Deliverable D1.2

Preparedness Data Hub

Project Title (grant agreement No)	BeYond-COVID Grant Agreement 101046203		
Project Acronym (EC Call)	BY-COVID		
WP No & Title	WP1: Support for virological analyses in emerging disease outbreaks		
WP Leaders	Guy Cochrane (EMBL-EBI), Clara Amid (Erasmus MC)		
Deliverable Lead Beneficiary	EMBL-EBI		
Contractual delivery date	30/09/2023	**Actual Delivery date**	31/10/2023
Delayed	Yes - 1 month delay		
Partner(s) contributing to this deliverable	EMBL-EBI, Erasmus MC, DTU, ELTE		
Authors	Nadim Rahman (EMBL-EBI), Colman O’Cathail (EMBL-EBI), Hannah-Marie Martiny (DTU), Marianna Ventouratou (EMBL-EBI), Zahra Waheed (EMBL-EBI), Clara Amid (EMC)		
Contributors			
Acknowledgements (not grant participants)			
Reviewers	Aitana Neves (SIB) Mari Kleemola (TAU-FSD)		

Disclaimer
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.

Funded by
the European Union

Table of contents

Funded by
the European Union

1. Executive Summary

The scope of the deliverable falls under Work Package Task 1.3 “Rapid deployment of the “preparedness” Data Hub” and the related subtasks for developing the tools (technical implementation) to allow the rapid deployment and configuration of a disease X scenario preparedness Data Hub. The system is intended to allow the rapid configuration of functions from a checklist of technical elements, including viral biology (genome browser, related viruses), surveillance (upload, data standards, integration tools), cohort data capabilities, computational processing and analytical workflows, Notebook visualisation, variation discovery and impact prediction and phylogeography.

In the past 24 months, work on this deliverable included the development of pathogen data classification based on taxonomy and tagging within the Pathogens Portal. To ensure consistency in pathogens classification, we adopted the UK’s Health and Safety Executive’s (HSE) list of approved biological agents which provides a definitive list on what constitutes a pathogen. At the same time we plan to expand beyond pathogens affecting humans, to plants for example. The accompanying tagging system is a simple ‘tag=pathogen’ query which overcomes the need to specify a very large number of taxonomic IDs. As part of pandemic preparedness, an Outbreaks page was developed within the Pathogens Portal to identify pathogens that can cause outbreaks or pandemics.

To better support users submitting pathogen data to a Data Hub, we developed a dedicated Pathogens Submission Guide which includes a list of six pathogen sample checklists, spanning prokaryotes, parasites and viruses. In addition, we maintain a helpdesk queue for submission-related queries, and continuation of the Contextual Data Clearing House, which allows the scientific community to extend or better-annotate pathogen metadata, such as via a Data Hub. WP1 partners The Arctic University of Norway (UiT) submitted a valuable dataset of over 27 million SARS-CoV-2 curations allowing us to identify areas of further development for the Clearing House.

As part of the analysis pipeline exploration task, we tested a pipeline for antimicrobial resistance (AMR). WP1 partner DTU developed an AMR pipeline, called ARGprofiler (antimicrobial resistance genes) which was explored for its potential to be integrated within the Pathogens Platform, constituting part of a preparedness Data Hub. In addition, two further community developed pipelines (Bactopia and nf-core/funcscan) are being benchmarked by EMBL-EBI and assessed for integration into the Data Hubs system. Furthermore, publicly developed and maintained viral metagenomic pipelines are being assessed for integration into the Data Hub system.

Finally, as part of the development of visualisation tools, Nextstrain, an open source project supporting real-time tracking of pathogen evolution was integrated in mid-2023. It includes Mpox (Monkeypox), Zika and West Nile Virus reports, and allows configuration (in the integration) of the visualisation results through various facets, including a phylogenetic tree, a map of the geographical distribution of the sequences and thor clade classification, as well as a genome browser presenting viral diversity. Note these reports are currently running for public data, and are not available for pre-publication/private data.

2. Contribution towards project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

	Key Result No and description	Contributed
Objective 1 Enable storage, sharing, access, analysis and processing of research data and other digital research objects from outbreak research	1.A research data management practice in European research infrastructures practice that drives discovery, access and reuse of outbreak data and directly links experimental data from HORIZON-INFRA-2021-EMERGENCY-02 transnational access projects into the COVID-19 Data Portal.	Yes
	2. Workflows and processing pipelines that integrate transparent quality management and provenance and are openly shared.	Yes
	3. Research infrastructures on-target training so that users can exploit the platform	No
	4. Engagement so that stakeholders (RI, national centres, policy makers, intergovernmental organisations, funders and end-users) incorporate FAIR and open data in infectious disease guidelines and forward planning.	No
Objective 2 Mobilise and expose viral and human infectious disease data from national centres	1. A comprehensive registry of available data with established procedures to collate data governance models, metadata descriptions and access mechanisms in a pandemic scenario.	Yes
	2. Mechanisms for the initial discovery across data sources based on available metadata at the reference collection.	Yes
	3. Demonstrated transnational linking of real-world data from national surveillance, healthcare, registries and social science data that allow the assessment of variants to serve the research needs of epidemiology and public health.	Yes
	4. Demonstrated assessment of emerging SARS-CoV-2 variants against data generated in the on-going European VACCELERATE clinical trials project to investigate vaccine efficacy.	No

