_detections_variant": 0
            }
        },
        {
            "attrs": {
                "line_number": 12058,
                "source": "ecdc-covid19-variants",
                "file": "3197/in/3197_url_ecdc-covid19-variants__covid19_virusvariant_csv_data.csv",
                "geo_code": "FR",
                "reporting_period": "2020-01-02",
                "period_type": "isoweek",
                "variant": "BA.1",
                "pathogen_code": "U07.1"
            },
            "obs": {
                "number_detections_variant": 0
            }
        },
        {
            "attrs": {
                "line_number": 12059,
                "source": "ecdc-covid19-variants",
                "file": "3197/in/3197_url_ecdc-covid19-variants__covid19_virusvariant_csv_data.csv",
                "geo_code": "FR",
                "reporting_period": "2020-01-02",
                "period_type": "isoweek",
                "variant": "BA.2",
                "pathogen_code": "U07.1"
            },
            "obs": {
                "number_detections_variant": 0
            }
        },
        {
            "attrs": {
                "line_number": 12060,
                "source": "ecdc-covid19-variants",
                "file": "3197/in/3197_url_ecdc-covid19-variants__covid19_virusvariant_csv_data.csv",
                "geo_code": "FR",
                "reporting_period": "2020-01-02",
                "period_type": "isoweek",
                "variant": "BA.2+L452X",
                "pathogen_code": "U07.1"
            },
```

- Influenza net number of participants declaring symptoms



```
{
    "tuples": [
        {
            "attrs": {
                "line_number": 1,
                "source": "influenza-net",
                "file": "3151/in/3151_url_influenza-net__data_metadata_indicators_DK_incidence.csv",
                "case_status": "participatory_surveillance",
                "reporting_period": "2013-10-17",
                "period_type": "isoweek",
                "geo_code": "DK",
                "pathogen_code": "U07.1"
            },
            "obs": {
                "number_of_cases": 1
            }
        },
        {
            "attrs": {
                "line_number": 2,
                "source": "influenza-net",
                "file": "3151/in/3151_url_influenza-net__data_metadata_indicators_DK_incidence.csv",
                "case_status": "participatory_surveillance",
                "reporting_period": "2013-10-24",
                "period_type": "isoweek",
                "geo_code": "DK",
                "pathogen_code": "U07.1"
            },
            "obs": {
                "number_of_cases": 9
            }
        },
        {
            "attrs": {
                "line_number": 3,
                "source": "influenza-net",
                "file": "3151/in/3151_url_influenza-net__data_metadata_indicators_DK_incidence.csv",
                "case_status": "participatory_surveillance",
                "reporting_period": "2013-10-24",
                "period_type": "isoweek",
                "geo_code": "DK",
                "pathogen_code": "J09"
            },
            "obs": {
                "number_of_cases": 0
            }
        },
        {
            "attrs": {
                "line_number": 4,
                "source": "influenza-net",
                "file": "3151/in/3151_url_influenza-net__data_metadata_indicators_DK_incidence.csv",
                "case_status": "participatory_surveillance",
                "reporting_period": "2013-10-31",
                "period_type": "isoweek",
                "geo_code": "DK",
                "pathogen_code": "U07.1"
            },
```

- The following example shows a dataset containing non numeric information, such as the relationship between country names and codes



```
{
    "tuples": [
        {
            "attrs": {
                "line_number": 5,
                "source": "pandemsource",
                "file": "default.json",
                "geo_code": "AT",
                "geo_level": "Country"
            },
            "attr": {
                "geo_name": "AUSTRIA"
            }
        },
        {
            "attrs": {
                "line_number": 6,
                "source": "pandemsource",
                "file": "default.json",
                "geo_code": "DK",
                "geo_level": "Country"
            },
            "attr": {
                "geo_name": "DENMARK"
            }
        },
        {
            "attrs": {
                "line_number": 7,
                "source": "pandemsource",
                "file": "default.json",
                "geo_code": "ES",
                "geo_level": "Country"
            },
            "attr": {
                "geo_name": "SPAIN"
            }
        },
        {
            "attrs": {
                "line_number": 8,
                "source": "pandemsource",
                "file": "default.json",
                "geo_code": "FI",
                "geo_level": "Country"
            },
            "attr": {
                "geo_name": "FINLAND"
            }
        },
```

6.6.2. Database schema for integrating into PANDEM-2 database

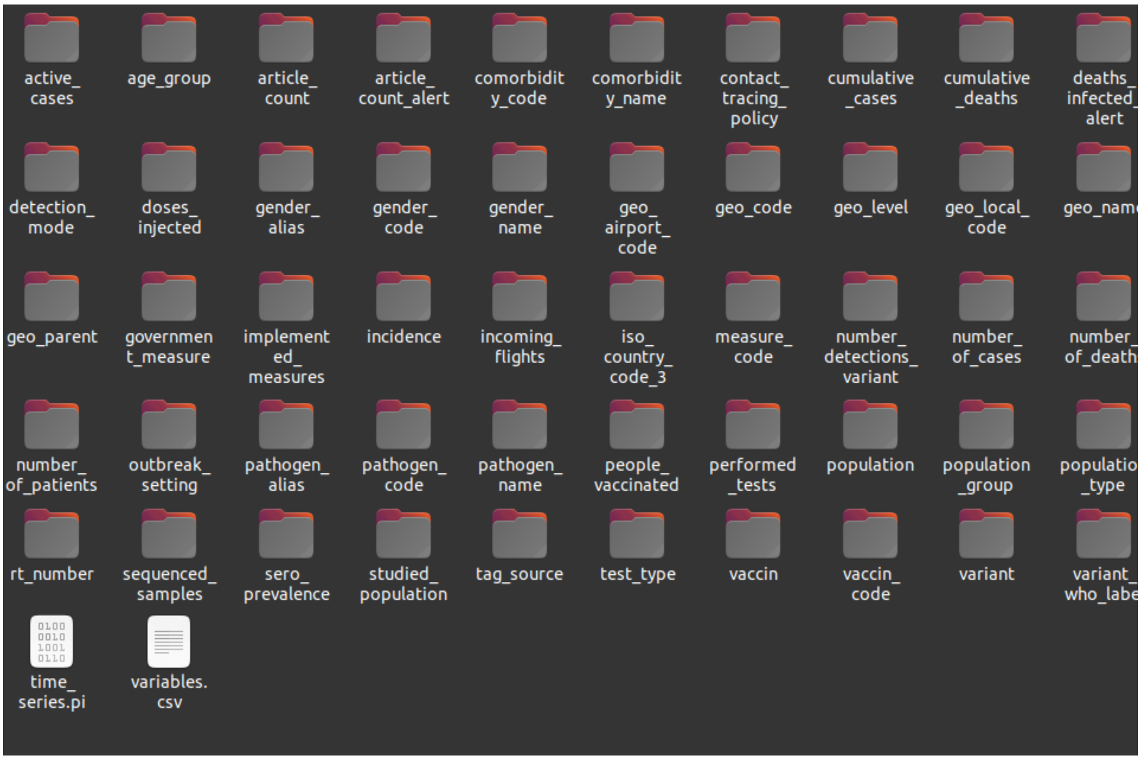
As explained in the previous sections, integrating data is simplified to three json-like tables thanks to the PANDEM-2 DLS.

