D1.3. Tracking and open analytics tools

Details of the tracking and open analytics tools that support supervision and monitoring of the network of locally installed SARS-CoV-2 Data Hubs.

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	31/03/2024	**Actual Delivery date**	22/07/2024
Delayed	[Yes]		
Partner(s) contributing to this deliverable	EMBL-EBI		
Authors	Nadim Rahman (EMBL-EBI)		
Contributors	Ahmad Zyoud (EMBL-EBI), Zahra Waheed (EMBL-EBI)		
Acknowledgements (not grant participants)			
Reviewers	Aitana Neves (SIB, ELIXIR-CH) Henning Hermjakob (EMBL-EBI) Ilaria Colussi (BBMRI-ERIC) - ELSI compliance		

Funded by the European Union

Table of contents

1. Executive Summary

Deliverable 1.3 details technical aspects related to the Data Hubs addressing key areas with regards to open data sharing. The deliverable scope addresses primarily tracking, analytical and other tools developed for Data Hubs that support localised instances of data hubs. In turn, this translates to work linked to the automation and maturation of the core Data Hubs at EMBL-EBI, also known as the Pathogen Data Hubs; a prerequisite for the second task within the WP - Local Data Hubs.

Local Data Hubs focus on generating a collection of standardised tools and workflows that follow standards at the European Nucleotide Archive (ENA), but can be used by collaborators using their own compute infrastructure and analysis subworkflows. The initial model detailed in this deliverable can be found on GitHub: https://github.com/enasequence/ena-local-datahub and has resulted in a standardised workflow template that has been created using DSL2 Nextflow[1]. The foundations set by the initial model allow for an advanced model to be developed in the future, that will include providing a workflow repository to deposit analysis pipelines, share and collaborate. In addition to the local data hubs, the deliverable report describes the Contextual Data Clearinghouse (CDCH) which is a data object store for community generated metadata curations associated with public International Nucleotide Sequence Database Coalition (INSDC) records. It supports the Data Hubs by improving the FAIRness and quality of submitted records, as curations are presented alongside the dataset in the ENA Browser. The ClearingHouse has undergone some significant improvements since the submission of SARS-CoV-2 curations by the Arctic University of Norway in 2021.

Furthermore, the report covers the extension of the core Data Hubs at EMBL-EBI which have been significantly extended during the lifetime of the BY-COVID project to support the localised instances of Data Hubs. Data Hubs users can have three main roles each with their own sets of responsibilities and actions: Coordinator, Data Provider, Data Consumer. To handle requests in setting up data hubs, a registration procedure was created. Following the registration process, a browser-based Setup form is sent to the coordinator to complete. The Data Hubs also include a Life Cycle Policy (LCP) that includes three separate statuses which define the life cycle of usage for an ENA Data Hub: Active, Dormant, Recycled.

Next steps in the final months of BY-COVID will be the release of frontend elements that are supporting the extension and greater automation of the Data Hubs:
- the release of the Setup form, which would be sent to approved requests for Data Hubs by coordinators, and

[1] https://www.nextflow.io/index.html

- the release of a data hub management interface, enabling coordinators to add/remove users to the Data Hub and edit aspects related to the Data Hub.

With the first version of the workflow template released, testing is the next main next step relating to the tool, which is envisaged to be carried out by WP1 partners, UiT and DTU. This will help identify any issues and fixes required, in addition to supporting planning for future work related to the Local Data Hubs.

The next steps related to the Clearinghouse, involve widely advertising its functionality through a supplementary document containing more general guidance for Clearinghouse users, and linking it out from an appropriate section of the existing ENA documentation. A new dataset of monkeypox curations is also planned to be submitted soon by WP1 partners, and there is ongoing work to index curations in the ENA Browser and Portal API.