Funded by
the European Union

Objective 3 Link FAIR data and metadata on SARS-CoV-2 and COVID-19	1. A platform that links normative pathogen genomes and variant representations to research cohorts and mechanistic studies to understand the biomolecular determinants of variant response on patient susceptibility, and disease pathways.	Yes
	2. An open and extensible metadata framework adopted cross-domain that supports comprehensive indexing of the infectious disease resources based on mappings across resources and research domains.	Yes
	3. A provenance framework for researchers and policy-makers that enables trust in results and credit to data submitters, workflow contributors and participant resources.	Yes
Objective 4 Develop digital tools and data analytics for pandemic and outbreak preparedness, including tracking genomics variations of SARS-CoV-2 and identifying new variants of concern	1. Broad uptake of viral *Data Hubs* across Europe deliver an order-of-magnitude increase in open viral variant detection and sharing.	Yes
	2. Infrastructure and quality workflows mobilised and shared to produce open, normative variant data that is incorporated into national and regional data systems and decision making.	Yes
Objective 5 Contribute to the Horizon Europe European Open Science Cloud (EOSC) Partnership and European Health Data Space (EHDS)	1. Guidelines and procedures for FAIR data management and access will be established, building on work of other guideline producing consortia such as the Global Alliance for Genomics and Health (GA4GH), the 1Mio Genomes Initiative (1MG) and the Beyond One Million Genomes project (B1MG).	No
	2. Services, software, protocols, guidelines and other research objects that are openly accessible for reuse by the EOSC Association and the community at large as a foundation for European preparedness for infectious diseases, leveraging developments in EOSC-Life, SSHOC, EOSC-Future, EGI-ACE and other EOSC projects.	Yes

	3. Alignment (both policy and implementation routes) will have been achieved between the data governance strategies for routinely collected health data in the EHDS initiative, including the TEHDAS Joint Action and future EHDS Pilot Actions.	No
	4. To empower national centres to build capacity and train platform users and data providers (e.g., from life, social or health sciences), and with experts from across partner institutions collaborating to create training materials for the identified gaps, and to exchange experiences and knowledge.	Yes

3. Methods

This deliverable details tools and extensions that together, allow for the rapid deployment and configuration of a disease X scenario preparedness Data Hub. This work is part of the Pathogens Platform, which includes the Pathogens Portal - supported by both WP1 and WP3 in BY-COVID, but also other projects, listed in the funders page here: https://www.pathogensportal.org/funding-projects.

For preparedness, work under this deliverable focussed on a set of priority pathogens that are most likely to be the cause of a future outbreak/epidemic/pandemic.

4. Description of work accomplished

4.1 Preparedness Data Hubs

Preparedness Data Hubs use the concept developed under the COMPARE Data Hubs[1], and SARS-CoV-2 Data Hubs[2]. This includes a set of **submission**, **analysis**, **visualisation** and **presentation** tools to help users in sharing, analysing, retrieving and interpreting sequence data amongst a group of collaborators. This can be done in pre-publication, i.e. private, or public modes. By using this concept, we aim to provide a set of tools spanning these areas in order to support researchers in preparedness, focusing on a select group of priority pathogens that are likely to cause the next infectious disease outbreak. In some cases, such as submission tools, tools are relatively generalised to support sequence data