- List of Variables
- Data Sources
- Datasets

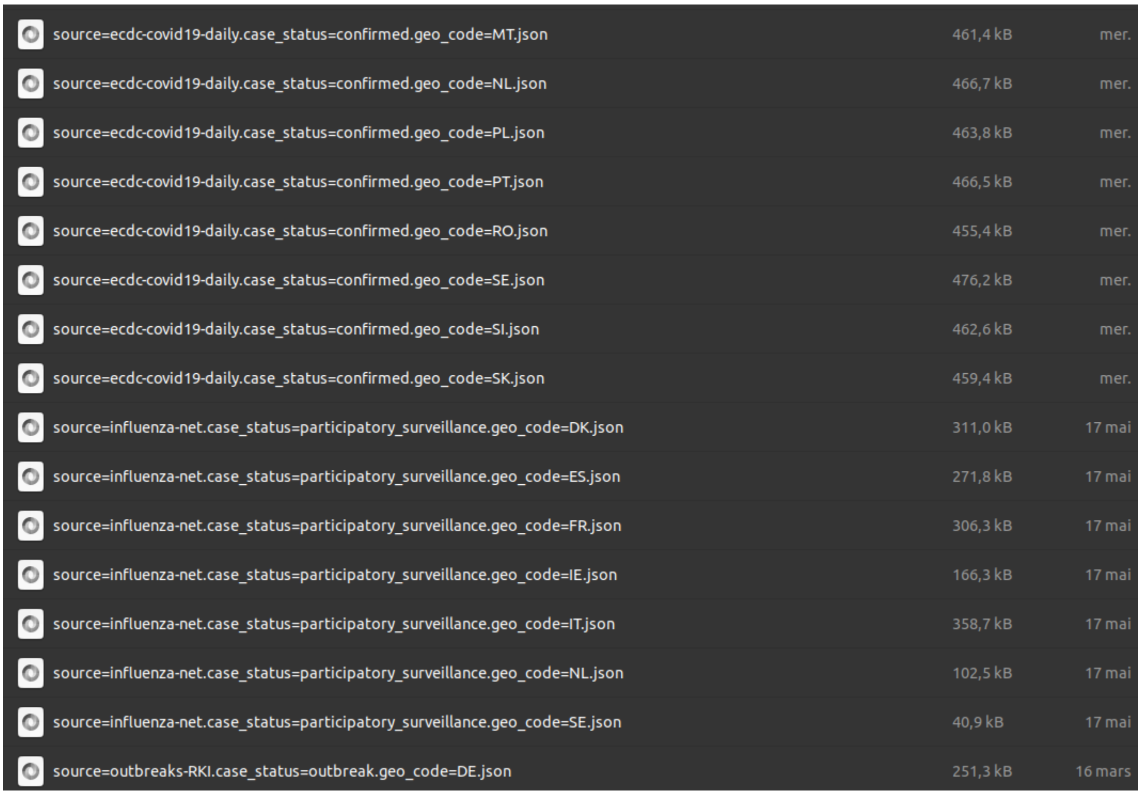
The fields of each one of these tables are described in section 6.6.1.

The final destination for each variable in the existing database model is presented in the Global database Model shown in PANDEM-2 deliverable D3.1, where each data family is related to a group of tables. Nevertheless, the abstraction layer of the DLS allows the model to evolve independently of the data families notion.

The datasets are stored in a dedicated folder. Each folder contains JSON datasets for a single main variable (observation or referential) with its respective attributes. To exemplify this, we display the processed variables folders from our PANDEM development version.



Each folder contains JSON files with the processed tuples following the partitioning schema defined on the list of variables. By example for the variable number of cases you can see the JSON files as follows:



source=ecdc-covid19-daily.case_status=confirmed.geo_code=MT.json	461,4 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=NL.json	466,7 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=PL.json	463,8 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=PT.json	466,5 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=RO.json	455,4 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=SE.json	476,2 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=SI.json	462,6 kB	mer.
source=ecdc-covid19-daily.case_status=confirmed.geo_code=SK.json	459,4 kB	mer.
source=influenza-net.case_status=participatory_surveillance.geo_code=DK.json	311,0 kB	17 mai
source=influenza-net.case_status=participatory_surveillance.geo_code=ES.json	271,8 kB	17 mai
source=influenza-net.case_status=participatory_surveillance.geo_code=FR.json	306,3 kB	17 mai
source=influenza-net.case_status=participatory_surveillance.geo_code=IE.json	166,3 kB	17 mai
source=influenza-net.case_status=participatory_surveillance.geo_code=IT.json	358,7 kB	17 mai
source=influenza-net.case_status=participatory_surveillance.geo_code=NL.json	102,5 kB	17 mai
source=influenza-net.case_status=participatory_surveillance.geo_code=SE.json	40,9 kB	17 mai
source=outbreaks-RKI.case_status=outbreak.geo_code=DE.json	251,3 kB	16 mars

6.7. List of open available variables

In this section we propose a detailed list of variables and list of sources that could be available within any PANDEM-2 installation. After refinement the PANDEM-2 team will develop and publish DLS for each of these sources, allowing any user to access this data “out of the box”.

This list of sources is not final or exhaustive in terms of within or among pathogens. It is built as a representation of well-known and reliable sources that are a result of epidemiological monitoring of existing epidemics. Most of them are associated with a particular contingency e.g., COVID-19 pandemic and are not expected to last in the long term. This nature of data shapes PANDEM-2 as a tool that can adapt very quickly to new challenges imposed from future threats.