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.	Yes
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

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.	No
	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.	No
	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).	Yes
	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. 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.	No

3. Methods

The European Nucleotide Archive (ENA), an open access repository for sequence data, hosts the Data Hubs. This was developed to support open data sharing, by emphasising collaboration across multiple institutes in a pre-release environment. The Data Hubs include data submission, analysis, visualisation, presentation, search and retrieval of sequence data and its analysed products. As the Data Hubs mature, a number of aspects are required and described as part of this deliverable. This includes:

1. More automated processes and architecture is required behind the scenes to expand and future-proof them. This supports users in registering interest, setup and management of their Data Hubs with little intervention from the ENA team.
2. A further angle to support open data sharing has been to improve support to data sharing and connectivity in local settings.
3. In both of the above scenarios, data sharing is beneficial with greater context and more complete metadata, providing better insight and interpretation of data.

To address 1, Data Hub extensions have been worked on - including frontend portal updates, mapping out user interactions and their downstream effects, backend database updates and intermediate API/pipeline improvements and developments. The Local Data Hubs analysis component was developed, with focus on workflow templating. Finally the Contextual Data Clearinghouse (CDCH) was expanded to support further curations and improve presentation of third-party curations.

All frontend improvements have been implemented on the Pathogens Portal[2]. This has been described in other deliverables and reporting, and is the interface to infectious disease related life sciences data at EMBL-EBI.

4. Description of work accomplished

4.1 Data Hub Extensions

4.1.1 Background

To support localised instances of Data Hubs, the core Data Hubs at EMBL-EBI have been significantly extended over the period of the BY-COVID project. Much of the work has included planning for future-proofing and user interactions with the Data Hubs, which has driven significant automation work. This overall provides a solid foundation for the frontend changes that we are implementing.

4.1.2 Data Hub Roles

Data Hubs define three main roles, each with their own sets of responsibilities and actions. This has been depicted by the schematic in fig.1.

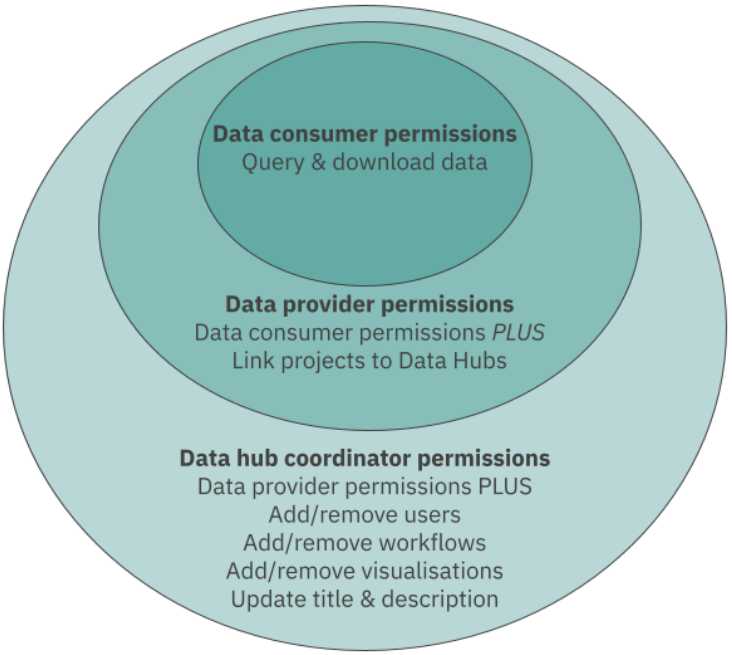

Figure 1. *Schematic presenting the permissions and user actions depending on the role a user has related to the Data Hub.*

[2] https://www.pathogensportal.org/

A **Coordinator** is a user who is responsible for the Data Hub as a whole. They are the main contact and will have communication channels with all users associated with the Data Hub, and the ENA helpdesk (through request forms, etc.). The coordinator must have a Webin[3] account, and use their Webin credentials to access Data Hub management features. A Webin account is a user account at the ENA, mainly used to submit data.

A **Data Provider(s)** is a user who is responsible for sharing data to a Data Hub. They must have a Webin account, and use their Webin credentials to share projects to the Data Hub(s) that they have access to. A Data Provider can have access to more than one Data Hub, and choose which projects they want to share with a given Data Hub using the 'Share' tab once logged into the Pathogens Portal: https://www.pathogensportal.org/datahubs#share. Therefore, not all projects within the Data Provider's Webin account are available to the Data Hub unless the Data Provider specifies, as shown by fig.2.