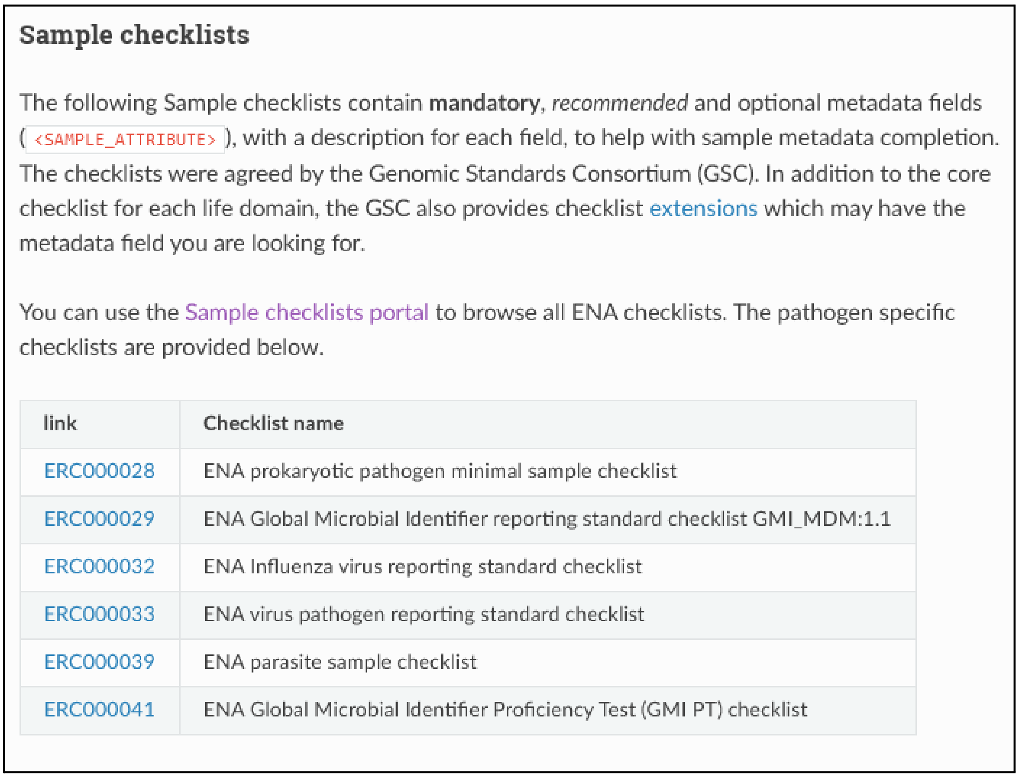
[1] COMPARE Data Hubs: https://academic.oup.com/database/article/doi/10.1093/database/baz136/5685390 [accessed 31.10.2023]

[2] SARS-CoV-2 Data Hubs: https://www.covid19dataportal.org/data-hubs [accessed 31.10.2023]

submissions of any pathogen. However for the majority of aspects, we will focus on a narrower set of pathogens to help tailor a user's experience and provide more appropriate tools. Below we delve into each of the main aspects mentioned above in greater detail.

4.2 Submissions

To provide more tailored support to users submitting pathogen data to a Data Hub, a dedicated Pathogen's Submission Guide[3] (https://ena-docs.readthedocs.io/en/latest/faq/pathogen-subs-guide.html#general-pathogens-submissions-guide) was created, bringing together both new and existing documentation. New guidance includes a list of 6 pathogen sample checklists[4] (fig.1) spanning prokaryotes, parasites and viruses (all developed in conjunction with the Genomic Standards Consortium, COMPARE project or Global Microbial Identifier), as well as information on completing host metadata fields, including for samples obtained from cell lines.



Sample checklists

The following Sample checklists contain **mandatory**, *recommended* and optional metadata fields (`<SAMPLE_ATTRIBUTE>`), with a description for each field, to help with sample metadata completion. The checklists were agreed by the Genomic Standards Consortium (GSC). In addition to the core checklist for each life domain, the GSC also provides checklist extensions which may have the metadata field you are looking for.

You can use the Sample checklists portal to browse all ENA checklists. The pathogen specific checklists are provided below.

link	Checklist name
ERC000028	ENA prokaryotic pathogen minimal sample checklist
ERC000029	ENA Global Microbial Identifier reporting standard checklist GMI_MDM:1.1
ERC000032	ENA Influenza virus reporting standard checklist
ERC000033	ENA virus pathogen reporting standard checklist
ERC000039	ENA parasite sample checklist
ERC000041	ENA Global Microbial Identifier Proficiency Test (GMI PT) checklist

Figure 1. *Screenshot of sample checklists on the Pathogen's Submission Guide*

[3] Pathogen's Submission Guide: https://ena-docs.readthedocs.io/en/latest/faq/pathogen-subs-guide.html#general-pathogens-submissions-guide [accessed 31.10.2023]

[4] Pathogen sample checklists: https://ena-docs.readthedocs.io/en/latest/faq/pathogen-subs-guide.html#sample-checklists [accessed 31.10.2023]

Funded by
the European Union

To supplement this we also continue to actively maintain the ena-pathogen help desk queue, which now sees 20 tickets a month (an average reduction of 3 tickets per month from 2022 and a reduction of 24 tickets per month in 2021, due to better documentation, guides, training material, as well as improved general functionality) with queries ranging from not only data submission, but post-submission processing, data access/retrieval, updates and removal.

The Contextual Data Clearing House (CDCH)[5] is a further development which offers a way for the scientific community to extend pathogen (and other types of) metadata after submission to the International Nucleotide Sequence Database Collaboration (INSDC), such as via a Data Hub. ‘Extending’ refers to curations such as: correcting inaccuracies in metadata, adding additional metadata fields, or adding ontological terms, and all curations will present alongside the record (without permanent modification of it). The Clearing House acts as a ‘store’ for these curations, and has an API component[6] for users to submit to, along with an assertion method and provenance information. Curations can also be submitted by individuals who are not the original submitters of the data.