This list represents the first variables feeding the PANDEM-2 database and that in a few months, together with end-user data uploaded in their own PANDEM-2 installations (restricted or not) will allow the use of the PANDEM-2 tools for pandemic management to be tested within the PANDEM-2 project within WP6.

6.7.1. Cases + Patient + Deaths (COVID)

- **John Hopkins University COVID-Dataset**
  - **PANDEM-2 use case:** Get reference COVID-19 epidemiological metrics (cases, deaths, recovered, incidence rate, fatality ratio) at global level by country and finer levels when available (NUTS2 for EU states).
  - Compilation: Yes
  - Institutions
    - Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).
    - ESRI Living Atlas Team
    - Johns Hopkins University Applied Physics Lab (JHU APL).
  - Project Leaders:
    - Ensheng Dong
    - Hongru Du
    - Lauren Gardner
  - Data[32]Frequency: Daily
  - License: Open Data (CC BY 4.0)
  - Data Dictionary[33]Protocol [34]API Access: git[35]
  - Geographic scope:
    - Coverage: Global
    - Detail: Multiple (NUTS3 for europe)
  - Variables:
    - Date
    - FIPS
    - Admin2
    - Province_State
    - Country_Region
    - Last_Update
    - Lat
    - Long
    - Confirmed
    - Deaths
    - Recovered
    - Active
    - Combined_Key
    - Incident_Rate
    - Case_Fatality_Ratio

[32] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports

[33] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

[34]https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext

[35] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports

- **John Hopkins COVID-Unified Dataset**
    - **PANDEM-2 use case:** Doing modeling and predictions with John Hopkins University COVID-Dataset dataset enriched with more variables (ICU, Tests) weather, measures, and comorbidity statistics by locality.
    - Compilation: Yes
    - Institutions
        - Johns Hopkins University
        - NASA Health & Air Quality project
        - National Institute of Health (NIH)
    - Project Leaders:
        - Lauren M. Gardner
        - Hamada S. Badr
    - Data[36]Frequency: Daily
    - License: Open data
    - Data Dictionary[37]Protocol[38]API Access: git[39]Geographic scope:
        - Coverage: Global
        - Detail: Variable
    - Variables:
        - Epidemiological (on COVID-Unified Dataset)
            - Active Cases
            - Confirmed Cases
            - Deaths
            - Recovered
            - Hospitalized
            - Home_Confinement
            - Hospitalized_Sym
            - Hospitalized_Now
            - Ventilator
            - Ventilator_Now"
            - ICU
            - ICU_Now
            - Tested
            - Tests
            - Pending
            - Positive
            - Negative
            - Positive_Dx
            - Positive_Sc

[36] https://github.com/CSSEGISandData/COVID-19_Unified-Dataset

[37] https://github.com/CSSEGISandData/COVID-19_Unified-Dataset/blob/master/README.md

[38] https://www.medrxiv.org/content/10.1101/2021.05.05.21256712v1

[39] https://github.com/CSSEGISandData/COVID-19_Unified-Dataset

- Hospitalized_Now
- Ventilator
- Ventilator_Now

6.7.2. Weather (on COVID-Unified Dataset)

- Date of data record
- Daily average near-surface air temperature
- Daily maximum near-surface air temperature
- Daily minimum near-surface air temperature
- Daily average near-surface dew point temperature
- Daily average near-surface dew point depression
- Daily average near-surface relative humidity
- Daily average near-surface specific humidity
- Daily average moisture availability (NLDAS)
- Daily average root zone soil moisture content (NLDAS)
- Daily average soil moisture content (NLDAS)
- Daily average volumetric soil water layer 1 (ERA5)
- Daily average volumetric soil water layer 2 (ERA5)
- Daily average volumetric soil water layer 3 (ERA5)
- Daily average volumetric soil water layer 4 (ERA5)
- Daily average surface pressure
- Daily average surface downward solar radiation (ERA5)
- Daily average surface downward longwave radiation flux (NLDAS)
- Daily average surface downward shortwave radiation flux (NLDAS)
- Daily average surface latent heat flux (ERA5)
- Daily average surface latent heat flux (NLDAS)
- Daily average potential evaporation / latent heat flux (ERA5)
- Daily average potential evaporation / latent heat flux (NLDAS)
- Daily total precipitation
- Daily average 10-m above ground Zonal wind speed
- Daily average 10-m above ground Meridional wind speed
- Data source: ERA5, NLDAS ± CIESIN*

- Risk factors (on COVID-Unified Dataset)
  - Fine particulate matter (PM2.5) concentration
  - Nitrogen dioxide (NO2) concentration (2014-2018 mean)
  - Travel time to nearest cities
  - Travel time to health care facilities, with motorized transport
  - Travel time to health care facilities, without motorized transport
  - Prevalence of adults with diagnosed diabetes

        - Percent of obese adults who are current smokers
        - Prevalence of chronic obstructive pulmonary disease by sex
        - Prevalence of cardiovascular disease by sex
        - Prevalence of HIV/AIDS by sex
        - Percent of adults with hypertension by sex
        - Proportion of individuals with at least 1 risk conditions for COVID-19
        - Proportion of individuals that would require hospital admission if infected.
        - Total MERS cases by country (October 2012 - February 2018)
        - Total SARS cases by country (1 November 2002 - 7 August 2003)
        - Total population from WorldPop
        - Population density from WorldPop
        - Population proportion over 65 years old from WorldPop
        - Population proportion by sex (Female) from WorldPop
        - Population proportion by sex (Male) from WorldPop
        - Sex ratio (Male / Female) from WorldPop
    - Policies (on COVID-Unified Dataset)
        - Date
        - Type of the policy
        - Value of the policy
        - Logical flag for geographic scope
        - Notes on the policy record

6.7.3. Sero prevalence (COVID-19)

- SeroTracker
    - PANDEM-2 use case: Get updated **COVID-19 seroprevalence** results for multiple and heterogeneous studies compiled globally.
    - Compilation: Yes
    - Institutions
        - Public Health Agency of Canada
        - COVID-19 Immunity Task Force
        - University of Calgary's Centre for Health Informatics
        - World Health Organization
    - Project Leaders:
        - Rahul Arora: rahul.arora@balliol.ox.ac.uk
        - Tingting Yan: tingting.yan@mail.utoronto.ca.