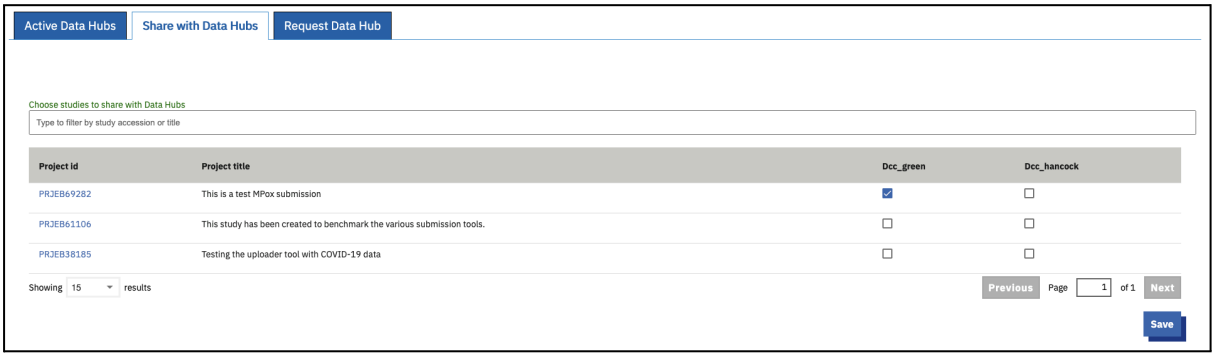

Figure 2. *Screenshot of a 'Share' page for a Data Provider that has been associated to more than one Data Hub.*

A **Data Consumer** is a user who is only consuming (searching and retrieving) metadata and data associated with the Data Hub. They use Data Hub credentials (dcc_XXXX) to access the Data Hub. By default, all users associated with a Data Hub are classified as Data Consumers, and so can use dcc credentials to access the Data Hub, however usage in this manner is limited to consuming data and not the added privileges mentioned for coordinators and data providers.

4.1.3 Registration

To handle requests in setting up data hubs, a registration procedure was created. This manifests through a 'Registration of Interest' form for a user (most likely a coordinator) who would like a Data Hub: https://www.pathogensportal.org/datahubs#request-data-hub. Here a user completes the following information, as shown in fig.3:

1. The coordinator Webin account - this is a requirement to help track Data Hub requests and provide additional privileges related to Data Hub management.
2. The coordinator email address

[3] https://www.ebi.ac.uk/ena/submit/webin/login

3. Estimated number of data providers
4. Estimated number of data consumers
5. Estimated period of Data Hub usage (months)
6. Consortium (if applicable) - if the Data Hub is associated with a consortium or project, then this can be mentioned here.
7. Reason for request