The first SARS-CoV-2/pathogen curations were submitted to CDCH by the Arctic University of Norway (UiT), reaching 27,566,814 SARS-CoV-2 curations by 2022. This was a valuable dataset which allowed us to identify further areas of development for the Clearing House. Firstly through a collaborative mapping exercise between UiT-generated curations and existing metadata within INSDC records, we noticed some redundancy where some curations attributes were identical to those in the INSDC record, despite not being fully INSDC-compliant. We thus identified the need to implement a validation component within the CDCH system to prioritise value-add curations, and to create user documentation for this service. The former is currently in progress and the latter planned for the future. Other outcomes from this collaboration were the creation of an ENA Clearing House taskforce, and a roadmap for wider CDCH development (spanning presentation, API functionality and searchability).

4.3 Analysis

4.3.1 Antimicrobial Resistance

An AMR pipeline developed by DTU was explored in its potential integration within the Pathogens Platform, constituting part of a preparedness Data Hub. The pipeline is called ARGprofiler, and the aim is to retrieve metagenomic sequencing datasets from ENA and

[5] ELIXIR Contextual Data ClearingHouse webpage:
https://elixir-europe.org/internal-projects/commissioned-services/establishment-data-clearinghouse [accessed 31.10.2023]

[6] Contextual Data ClearingHouse
https://www.ebi.ac.uk/ena/clearinghouse/api/swagger-ui/index.html [accessed 31.10.2023]

analyse the presence, abundance, and genomic locations of antimicrobial resistance genes (ARGs). ARGprofiler has been designed to be a robust, efficient, and fast workflow for processing large quantities of metagenomic sequencing reads. Sequencing reads are downloaded from ENA and undergo quality checking and trimming before being used for three downstream tasks. The first task is the mapping and alignment of the reads against reference sequence databases. Secondly, the trimmed reads are used to create targeted de-novo assemblies by including a new approach in the pipeline, which uses the ARGs as seeds and assembles the surrounding genome. Thirdly, Mash sketches are produced to allow the re-use of the metagenomic reads for other tasks unrelated to AMR. Each step of the pipeline was carefully evaluated to optimise computational requirements and usability, as the initial exploration revealed several significant computational bottlenecks. Testing state-of-the-art tools and reference databases greatly reduced the impact of these identified bottlenecks, making the pipeline as streamlined as possible. ARGprofiler provides a simple solution to integrate metagenomic datasets into AMR surveillance and pandemic preparedness. Recommendations were discussed around improvements and adaptability to support full integration into the platform. A manuscript detailing the ARGprofiler pipeline is currently undergoing review in the journal *Bioinformatics*; however, the pipeline has already been made accessible at https://github.com/genomicepidemiology/ARGprofiler.

The need for additional pipeline exploration and support is clear, as such EMBL-EBI has begun benchmarking of community developed and widely used AMR pipelines. Two such pipelines Bactopia[7] and nf-core/funcscan[8] have been benchmarked to date. The benchmarking approach is summarised as follows; for each of the bacterial pathogens recognised by the WHO as posing a high risk due to AMR, 100 samples were randomly selected for each pathogen and run through each pipeline. The pipelines are then assessed for various factors including compute resources used, total runtime and AMR profiles produced.

Each pipeline takes different inputs. Bactopia, takes raw read data as an input. nf-core/funcscan by contrast takes assembled genomes as input, which require assembly into a genome as a consequence. However, both pipelines showed good performance and reproducible results across the pathogens selected for benchmarking. Thus, EMBL-EBI feels it is appropriate to scope both for integration into the Data Hubs model, as each offers similar outputs depending on what data type the user starts with (raw reads or assemblies). At this stage, EMBL-EBI is currently piloting integration of these pipelines into the Data Hubs system, which should lead to a model framework for integrating more pipelines in the future.