- Data[40]Frequency: Weekly
- License: Open Data
- Data Dictionary[41]
- Protocol[42]API Access: On request to Rahul Arora at rahul.arora@balliol.ox.ac.uk
- Geographic scope:
    - Coverage: Global
    - Detail: Multiple
- Variables:
    - Prevalence Estimate Name
    - Publication Date
    - Grade of Estimate Scope
    - Country
    - Specific Geography
    - Sampling Start Date
    - Sampling End Date
    - Sample Frame (groups of interest)
    - Sample Frame (age)
    - Denominator Value
    - Serum positive prevalence
    - Serum pos prevalence, 95pct CI Lower
    - Serum pos prevalence, 95pct CI Upper
    - Sampling Method

    1. Test Manufacturer
    2. Test Type
    3. Isotype(s) Reported
    4. Sensitivity
    5. Specificity
    6. Overall Risk of Bias (JBI)
    7. Source Type
    8. First Author Full Name
    9. Lead Institution
    10. URL
    11. Date Created
    12. Last modified time
    13. Data Quality Status

[40] https://airtable.com/shraXWPJ9Yu7ybowM/tbljN2mhRVfSlZv2d?backgroundColor=blue&viewControls=on

[41] https://docs.google.com/spreadsheets/d/1KQbp5T9Cq_HnNpmBTWY1iKs6Etu1-qJcnhdJ5eyw7N8/edit#gid=0

[42] https://docs.google.com/document/d/1NYpszkr-u__aZspFDFa_fa4VBzjAAAAxNxM1rZ1txWU/edit

6.7.4. Vaccination (COVID-19)

- Vaccination (EU)
    - PANDEM-2 use case: Get vaccination progress at EU/EEA level
    - Compilation: No
    - Institutions
        - TESSy (ECDC)
    - Project Leaders:
        - Yet to identify
    - Data[43]Frequency: Weekly
    - License: Open Data (compatible with CC BY 4.0 license)
    - Data Dictionary[44]Protocol[45]API Access: Download[46]Geographic scope:
        - Coverage: EU/EEA
        - Detail: Country + some regions
    - Variables:
        - YearWeekISO
        - FirstDose
        - FirstDoseRefused
        - SecondDose
        - UnknownDose
        - NumberDosesReceived
        - Region
        - Population
        - ReportingCountry
        - TargetGroup
        - Vaccine
        - Denominator
- Vaccination (World)
    - PANDEM-2 use case: Get vaccination progress at world level
    - Compilation: Yes
    - Institutions
        - Our World in Data[47]Project Leaders:
        - Edouard Mathieu
    - Data[48]Frequency: Daily
    - License: Open Data (CC BY 4.0 license)
    - Data Dictionary[49]Protocol:[50]

[43] https://opendata.ecdc.europa.eu/covid19/vaccine_tracker/csv/data.csv

[44] https://www.ecdc.europa.eu/sites/default/files/documents/Variable_Dictionary_VaccineTracker-20-08-2021.pdf

[45] https://www.ecdc.europa.eu/en/covid-19/data-collection

[46] https://opendata.ecdc.europa.eu/covid19/vaccine_tracker/csv/data.csv

[47] https://ourworldindata.org/

[48] https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations

[49] https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/README.md

- API Access: git or download[51]Geographic scope:
    - Coverage: World
    - Detail: Country
- Variables:
    - location
    - date
    - manufacturer
    - daily_vaccinations
    - daily_vaccinations_per_million
    - total_vaccinations
    - total_vaccinations_per_hundred
    - age_group
    - people_vaccinated
    - people_vaccinated_per_hundred
    - people_fully_vaccinated
    - people_fully_vaccinated_per_hundred
    - location
    - iso_code
    - date
    - total_boosters
    - total_boosters_per_hundred

6.7.5. Government Measures

- Oxford Covid-19 Government Response Tracker
    - PANDEM-2 use case: Evaluate impact of government measures on epidemiological indicators
    - Compilation: Yes
    - Institutions
        - Blavatnik School of Government (Oxford)
    - Project Leaders:
        - Thomas Hale
    - Data[52]Frequency: Daily
    - License: Open Data (CC BY 4.0 license)
    - Data Dictionary[53]Protocol[54]API Access: git[55]
    - Geographic scope:
        - Coverage: World
        - Detail: Country

[50] https://www.nature.com/articles/s41562-021-01122-8
[51] https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations
[52] https://github.com/OxCGRT/covid-policy-tracker/tree/master/data
[53] https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md
[54] https://www.nature.com/articles/s41562-021-01079-8
[55] https://github.com/OxCGRT/covid-policy-tracker

- Variables:
    - Country
    - Date of start
    - Date of end
    - Comments
    - Containment and closure
        - C1 School closing
        - C2 Workplace closing
        - C3 Cancel public events
        - C4 Restrictions on gathering size
        - C5 Close public transport
        - C6 Stay-at-home requirements
        - C7 Restrictions on internal movement
        - C8 Restrictions on international travel
    - Economic response
        - E1 income support for households
        - E2 Debt/contract relief for households
        - E3 Fiscal measures
        - E4 Giving international support
    - Health systems
        - H1 Public information campaign
        - H2 Testing policy
        - H3 Contact tracing
        - H4 Emergency investment in health care
        - H5 investment in COViD-19 vaccines
    - H6 Facial coverings
    - H7 Vaccination policy Ordinal Funding
    - H8 Protection of elderly people
- Vaccines
    - V1 Vaccine prioritisation
        - By Age Group
        - By Risk and age group
        - By Occupation
    - V2 Vaccine eligibility
        - By Age Group
        - By Risk and age group
        - By Occupation
    - V3 Vaccine financial support
        - By Age Group
        - By Risk and age group
        - By Occupation

6.7.6. News aggregator

- European Media Monitor[56]: In EMM we can retrieve articles from around 11,200 international sources. The themes of these articles are more general such as terrorism, crime or conflict.
  - PANDEM-2 use case: identify alerts by country on user defined topics for natural disasters or other hazards.
  - Compilation: Yes
  - Institutions
    - European Commission's Joint Research Centre
  - Project Leaders (paper authors):
    - Guillaume Jacquet
    - Ralf Steinberger
  - Data[57]Frequency: updated every 10 minutes, 24 hours per day
  - License: Open Data
  - Data Dictionary: Not Found
  - Protocol: Not found
  - API Access: RSS[58]Geographic scope:
    - Coverage: World
    - Detail: Country
  - Variables:
    - Title
    - Link
    - Description
    - Keywords
    - Publication date
    - ID
    - Source
    - Language
    - Categories
    - Geolocation country

- Medisys:[59]: MEDISYS is a media monitoring system providing event-based surveillance to rapidly identify potential public health threats using information from media reports. The system displays only those articles with interest to public health (e. g. diseases, plant pests, psychoactive substances), analyses news reports and warns users with automatically generated alerts.