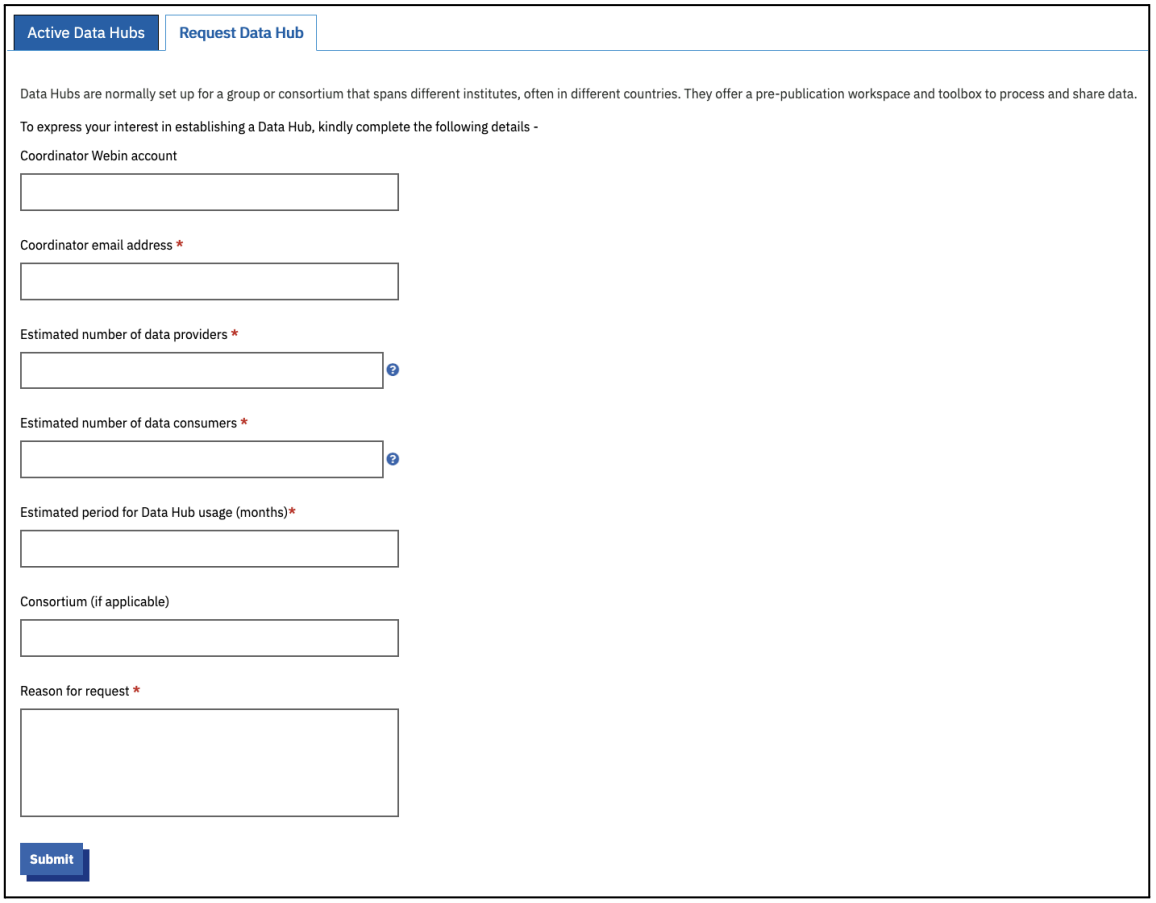

Figure 3. *Screenshot of the Data Hub Registration form for a user to register interest in a Data Hub*[4].

Once a user presses 'Submit', a helpdesk ticket is created with the Registration form information, for a member of the ENA team to review. The criteria used to assess include the following:

- Will data be private for a period of time?
- How many users are associated with the Data Hub?
- Are Data Hub users geographically distinct?
- Is the Data Hub required for excessively long periods of time?

[4] As regards data management and privacy issues related to which data are stored, how long and where, see the privacy policy at https://www.ebi.ac.uk/data-protection/privacy-notice/covid-19-data-portal

- Is the Data Hub associated with a consortium? If so, then Data Hubs can live for the life cycle of the project, and so do not need to adhere strictly to the Data Hub LCP mentioned below.
- Does the Data Hub need any specific analysis or visualisation associated with it?
- Does the reason for request suit the need for a Data Hub? Are there alternatives that may be useful, e.g. Umbrella Projects at the ENA.

Once approved, a user is then sent a 'Setup' form, described below.

4.1.4 Setup

Following the Registration process, a browser-based Setup form is sent to the coordinator to complete. This captures information required by the ENA to set up a Data Hub. This form is not accessible unless a helpdesk member provides it to the coordinator. The following information is captured by the form:

1. Coordinator full name
2. Coordinator affiliation and address
3. Coordinator Webin account
4. Coordinator email address
5. Data Hub description - longer abstract description
6. Data Hub topic - shorter description

The information retrieved is then fed into an automated pipeline that (1) assigns a Data Hub from an available pool of Data Hubs; (2) links data providers to the Data Hub so shared data can be associated with the Data Hub; (3) emails Data Hub credentials to all users. Alongside this pipeline, changes were made to an existing Data Hubs API to support the automated setup above, and also include management related features (such as adding or removing Data Hub providers/consumers).