[7] Bactopia pipeline: https://bactopia.github.io/v3.0.0/ [accessed 31.10.2023]
[8] nf-core/funcscan pipeline: https://nf-co.re/funcscan/1.1.3 [accessed 31.10.2023]

4.4 Visualisations

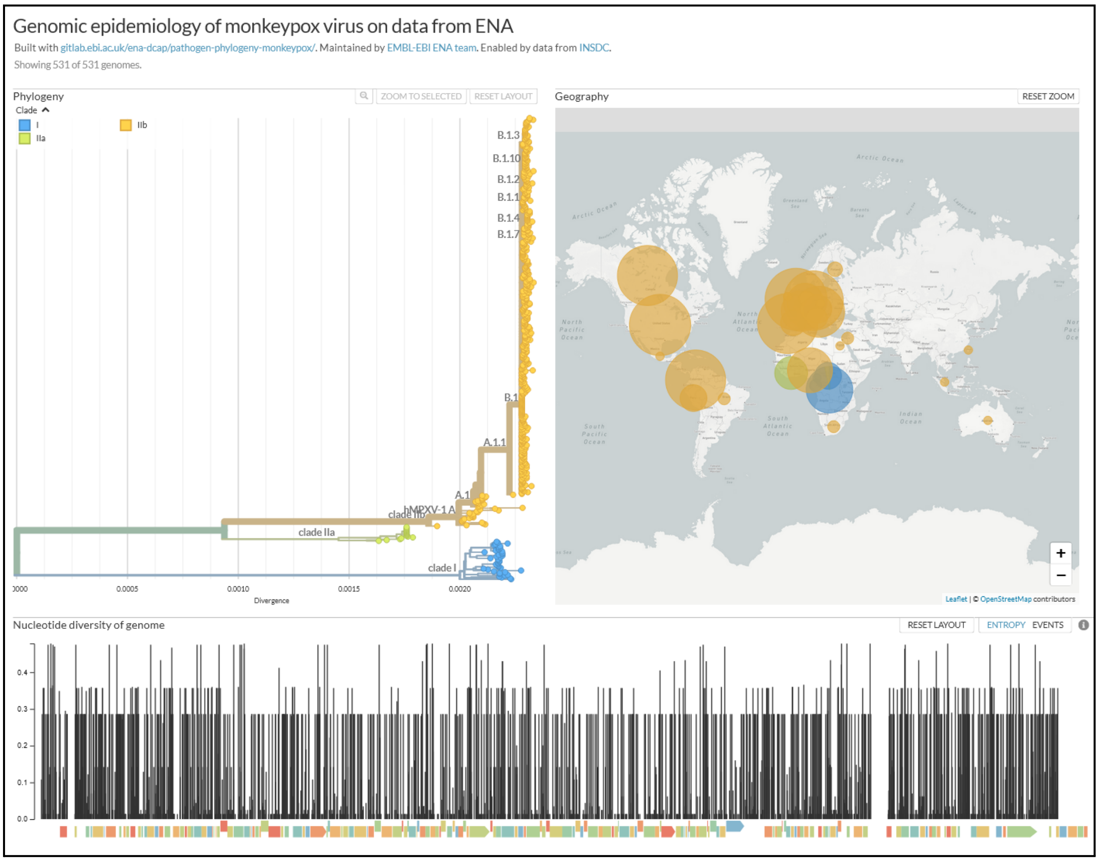
4.4.1 Nextstrain Reports

Nextstrain[9], an open-source project supporting real-time tracking of pathogen evolution, was successfully integrated mid-2023. The initial use-case to help drive the integration of the tool, along with its accompanying packages (such as augur and auspice), centred around Monkeypox (MPox). However since then, two new reports have been added to the integrated Nextstrain report: https://www.pathogensportal.org/priority-pathogens. Click the 'Nextstrain Reports' tab to view the reports. The current integrated versions are not fully automated to pick up incoming sequences in real-time from the relevant taxa but automation is under consideration. The reports can be however updated if needed.

The reports (fig.2) present a phylogenetic tree of analysed sequences (for MPox) and submitted sequences (for West Nile Virus - WNV- and Zika) based on data holdings at the ENA. MPox includes analysed sequences from raw reads as part of the repurposing of COVID -19 workflows previously carried out. The reports include customisability in presentation through the facets, e.g. colour-coding by clades, country of origin, etc. that is provided by Nextstrain and enabled within the integration. Along with a phylogenetic tree, a map presents the geographic distribution of sequences, along with their clade classifications. Scrolling towards the very bottom, a genome browser can be found, presenting the observed diversity across the data.

[9] Nextstrain: https://nextstrain.org/ [accessed 31.10.2023]

Funded by
the European Union



Genomic epidemiology of monkeypox virus on data from ENA

Built with gitlab.ebi.ac.uk/ena-dcap/pathogen-phylogeny-monkeypox/. Maintained by EMBL-EBI ENA team. Enabled by data from INSDC.

Showing 531 of 531 genomes.

Figure 2. *Screenshot of the Nextstrain report of monkeypox virus on the Pathogens Portal*

The exploration highlighted some issues, which were overcome following extensive work. One of the issues included difficulties in mapping URLs for reports to public URLs associated with the portal. This was overcome through deploying the nextstrain report code within the existing Kubernetes cluster used by the ENA browser. Due to the lightweight nature of these applications, they did not need a dedicated web VM and did not take much resource away from the ENA browsers Kubernetes cluster. This has exposed a useful route for rapidly deploying lightweight web applications in the future.