[56] https://ec.europa.eu/jrc/en/scientific-tools/emm

[57] https://emm.newsbrief.eu/NewsBrief/clusteredition/fr/latest.html

[58] https://emm.newsbrief.eu/rss/rss?type=rtn&language=fr&duplicates=false

[59] https://ec.europa.eu/jrc/en/scientific-tool/medical-information-system

- PANDEM-2 use case: PANDEM-2 use case: identify public health by country on user defined topics for natural disasters or other hazards.
- Compilation: Yes
- Institutions
    - European Commission's Joint Research Centre
- Project Leaders:
    - Yet to be identified
- Data[60]Frequency: Real time
- License**:** Open Data
- Data Dictionary: Not found
- Protocol: Yet to be identified
- API Access: RSS[61]
- Geographic scope:
    - Coverage: World
    - Detail: Country
- Variables:
    - Title
    - Link
    - Description
    - Keywords
    - Publication date
    - ID
    - Source
    - Language
    - Categories
    - Geolocation
        - Country (derived)

6.7.7. Social Networks

- Sources
    - Twitter: Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read those that are publicly available.
    - Reddit: Reddit is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, images, and videos, which are then voted up or down by other members.

[60] https://medisys.newsbrief.eu/rss/?type=search&mode=advanced&atLeast=covid

[61] https://medisys.newsbrief.eu/rss/?type=search&mode=advanced&atLeast=covid

- PANDEM-2 use case: identify alerts on people's shared content by user defined topics extracting targeted information from message content using novel natural language processing algorithms, such as symptoms, suggestions, sentiment.
- Data:
    - Twitter[62]Reddit[63]License (end user agreement)
    - Twitter[64]Reddit[65]Frequency
    - Real time
- Data dictionary (see data)
- API Access: HTTP GET/POST
- Geographic scope:
    - City (using epitweetr)[66]Variables
    - Number of messages
    - Country
    - City
    - Suggestions
    - Sentiment
    - Symptoms
    - Emotions
    - Topic
    - Date

6.7.8. NGS Sequencing

- Sources:
    - NCBI - SRA:[67] SRA is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. It is produced by NCBI, The National Center for Biotechnology Information (US).
    - GISAID:[68] The GISAID Initiative promotes the rapid sharing of data from all influenza viruses and the coronavirus causing COVID-19. This includes genetic sequence and related clinical and epidemiological data associated with human viruses, and geographical as well as species-specific data associated with avian and other animal viruses, to help researchers understand how viruses evolve and spread during epidemics and pandemics. GISAID does so by overcoming disincentive hurdles and restrictions, which discourage or prevent sharing of virological data prior to formal publication. The Initiative ensures that open access to data in GISAID is provided free-of-charge to all individuals that

[62] https://developer.twitter.com/en/docs/twitter-api

[63] https://www.reddit.com/dev/api

[64] https://developer.twitter.com/en/developer-terms/agreement-and-policy

[65] https://www.redditinc.com/policies/user-agreement

[66] https://www.ecdc.europa.eu/en/publications-data/epitweetr-tool

[67] https://submit.ncbi.nlm.nih.gov/about/sra/

[68] https://www.gisaid.org/

agreed to identify themselves and agreed to uphold the GISAID sharing mechanism governed through its Database Access Agreement.

- PANDEM-2 use case: Monitor real time evolution of mutations and variants shared on public repositories by country and other shared metadata. Also, by adding additional fictitious metadata. PANDEM-2 will prove potential benefits of improving the quality and quantity of metadata.
- Data:
  - NCBI:[69]
  - GISAID: (to be identified)
- License
  - NCBI: Open data
  - GISAID: Database agreement[70]
- Frequency
  - Daily
- Data Dictionary
  - NCBI[71]GISAID[72]API Access
  - Yet to be identified
- Geographic scope
  - Country
- Variables: The following subset of variables has been identified to work on the first version due to the lack of standardisation of different data sources.
  - Bio Sample
    - Bio Sample Id
    - Collection Date
    - Location
    - Host
    - Host age
    - Host sex
    - prior vaccination status
    - prior infection status
  - NGS Raw data (Fastq file)
    - Sequence accession
    - Instrument
    - Strategy
    - Selection
    - Protocol
  - NGS Assembled data (consensus fasta sequence)
    - Nucleotide accession

[69] https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud/

[70] https://www.gisaid.org/about-us/acknowledgements/sharing-policy/

[71] https://www.protocols.io/view/guidance-for-populating-genometrakr-metadata-templ-bxcipiue

[72] https://github.com/CDCgov/SARS-CoV-2_Sequencing/blob/master/templates/20200315_EpiCoV_BulkUpload_Template.xls

        - Assembly method
    - NGS derived features
        - List of substitution
        - List of insertion/deletion
        - Clade
        - Genome quality score
        - Annotation algorithm
        - Reference genome
    - PCR Result
        - PCR main result
        - PCD Output ID
        - PCR protocol
        - PCR Ct

6.7.9. Surveillance Atlas of Infection diseases (TESSy)

- PANDEM-2 use case: Get official indicators about infectious diseases such as Influenza, Dengue or Ebola
- Compilation: Yes
- Institutions
    - ECDC
- Project Leaders:
    - Joana Gomes Dias
- Data[73]Frequency:
    - Dengue: Yearly (data prior to 2019)
    - Influenza: Weekly
    - Ebola: Yearly (data prior to 2019)
- License: European Commision reuse notice[74]Data Dictionary (Found in a metadataset downloaded on ECDC website[75], file: MetaDataset_48(2021-03-12).xlsx, sheet: Coded values)
- Protocol:
    - Tessy Protocol[76] (outdated) (see ecdc article[77])
- API Access:
    - HTTP (scraping)[78]
- Geographic scope:

[73] http://atlas.ecdc.europa.eu/public/index.aspx

[74] http://data.europa.eu/eli/dec/2011/833/oj

[75] https://www.ecdc.europa.eu/en/publications-data/tessy-metadata-report

[76]https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/1003_TER_TESSy_Transport_Protocol_CSV.pdf

[77] https://www.ecdc.europa.eu/en/publications-data/transport-protocol-specification-csv-comma-separated-value-tessy

[78] https://github.com/behrica/ecdcatlas/blob/master/R/getData.R

        - Coverage: Europe
        - Detail: Country
            - (Region level for the PlaceOfInfectionEVD variable. If the probable case of infection is not an EU/EEA country, then the NUTS nomenclature is used.)