4.1.5 Data Hub Tables

To support the work mentioned above and to future-proof the Data Hubs, several database tables were created or adapted. The overall aim was to remove dependencies on manual aspects, for example remove the requirement of handling and maintaining documents that detailed data providers and consumers.

The following database tables were created/updated:

1. List of active data hubs, along with descriptions, etc.
2. List of projects linked to data hubs
3. List of data providers linked to data hubs
4. Tracking of users for data hubs with their defined roles

5. List of data hub-compliant workflows
6. Workflows linked to data hubs

3 and 4 support logging actions of when a user has been added or removed by the coordinator and specifically remove the dependency of manual document maintenance. 5 and 6 enable for analysis workflow support and ability to add/remove any in the future.

4.1.6 Data Hub Life Cycle

The Data Hubs include a Life Cycle Policy (LCP) that includes three separate statuses which define the life cycle of usage for an ENA Data Hub.

1. **Active** - this is the default state for an ENA Data Hub that has been setup. An ENA Data Hub will remain in the Active state so long as there has been data submitted to an associated private project in the previous 6 months.
2. **Dormant** - if, after 6 months, an ENA Data Hub's private projects have had no data submitted to them and those projects are public or set to go public within 6 months, the status becomes dormant. A dormant ENA Data Hub can be moved back to Active if it receives submissions during this period. Submissions to associated public projects will not alter the ENA Data Hub's state.
3. **Recycled** - if no intervention is made by the coordinator after an ENA Data Hub has been placed in the Dormant state, the status is updated to Recycled after 90 days. In this state, all associations to Webin accounts and projects will be removed and the password will be changed to lock out old users. The ENA Data Hub can be assigned to a different set of users in the future, thus entering the life cycle again.

The ENA team is currently working on assigning statuses to the existing Data Hubs, and also cleaning up the Active Data Hubs page: https://www.pathogensportal.org/datahubs.

4.2 Local Data Hubs

4.2.1 Introduction

We are developing a collection of standardised tools and workflows that follow standards at the ENA but can be used by collaborators using their own compute infrastructure. We aimed to:

1. Maintain the ENA as the main repository to consume raw data and submit analysed data to.
2. Create a standardised workflow template that is suitable for the majority of analysis types.

3. Generate a user friendly workflow that can be implemented simply on a user's compute infrastructure.

The tools target bioinformaticians and individuals who have experience in setting up tools and workflows on compute infrastructure. This type of person is a key member in national and institutional sequencing and analysis efforts and helps facilitate surveillance and discovery of patterns and data interpretations. With the work described, the focus is on a single component of the core Data Hubs - analysis and its management. The collection of tools will expand in the future to bring about further extension of the Local Data Hubs.

4.2.2 Initial Model

The initial model, which has been achieved, is available on GitHub: https://github.com/enasequence/ena-local-datahub . This model has led to the creation of a standardised workflow template using DSL2 Nextflow[1]. This is a workflow management tool that follows the nf-core standard. Nf-core is a community effort to collect a curated set of analysis pipelines built using Nextflow and used by the wider scientific community.

In general, the initial model includes the following main features:

- ***Modularised*** - the workflow template is divided into a minimum of three modules:
    - A raw data fetching module - pulls data that a user would like to analyse from the ENA.
    - An analysis submission module - submits analysis data generated from the analysis of the raw data to the ENA.
    - A user module - includes at least one sub-workflow/process that constitutes an analysis workflow.
- ***Comprehensive documentation*** - associated with the workflow template, documentation on how to run and use it has been incorporated in the GitHub repository.
- ***Support through Docker, Singularity and Conda environments*** - these are tools that have a high level of integration with Nextflow, and are often already utilised by the community in sharing and running workflows whilst maintaining software dependencies.