As mentioned above, currently, there are three reports for three separate organisms - MPox, WNV and Zika. This showcases Nextstrain's flexible utility where the integrated instance does not require extensive changes to add reports. The addition occurs through Nextstrain 'dataset' processing, so long as Nextstrain-analyses already exist (these are spun up by the Nextstrain team relatively quickly in the case of novel outbreaks) to which augur and auspice are used for integrated reports. This functionality is possible thanks to the ongoing support and maintenance for Nextstrain going into the future, providing a sustainable option for an integrated visualisation in the Pathogens Portal.

Funded by
the European Union

4.5 Presentation

4.5.1 Pathogen Data Classification

Taxonomy

The definition of what constitutes a pathogen, and then applying this definition posed some challenges in the resulting Pathogens Portal. First, to various researchers, health officials and individuals, the definition of a pathogen often differs. A specific pathogen can also be missed out within any list. There are many sources of information, so the question then becomes how best to choose a list to define pathogens. Prior to the update of the Pathogens Portal (and official launch), the definition of a pathogen was very broad, it included all bacteria, viruses, fungi and some eukaryotes. However, not all of the organisms underneath these broad domain categorisations are pathogens, this is at least the case for bacteria, eukaryotes and fungi. The definition of a pathogen also came up during the BY-COVID mid-term review, so we have taken on board some of the comments while developing further the current Pathogens Portal.

Therefore, a more holistic definition was taken from the UK Health and Safety Executive (UKHSE) which provides a definitive list of approved biological agents[10]. This is a core reference that can be kept up-to-date on what constitutes a pathogen in the Pathogens Portal. We do recognise that there may be other resources and also other pathogens defined, which has prompted planning for the ability for users to request addition of specific pathogens (through the National Center for Biotechnology Information (NCBI) Taxonomy[11]) to the list of pathogens used to define a pathogen. Overall ,this will help provide a more comprehensive portal, and appeal to a broader group of users, who are part of various pathogen communities.

The pathogens defined in the portal are those that are pathogenic to humans. In order to encourage collaborations, support other communities, and expand the portal, for example to plant pathogens, or others, we will seek to explore opportunities in this area.

As part of the Pathogens Portals ongoing efforts to improve transparency, the list of pathogens used to populate the portal is also available within the portal as a set of tables: https://www.pathogensportal.org/pathogen-classifications., and it can be retrieved from GitHub[12]. These tables include a “Source” column, attributing the external source/authority from which the pathogen definitions are derived.

[10] UKHSE list of approved biological agents: https://www.hse.gov.uk/pubns/misc208.pdf [accessed 31.10.2023]
[11] NCBI Taxonomy: https://www.ncbi.nlm.nih.gov/taxonomy [accessed 31.10.2023]
[12] Ena-content-dataflow GitHub: https://github.com/enasequence/ena-content-dataflow/tree/master/classifications/pathogen_taxa [accessed 31.10.2023]