    - Variables:
        - DENGUE and EBOLA Common variables:
            - RecordId
            - RecordType
            - RecordTypeVersion
            - Subject
            - Status
            - Datasource
            - ReportingCountry
            - DateUsedForStatistics
            - Age
            - Classification
            - ClinicalCriteria
            - DateOfDiagnosis
            - DateOfNotification
            - DateOfOnset
            - Gender
            - Imported
            - LaboratoryResult
            - Outcome
            - PlaceOfInfectionEVD
            - PlaceOfNotification
            - PlaceOfResidence
        - DENGUE specific variables (Record type DENGUE)
            - ClinicalManifestation
            - Hospitalization
            - LabMethod
            - Serotypes
        - EBOLA specific variables (Record type FILO)
            - EpiLinked
            - Pathogen
        - INFLUENZA
            - INFLUENZA Common variables
                - RecordType
                - RecordTypeVersion
                - Subject
                - Datasource

  - ReportingCountry
  - DateUsedForStatistics
- AH1N1 (Pandemic influenza, A(H1N1), RecordType AH1N1HAGGR) specific variables:
  - Age00-04
  - Age05-14
  - Age15-64
  - Age65+
  - AgeUnk
  - Deaths00-04
  - Deaths05-14
  - Deaths15-64
  - Deaths65+
  - DeathsUNK
  - NumberOfCases
  - NumberOfDeaths
  - TestAll (Are all suspect cases tested?)
- INFLANTIVIR (Influenza - Antiviral susceptibility data) specific variables:
  - RecordID
  - Age
  - AgeMonth
  - Amantadine
  - AntigenicGroup
  - Comment
  - CommentAG
  - CommentGC
  - ComplicationDiagnosis
  - ComplicationDiagnosisOther
  - DateOfOnset
  - ExposureDrug2weeksHouse
  - ExposureDrug2weeksHouseType
  - ExposureDrug2weeksPatient
  - ExposureDrug2weeksPatientType
  - Gender

  - GeneticClade
  - HAAAMutations
  - HAISD
  - Hospitalisation
  - ImmunoCompromised
  - IMOVE
  - InterpretationM2BlockerResistanceTesting
  - InterpretationOseltamivirResistanceTesting
  - InterpretationZanamivirResistanceTesting
  - M2AAMutations
  - M2ISD
  - NAAAMutations
  - NAISD
  - OseltamivirMUNANA
  - OseltamivirNAStar
  - Outcome
  - ProbableCountryOfInfection
  - Progress4weeks
  - Rimantadine
  - Subtype
  - VaccStatus
  - VirusCategoryIfNonSentinel
  - VirusSource
  - ZanamivirMUNANA
  - ZanamivirNAStar
- INFCLIN (Influenza - Clinical weekly data (ILI / ARI), RecordType INFLCLINAGGR) specific variables:
  - ARI_Denominator00-04
  - ARI_Denominator05-14
  - ARI_Denominator15-64
  - ARI_Denominator65+
  - ARI_DenominatorNumberOfCases
  - ARI_DenominatorUnk
  - ARI00-04
  - ARI05-14

  - ARI15-64
  - ARI65+
  - ARINumberOfCases
  - ARIUnk
  - CommentNonPublic
  - CommentPublic
  - GeographicSpread
  - ILI_Denominator00-04
  - ILI_Denominator05-14
  - ILI_Denominator15-64
  - ILI_Denominator65+
  - ILI_DenominatorNumberOfCases
  - ILI_DenominatorUnk
  - ILI00-04
  - ILI05-14
  - ILI15-64
  - ILI65+
  - ILINumberOfCases
  - ILIUnk
  - Impact
  - Intensity
  - NumberOfPhysicians
  - Trend
- INFLSARI (Influenza - SARI and fatal case data, Severe acute respiratory infections (SARI) case-based, RecordType INFLSARI) specific variables:
  - Age
  - AgeMonth
  - CauseOfDeath
  - Classification
  - ClinicalPresentation
  - ComplicationDiagnosis
  - ComplicationDiagnosisOther
  - DateOfDeath
  - DateOfHospitalDischarge

  - DateOfHospitalisation
  - DateOfNotification
  - DateOfOnset
  - DateOfTreatment
  - DateOfVacc
  - DrugUsedProphilaxis
  - DrugUsedTreatment
  - Gender
  - HospitalUnitType
  - Outcome
  - PlaceOfNotification
  - PlaceOfResidence
  - Precondition
  - PreconditionOther
  - Resistance
  - RespSupport
  - Subtype
  - TypeOther
  - VaccStatus
- INFLSARI (Influenza - SARI and fatal case data, Severe acute respiratory infections (SARI) aggregated, RecordType INFLSARIAGGR) specific variables:
  - DenomHospAdmissionsAge00-04STL
  - DenomHospAdmissionsAge05-14STL
  - DenomHospAdmissionsAge15-29STL
  - DenomHospAdmissionsAge15-64STL
  - DenomHospAdmissionsAge30-64STL
  - DenomHospAdmissionsAge65+STL
  - DenomHospAdmissionsSTL
  - DenomHospAdmissionsUnkSTL
  - DenomHospPopulationAge00-04STL
  - DenomHospPopulationAge05-14STL
  - DenomHospPopulationAge15-29STL
  - DenomHospPopulationAge15-64STL
  - DenomHospPopulationAge30-64STL

- DenomHospPopulationAge65+STL
- DenomHospPopulationUnkSTL
- DenomHospPopulationSTL
- DescriptionSARI
- NumSariDeathsAge00-04STL
- NumSariDeathsAge05-14STL
- NumSariDeathsAge15-29STL
- NumSariDeathsAge15-64STL
- NumSariDeathsAge30-64STL
- NumSariDeathsAge65+STL
- NumSariDeathsAgeUnkSTL
- NumSariHospitalisationsAge00-04STL
- NumSariHospitalisationsAge05-14STL
- NumSariHospitalisationsAge15-29STL
- NumSariHospitalisationsAge15-64STL
- NumSariHospitalisationsAge30-64STL
- NumSariHospitalisationsAge65+STL
- NumSariHospitalisationsAgeUnkSTL
- NumSariHospitalisationsDeathsSTL
- NumSariHospitalizationsSTL
- NumSariRepSitesSTL
- NumSpecimensAH1DetectSARI
- NumSpecimensAH1N1DetectSARI
- NumSpecimensAH3DetectSARI
- NumSpecimensAH3N2DetectSARI
- NumSpecimensAUnkDetectSARI
- NumSpecimensBDetectSARI