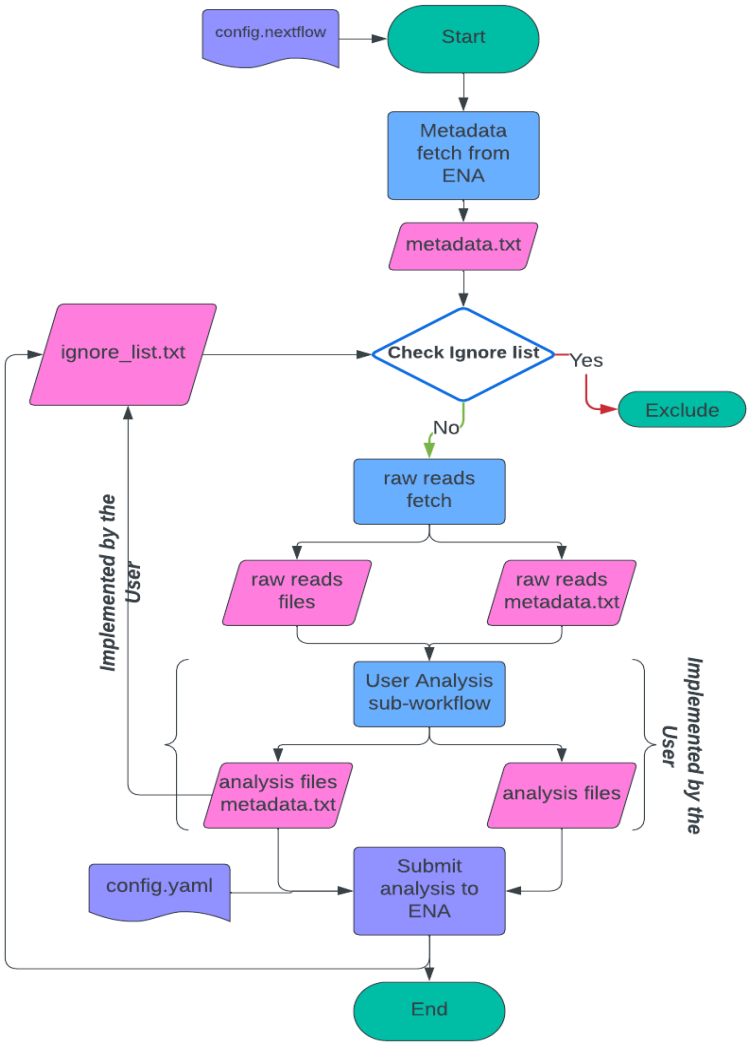

Figure 4. *Presents the architecture and steps of the standardised workflow template to support Local Data Hubs.*

Greater detail for the initial model is shown in fig.1. Here, a user defines a dataset that they wish to retrieve using a project accession and/or a taxonomic ID. In future, this could be adapted to include other fields to query on, for example country. The tool fetches the metadata based on this search criteria, saving it in metadata.txt. Following this, a check is done of the metadata file against a list of accessions to ignore (ignore_list.txt). This file can be populated by the user themselves, or by the pipeline overall to avoid duplicate processing of the same raw read. If there is an ignore list, rows that match this from the metadata.txt file will be excluded from the subsequent stages. The metadata.txt file is used to fetch the raw reads using links. This step also generates a metadata file associated with the raw reads (raw reads metadata.txt) available to support tracking. The user's analysis workflow is run, generating analysis files and a metadata.txt file relating to the analysis files (analysis files metadata.txt). Both of these components are fed to the final process of analysis submission to the ENA. The analysis metadata.txt file is also used to then add to the ignore list, to avoid duplicate analysis.

4.2.3 Advanced Model

The foundations set up by the initial model allow for an advanced model to be generated in the future should the opportunity arise. The advanced model includes providing a workflow repository to deposit analysis pipelines, share and collaborate. The model would use open source features provided from the nf-core repository and tools as a base for the platform, and integrate Pathogens Platform (and by extension ENA) specific features and tools. The following features would be available:

- Application of ENA standards alongside the nf-core community standards.
- Provision of tools specific for the needs of the Pathogens Platform (and ENA) user.
- Control status of pipelines, where they can be fully private, public or controlled access.
- Ability to share workflows easily between collaborators.