Tagging System

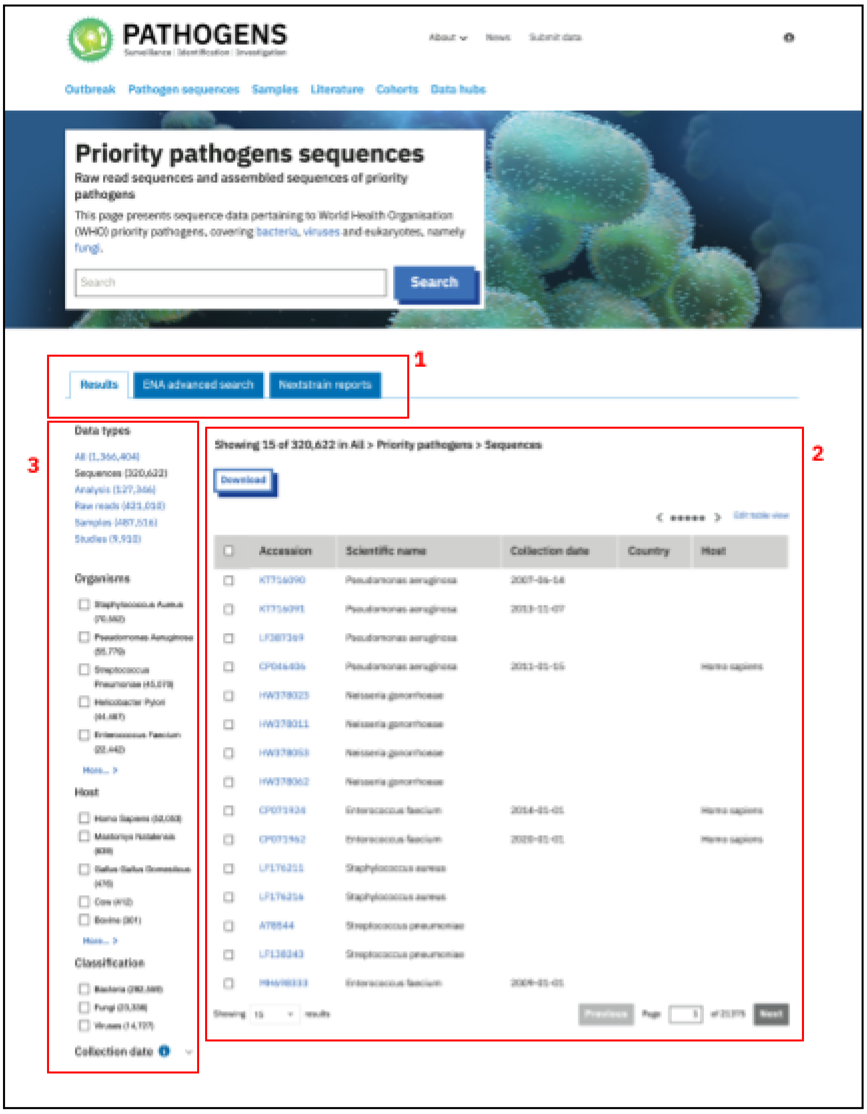
The list of pathogens described in 4.5.1, were collated into a CSV file (in the GitHub link above). This enables the European Nucleotide Archive (ENA) to tag datasets with the term 'pathogen' in the backend ElasticSearch system that supports presentation services with the Pathogens Portal. This tagging mechanism has been applied to the EBI Search indexing system[13,14] that is behind the availability of data in the COVID-19 Data Portal,[15] and now the Pathogens Portal. The beauty of this system is that behind the scenes a simple 'tag=pathogen' query can be done to pull all pathogenic records, without the need to specify a very large number of taxonomic IDs. This in turn, enables for a quick method in serving appropriate sequence data to the Pathogens Portal. Note, the system also enables datasets to have multiple tags, e.g. 'pathogen', 'datahub', the latter being used for data associated with specific Data Hubs.

4.5.2 Outbreaks Page

An 'Outbreaks' page has been developed as part of the Pathogens Portal: https://www.pathogensportal.org/priority-pathogens. This page provides a filtered view of the entire portal on a select group of pathogens that have been identified as potentially causing the next epidemic or pandemic (fig.3). As shown by the figure, the format remains very similar to the other pages on the portal, and to the COVID-19 Data Portal, as it utilises the same web framework and components.

[13] EBI Search: https://www.ebi.ac.uk/ebisearch/about [accessed 31.10.2023]
[14] EBI Search data coverage: https://www.ebi.ac.uk/ebisearch/data-coverage [accessed 31.10.2023]
[15] COVID-19 Data Portal: https://www.covid19dataportal.org/ [accessed 31.10.2023]

Funded by
the European Union



Priority pathogens sequences

Raw read sequences and assembled sequences of priority pathogens

This page presents sequence data pertaining to World Health Organisation (WHO) priority pathogens, covering bacteria, viruses and eukaryotes, namely fungi.

Data types

All (1,366,404)
Sequences (320,622)
Analysis (127,366)
Raw reads (421,010)
Samples (487,516)
Studies (9,910)

Organisms

□ Staphylococcus Aureus (70,662)
□ Pseudomonas Aeruginosa (55,770)
□ Streptococcus Pneumoniae (45,078)
□ Helicobacter Pylori (44,487)
□ Enterococcus Faecium (22,442)
More... >

Host

□ Homo Sapiens (32,050)
□ Mastomys Natalensis (839)
□ Gallus Gallus Domesticus (470)
□ Cow (412)
□ Bovine (301)
More... >

Classification

□ Bacteria (282,449)
□ Fungi (23,338)
□ Viruses (14,727)

Collection date

Showing 15 of 320,622 in All > Priority pathogens > Sequences

□	Accession	Scientific name	Collection date	Country	Host
□	KT716090	Pseudomonas aeruginosa	2007-06-18		
□	KT716091	Pseudomonas aeruginosa	2013-11-07		
□	LF387369	Pseudomonas aeruginosa			
□	CP046406	Pseudomonas aeruginosa	2011-01-15		Homo sapiens
□	HW378023	Neisseria gonorrhoeae			
□	HW378011	Neisseria gonorrhoeae			
□	HW378053	Neisseria gonorrhoeae			
□	HW378062	Neisseria gonorrhoeae			
□	CP071924	Enterococcus faecium	2014-01-01		Homo sapiens
□	CP071962	Enterococcus faecium	2020-01-01		Homo sapiens
□	LF176211	Staphylococcus aureus			
□	LF176214	Staphylococcus aureus			
□	A78544	Streptococcus pneumoniae			
□	LF138243	Streptococcus pneumoniae			
□	MH698333	Enterococcus faecium	2009-01-01		