- NumSpecimensBVICDetectSARI
- NumSpecimensBYAMDetectSARI
- NumSpecimensRSVDetectSARI
- NumSpecimensSARSCoV2DetectSARIAge00-04
- NumSpecimensSARSCoV2DetectSARIAge05-14
- NumSpecimensSARSCoV2DetectSARIAge15-64
- NumSpecimensSARSCoV2DetectSARIAge65+

  - NumSpecimensSARSCoV2DetectSARIAgeUNK+
  - NumSpecimensSARSCoV2DetectSARITotal
  - NumSpecimensSWOAH1DetectSARI
  - NumSpecimensSWOAH1N1DetectSARI
  - NumSpecimensTotSARI
  - SARITestedSARSCoV2Age00-04
  - SARITestedSARSCoV2Age05-14
  - SARITestedSARSCoV2Age15-64
  - SARITestedSARSCoV2Age65+
  - SARITestedSARSCoV2AgeUNK+
  - SARITestedSARSCoV2Total

- INFLVIR (Influenza - Virological weekly data,RecordType INFLVIRWAGGR) specific variables :
  - agAH1/Guangdong-Maonan/SWL1536/2019
  - agAH1/Victoria/2570/2019
  - agAH1NOCAT
  - agAH3/Hong Kong/2671/2019
  - agAH3/Kansas/14/2017
  - agAH3NOCAT
  - agBVicB/Brisbane/60/2008
  - agBVicB/Colorado/06/2017
  - agBVicB/Washington/02/2019
  - agBVicNOCAT
  - agBYamB/Phuket/3073/2013
  - agBYamNOCAT
  - CommentAG
  - CommentGC
  - CommentNonPublic
  - CommentPublic
  - CommentWHO
  - DominantType
  - genAH1/Guangdong-Maonan/SWL1536/2019
  - genAH1/Norway/3433/2018
  - genAH1/Slovenia/1489/2019

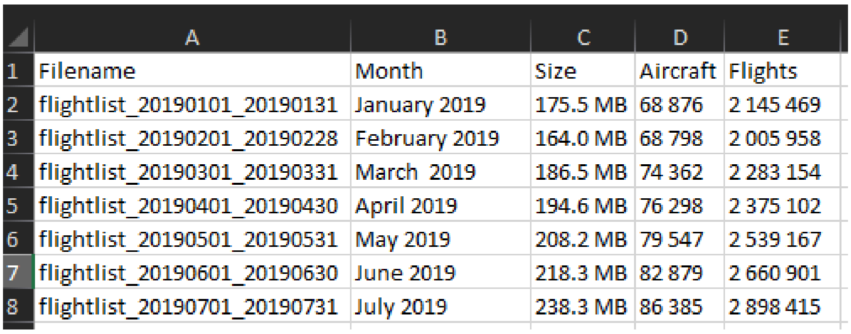
- genAH1/Switzerland/3330/2018
- genAH1/Victoria/2570/2019
- genAH1NOClade
- genAH1SubgroupNotListed
- genAH3/Bretagne/1323/2020
- genAH3/Denmark/3264/2019
- genAH3/Hong Kong/2671/2019
- genAH3/Kansas/14/2017
- genAH3/Slovenia/1637/2020
- genAH3NOClade
- genAH3SubgroupNotListed
- genBVicB/Colorado/06/2017
- genBVicB/Washington/02/2019
- genBVicCladeB/Brisbane/60/2008
- genBVicNOClade
- genBVicSubgroupNotListed
- genBYamB/Phuket/3073/2013
- genBYamNOClade
- genBYamSubgroupNotListed
- NumPatientsTestedAH5N1OTH
- NumPatientsTestedAH5N1STL
- NumSpecimensAH1DetectOTH
- NumSpecimensAH1DetectSTL
- NumSpecimensAH1N1DetectOTH
- NumSpecimensAH1N1DetectSTL
- NumSpecimensAH3DetectOTH
- NumSpecimensAH3DetectSTL
- NumSpecimensAH3N2DetectOTH
- NumSpecimensAH3N2DetectSTL
- NumSpecimensAH5DetectOTH
- NumSpecimensAH5DetectSTL
- NumSpecimensAH5N1DetectOTH
- NumSpecimensAH5N1DetectSTL
- NumSpecimensANTypableOTH

  - NumSpecimensANTypableSTL
  - NumSpecimensAUnkDetectOTH
  - NumSpecimensAUnkDetectSTL
  - NumSpecimensBDetectOTH
  - NumSpecimensBDetectSTL
  - NumSpecimensBVICDetectOTH
  - NumSpecimensBVICDetectSTL
  - NumSpecimensBYAMDetectOTH
  - NumSpecimensBYAMDetectSTL
  - NumSpecimensOtherDetectOTH
  - NumSpecimensOtherDetectSTL
  - NumSpecimensRSVDetectOTH
  - NumSpecimensRSVDetectSTL
  - NumSpecimensRSVTotOTH
  - NumSpecimensRSVTotSTL
  - NumSpecimensSARSCoV2DetectSTL
  - NumSpecimensSWOAH1DetectOTH
  - NumSpecimensSWOAH1DetectSTL
  - NumSpecimensSWOAH1N1DetectOTH
  - NumSpecimensSWOAH1N1DetectSTL
  - NumSpecimensTestedSARSCoV2STL
  - NumSpecimensTotOTH
  - NumSpecimensTotSTL
  - SpecimensOtherCommentOTH
  - SpecimensOtherCommentSTL

6.7.10. Mobility

- OpenSky Network
  - PANDEM-2 use case: Get an estimation of the flight number between countries in order to measure the impact of an epidemic on air traffic
  - Compilation: Yes
  - Institutions
    - OpenSky Network association
  - Project Leaders:
    - Martin Strohmeier

- Data[79] (prepared datasets[80])
  - Dataset structure expected: country of origin, destination country, date of flight
    - Infos about Data collection here: Impact of COVID-19 on worldwide aviation — traffic documentation[81]
    - List of files that is supposed to list flight traffic each month[82]

- 

	A	B	C	D	E
1	Filename	Month	Size	Aircraft	Flights
2	flightlist_20190101_20190131	January 2019	175.5 MB	68 876	2 145 469
3	flightlist_20190201_20190228	February 2019	164.0 MB	68 798	2 005 958
4	flightlist_20190301_20190331	March 2019	186.5 MB	74 362	2 283 154
5	flightlist_20190401_20190430	April 2019	194.6 MB	76 298	2 375 102
6	flightlist_20190501_20190531	May 2019	208.2 MB	79 547	2 539 167
7	flightlist_20190601_20190630	June 2019	218.3 MB	82 879	2 660 901
8	flightlist_20190701_20190731	July 2019	238.3 MB	86 385	2 898 415

- Frequency:
  - Real time: for free
  - Historical data: available for academics or after registration (+ license)
- License:
  - Real time: Open data
  - Historical: Paid license for non-academic use
- Data Dictionary
  - Not yet identified
- Protocol[83]
- API Access: Git[84]
  - openskynetwork/opensky-api: Python and Java bindings for the OpenSky Network REST API[85]
  - The OpenSky Network API documentation — The OpenSky Network API 1.4.0 documentation[86]
- Geographic scope:
  - Coverage: World (only ADS-B-equipped aircraft)
  - Detail: airport cities

[79] https://opensky-network.org/

[80] https://opensky-network.org/datasets/

[81] https://traffic-viz.github.io/scenarios/covid19.html

[82] https://zenodo.org/record/3928550#.YUtNx7gzbce

[83] https://essd.copernicus.org/articles/13/357/2021/essd-13-357-2021.html

[84] https://github.com/openskynetwork

[85] https://github.com/openskynetwork/opensky-api

[86] https://opensky-network.org/apidoc/

- Variables: Crowdsourced air traffic data from the OpenSky Network 2019–2020[87] . Datasets completed every month with the following variables should be reachable:
    - callsign
    - number
    - aircraft_uid
    - typecode
    - origin
    - destination
    - firstseen
    - lastseen
    - day
    - latitude_1, longitude_1, altitude_1
    - latitude_2, longitude_2, altitude_2

6.8. Final list of sources

Based on the groups of data families identified on the data survey and the data availability we have chosen the final list sources for PANDEM-2. The following rationale was followed.

- Important and available (Cases, Deaths, vaccination, patient, tests, Lab)

A majority of the variables on these data families can be obtained either from public sources or by aggregating data available from end users. The following strategy will be used

- Open data when available
- if not, user provided data (aggregated) when available
- if not synthetic data (not necessarily realistic, mainly to shown dashboard functionalities)

- Important with limited availability (NGS, Contact tracing, Population study)

Different approaches were taken from each of these data sources

- NGS: Realistic simulated data will be produced using a dedicated algorithm capable of using data from NGS public repositories in order to estimate figures not publicly available such as the number of cases by mutation/variant and age group.
- Contact Tracing: Following ECDC guidelines we will include aggregated indicators for measuring contact tracing performance. Since this data is hard to obtain for public agencies, we will produce this data using randomised extracts from GO.Data.
- Population studies: Because of the big variability in studies we will use machine learning algorithms to extract topics, suggestions, emotions and

[87] https://essd.copernicus.org/articles/13/357/2021/essd-13-357-2021.pdf

sentiment coming from social network and mass media news using novel machine learning approaches.

- Lack of data or low priority: First response (including staff), transport, referentials, measures (including flights), emergency calls, resources
  - These sources will be included only when public sources are found but with less priority
  - Concerning resources, they are going to be entered as manual input for resource planning models.
- Not in requirements: Weather, seroprevalence
  - These sources will be included only when public sources are found but with less priority. Currently sero prevalence was integrated using SeroTracker, but weather is currently out of scope. This decision was taken in coordination with internal experts on the consortium and the project executive team.

After applying the criterion mentioned above, we obtain the following list of sources

Subtype (Data Family)	Source	Sample variables	Type
Cases, Deaths, vaccination, patient, tests, Lab	Covid19Datahub[88] (similar to JHK covid-19 dataset but with finer granularity)	Per NUTS3: Number of cases Active Cases At Least One Dose Vaccinated Confirmed Cases InfectedDeaths Doses Injected Icu Patients Number Of Hospitalised Patients Number Of Icu Patients Number Of Patients With Ventilator People Fully VaccinatedPerformed Tests Rt Number	Open data
	ECDC (COVID19)	At national level: Number of cases per age group Number of cases per variant	Open data

[88] This table was added after the project midterm review and reflects a more advanced status than the rest of the document

	User data	Per NUTS3: Deaths in ltcf, Positivity rate, cases per variant	Users' data (public)
	ECDC (ATLAS)	Number of cases with influenza	Open data
	Ourworld in data	By country Excess mortality	Open data
	Synthetic data	Per NUTS3: Mortality rate at ICU, Hospitalizations by age group, hospitalizations by comorbidities, tests performed per type	Synthetic
Government measures (First response)	ECDC (COVID19)	At national level: Lockdown active, teleworking active, mass gathering restrictions	Open data
NGS	ECDC (Simulated)	At national level: Number of cases per variant and age group Number of cases per mutation and age group	Realistic simulation
Mass media (population study)	Twitter	At NUTS3 level: Number of posts per topic, suggestion, sentiment, emotion	Open data + Algorithm estimation
	Reddit	We intend to apply developed social media analysis (SMA) components on this data source in the future	-
	Medisys	At NUTS3 level Number of posts per topic, suggestion, sentiment, emotion	Open data + Algorithm estimation
Population surveys (population study)	Synthetic	At NUTS3 level: Number of surveys	Synthetic

		conducted People respecting policy X	
Participatory surveillance	Influenza.net	Number of participants declaring symptoms, number of participants visiting GP	Open data
Contact Tracing	Go Data	At national level: contacts found in Contact tracing, Current contact tracing policy Total cases that previously had been identified as contact	Realistic Simulation
Flights (measures)	Open Sky network covid19	Sampled number of flights entering to a country	Open data
Seroprevalence studies (population study)	SeroTracker	Per NUTS3 Sero positivity per age-groups	Open data
Resources (first response)	Parametric	Number of available ventilators, number of available PPE	manually introduced