Nf-core provides a template for a site that would achieve the above, and this has been utilised by the Sanger Institute through sanger-tol: https://pipelines.tol.sanger.ac.uk/. This is essentially a repository of tools and pipelines utilised by the Sanger Institute for the investigation of genomic diversity of complex organisms, providing an end-to-end suite of pipelines for analysis. The aim of the advanced model would be to achieve something similar, but for the ENA.

4.2 Contextual Data Clearinghouse (CDCH)

The Contextual Data Clearinghouse (CDCH) (API here[5]) is a data object store for community generated metadata curations associated with public INSDC records. It supports the Data Hubs system by improving the FAIRness and quality of submitted records, as curations are presented alongside the dataset in the ENA Browser (e.g. https://www.ebi.ac.uk/ena/browser/view/OM635134.1?show=curations).

The ClearingHouse has undergone some significant improvements since the first set of submissions of SARS-CoV-2 curations by the Arctic University of Norway in 2021. Mainly, this includes implementation of data validation for the first time, to better record valuable curations (and reject those which record information already captured by ENA but do not comply with existing ENA field restrictions). CDCH search functionality has also been expanded to include suppressed curations - those that are erroneous in some way (e.g. wrongly submitted) and thus have been withdrawn by the curation submitter. Suppressed curations do not present in the ENA browser.

[5] https://www.ebi.ac.uk/ena/clearinghouse/api/swagger-ui/index.html#/Document%20Specification/downloadDoc

5. Discussion

The deliverable has detailed aspects related to the Data Hubs, which together aim to address key areas with regards to open sequence data and metadata sharing currently and in the future.

The Data Hubs are a powerful tool, and it has been developed with applicability beyond data domains in mind. During the BY-COVID project, the increased automation and maturity of the Data Hubs has enabled them to broaden their scope beyond just SARS-CoV-2 or infectious disease. This will improve sustainability of the Data Hubs (which are based on ENA's infrastructure) through a greater number of applications.

The Local Data Hubs provide a step towards decoupling the Data Hubs from EMBL-EBI infrastructure and support localised hubs of data in sharing sequence data. This task within BY-COVID has had its scope focused down due to the broad nature and extensive effort requirements. By focusing on more bite-sized chunks, the overall task would be achieved in steps. The existing tool does require extensive background knowledge, however can be applied relatively easily across infrastructure.

Metadata curations support greater reusability of data, whilst improving descriptions of contextual data and metadata for submitted data objects. This empowers researchers downstream in their interpretation and development of tools to support the wider research community.

6. Next steps

6.1 Data Hubs

The next steps here relate to release of frontend elements that are supporting the extension and greater automation of the Data Hubs. Namely, this includes release of the Setup form, which would be sent to approved requests for Data Hubs by coordinators. Secondly, the release of a data hub management interface, enabling coordinators to add/remove users to the Data Hub and edit aspects related to the Data Hub.

Overall, the features that are being included here have been coupled with future-proofing in mind, and so the ENA team is planning the release of a new Data Hubs Portal (DHP). This is currently only available in development, as text and imaging is included, however expected to be released soon. The DHP will be generic, available for any domain of data at the ENA.

Its use-case arose from non-pathogenic requests for a Data Hub (e.g. marine), and with the current setting, users would have to be directed through a specific Pathogens Portal to access Data Hubs which was not appropriate. This is alongside greater sustainability of the Data Hubs as a product in itself.

6.2 Local Data Hubs

The first version of the workflow template has been released, and so testing is the next main next step relating to the tool. There has been interest within the working group through UiT and DTU partners. This will help identify any issues and fixes required, in addition to supporting planning for future work related to the Local Data Hubs. Additionally, future next steps would include focusing on 'localising' a different component to the analysis, such as submission tools or visualisations.

6.3 Contextual Data ClearingHouse

The next steps related to Clearinghouse involve first advertising the Clearinghouse more widely. To help with this, a supplementary document containing more general guidance for Clearinghouse users has been created, and this will be linked out from an appropriate section of the existing ENA documentation. A new dataset of monkeypox curations is also planned to be submitted soon by WP partners, and there is ongoing work to index curations in the ENA Browser and Portal API.