Showing 15 results

Page 1 of 21,375

Figure 3. *Screenshot of the Outbreaks page of the Pathogens Portal. (1)It includes sub-tabs to show a metadata table of information, ENA Advanced Search to create structured queries and Nextstrain reports on some of the priority pathogens. (2) It presents the metadata table of results, with the ability to download metadata and data files, and editing of the table view itself. (3) On the left, facet filters enable users to input parameters to filter the metadata table.*

Priority pathogens are defined by the World Health Organisation 's(WHO)[16,17,18] list of priority pathogens, covering bacteria, viruses, fungi and other eukaryotes. The bacterial list covers organisms where antimicrobial resistance (AMR) is becoming a pressing issue - with last line antibiotics becoming increasingly used or failing in their usage. Taking on feedback from the scientific review and other project partners across work packages, the list will be

[16] Viruses: https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts [accessed 31.10.2023]

[17] Bacteria: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed [accessed 31.10.2023]

[18] Fungi: https://www.who.int/publications/i/item/9789240060241 [accessed 31.10.2023]

updated to include a range of sources, namely the Centres for Disease Control (US, Europe, Africa, etc.). The list of priority pathogens can be found and retrieved from this file in GitHub: https://github.com/enasequence/ena-content-dataflow/blob/master/classifications/pathogen_taxa/Priority_pathogens_taxonomy.csv.

Currently the Outbreaks page[19] presents sequencing data holdings across the priority pathogens. Overall there are >1.3million (sequence) records for the priority pathogens, including >320,000 sequences, >127,000 analyses, >420,000 raw reads, >485,000 samples and >9,900 studies. Some organisms are represented more than others, as expected. The page highlights these organisms through the facets, where the total number of records per organism are shown, thereby providing a quick view of where efforts could be concentrated.

5. Results

Work on this deliverable included the development of pathogen data classification based on taxonomy and tagging within the Pathogens Portal.

For taxonomy purposes the UK Health and Safety Executive (UKHSE) list was used as a core reference on what constitutes a pathogen on the Pathogens Portal, allowing at the same time for additional requests from users for specific pathogens to be added. As a next step the list was collated and tagged within the ENA backend ElasticSearch system, allowing for a simple 'tag=pathogen' query rendering all pathogenic records as results.

To monitor upcoming threats, an 'Outbreaks' page was added to the Pathogens Portal[20], providing a filtered view of the entire portal on a select group of pathogens identified as potential causes of the next epidemic or pandemic. For the 'Outbreaks' page, the World Health Organisation's priority pathogens list, covering bacteria, viruses, fungi and other eukaryotes are used. In addition a list of bacteria, covering organisms where antimicrobial resistance (AMR) is becoming a pressing issue, is included.

Pathogen Data Sharing is supported with a dedicated Pathogens Submission Guide and a help desk queue for submission-related queries, in addition to the Contextual Data Clearing House which allows the extension of pathogen metadata via a Data Hub or generally for public records. Areas for further development for Clearing House were explored through the submission of over 27 million SARS-CoV-2 curations by WP1 partner, The Arctic University of Norway (UiT).

[19] Pathogens Portal Outbreaks page: https://www.pathogensportal.org/priority-pathogens [accessed 31.10.2023]
[20] Pathogens Portal Outbreaks page: https://www.pathogensportal.org/priority-pathogens [accessed 31.10.2023]

Funded by
the European Union

Three pipelines were developed or explored as part of the analysis pipeline exploration task for antimicrobial resistance (AMR). WP1 partner DTU developed an AMR pipeline, constituting part of a preparedness Data Hub, and a further two community developed pipelines are being assessed for integration into the Data Hubs system.

Finally, Nextstrain, a real-time tracking of pathogen evolution visualisation tool was integrated in mid-2023 in the Pathogens Portal, presenting Mpox, Zika and West Nile virus reports. The current integrated versions are not fully automated to pick up incoming sequences in real-time from the relevant taxa but automation is under consideration. The reports can be updated if needed.

6. Next steps and impact

Looking at the next steps for Work Package 1, we intend to continue connecting new benchmark tools in the Data Hubs model (AMR, and metagenomics from other sister projects, e.g. VEO) . We are also starting work on components that will help deliver Deliverable 1.3.

The greatest impact of 1.2 is that users have now in place analysis tools readily available for pandemic preparedness. It also promotes open data sharing in pandemic preparedness through tailored support to users submitting pathogen data to a Data Hub. Finally the Pathogens Portal and its Outbreaks page provide an entry point for pandemic preparedness to find sequences of publicly shared priority pathogens, tailored Nexstrain reports, tools, and cohort-associate data, all in one place.

Funded by
the European Union