

b

## Deliverable D9.2 Document where code from the various regional/national data hubs is available; training documentation on data submission by the platforms - relates to tasks 9.1-9.2

<b>Project Title (Grant agreement no.):</b>	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
<b>Project Acronym (EC Call):</b>	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
<b>WP No &amp; Title:</b>	WP9 Mobilisation of SARS-CoV-2 variant surveillance data tracking services and tools		
<b>WP leader(s):</b>	Aitana Neves (SIB) & Erik Hjerde (UiT) & Isabel Cuesta (ISCIII)		
<b>Deliverable Lead Beneficiary:</b>	3 - SIB		
<b>Contractual delivery date:</b>	31/05/2023	<b>Actual delivery date:</b>	07/06/2023
<b>Delayed:</b>	Yes		
<b>Partner(s) contributing to this deliverable:</b>	ALU-FR, CING, CNRS, DTL-Projets, EMBL-EBI, FPS, GÖG, IBCH, ISCIII, SIB, UiT, UNILU, CTU (UOCHB), UT, UU		
<b>Authors:</b> Aitana Neves (SIB), Lorenz Dolanski-Aghamanoukjan (GÖG), Wolfgang Maier (ALU-FR), Robert Pergl (CTU-UOCHB), Wolmar Nyberg Akerström (UU), Erik Hjerde (UiT), Isabel Cuesta (ISCIII), Marina Popleteeva (UNILU)			
<b>Contributors:</b> Aitana Neves (SIB), Anastasis Oulas (CING), Anliat Mohamed (CNRS), Arianna Tonazzolli (CNRS), Diana Pilvar (UT), Erik Hjerde (UiT), Erin Calhoun (UiT), Espen Åberg (UiT), Frédéric Erard (SIB), Helena Rasche (DTL-Projects), Imane Messak (CNRS), Isabel Cuesta (ISCIII), Jacques van Helden (CNRS), Lorenz Dolanski-Aghamanoukjan (GÖG), Maria Lara (FPS), Marina Popleteeva (UNILU), Nadim Rahman (EMBL-EBI), Nestoras Karathanasis (CING), Nils-Peder Willassen (UiT), Pawel Zmora (IBCH), Robert Pergl (UOCHB), Sara Monzón (ISCIII), Saskia Hiltmann (DTL-Projects), Sunny Singhroha (UiT), Terje Klemetsen (UiT), Thomas Denecker (CRNS), Wolfgang Maier (ALU-FR), Wolmar Nyberg Åkerström (UU), Zahra Waheed (EMBL-EBI)			
<b>Acknowledgments (not grant participants):</b> Kim Ng (STATENS SERUM INSTITUT, Denmark)			
<b>Reviewers:</b>	ELIXIR-CONVERGE Management Board (MB) members.		

## Log of changes

DATE	Mv m	Who	Description
15/05/2023	0v1	Aitana Neves (SIB)	Initial version
01/06/2023	0v2	Aitana Neves (SIB)	Sent to PMO after incorporating internal WP feedback
19/06/2023	0v3	Juan Arenas (ELIXIR Hub)	Circulated to the MB for final review before submission
26/06/2023	0v4	Aitana Neves (SIB)	MB comments addressed
05/07/2023	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

## Table of contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Contribution toward project objectives</b>	<b>2</b>
<b>3. Introduction</b>	<b>4</b>
<b>4. Description of work accomplished</b>	<b>5</b>
4.1 Set up an expert network of regional/national SARS-CoV-2 Data Hubs	5
4.2 Workshop and documentation on data brokering	7
4.3 Workshop and documentation on analysis pipelines	8
4.4 Workshop on legal aspects related to pathogen data sharing	9
<b>5. Results</b>	<b>10</b>
5.1 Mobilising SARS-CoV-2 data into the EU Covid-19 Data portal	10
5.2 A survey into the contribution of regional/national pathogen data hubs and on the resources needed to develop and maintain them	10
5.3 A Maturity Model for Federated One Health Surveillance Platforms	13
5.4 Advocating for Federated and International One Health Surveillance Platforms	13
5.5 ELIXIR CZ Proof-of-Concept FAIRification of Czech COVID-19 Virus Data	15
5.6 Data brokering documentation	16
5.7 Documentation of Analysis Pipelines and Visualisation Frameworks (WP9 Task Force - SARS-CoV-2 Analysis Pipelines)	16
<b>6. Conclusions</b>	<b>16</b>
<b>7. Impact</b>	<b>18</b>
<b>8. Next Steps</b>	<b>18</b>
<b>9. Deviation from Description of Action</b>	<b>19</b>



## 1. Executive Summary

Deliverable 9.2 relates to tasks 9.1 and 9.2 on strengthening the central Covid-19 Data Portal and coordinating nascent and established data hubs in mobilising data into the central portal. To achieve this, an active network of regional and national SARS-CoV-2 data hubs managers was established. To get existing data flowing, we first organised a technical workshop on data brokering to the ENA, from which an online tutorial and RDMkit documentation were published. Then, in order to support nascent and more established nodes in mobilising data, we started addressing the various challenges and gaps identified:

- (i) Importance to build a case for the regional/national data hub and brokering model, also for mid- and long-term funding: a survey was conducted on 11 countries to identify use cases, funding mechanisms and assess sustainability of the developed infrastructures and tools. We aim to publish it in F1000 (ELIXIR Gateway);
- (ii) Maturity assessment to oversee data hub development and identify potential areas of improvement: a maturity model for pathogen data hubs was developed and is readily available online.
- (iii) Sensitive data remains siloed and not shared on public open data repositories, calling for the need to federate pathogen data hubs through a central system enabling privacy-preserving data queries and controlled data access following FAIR principles: an opinion paper is being finalised to present and discuss a model of Federated One Health Surveillance Platforms, beyond Covid-19. A proof-of-concept of how a data hub might be “fairified” was also conducted in the Czech Republic.
- (iv) Training in analytical pipelines and documentation - several actively followed workshops were organised on Galaxy and available pipelines and tools have been documented.
- (v) Need to better understand the legal aspects of SARS-CoV-2 data sharing: a dedicated workshop was organised to identify the common needs and start addressing them. Legal experts should be further invited to the table given the many open questions remaining.

The pandemic has demonstrated the importance of genomic surveillance. This deliverable establishes the foundations to strengthen SARS-CoV-2 data hubs and importantly, scale and expand them to other pathogens and use cases serving both research and surveillance through a federation of hubs.

## 2. Contribution toward project objectives

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

Objective no. / Key Result no. Description	Contributed to:
<b>Objective 1:</b> Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities ( <b>WP1, WP5</b> )	

<b>Key Result 1.1:</b> Established European expert network of data stewards that connect national data centres and similar infrastructures and drive the development of interoperable solutions following international best practice, including national interpretations of the General Data Protection Regulation (GDPR)	<b>Yes</b>
<b>Key Result 1.2:</b> Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	<b>No</b>
<b>Key Result 1.3:</b> The catalogue of successful national business models incorporated into national strategies	<b>No</b>
<b>Key Result 1.4:</b> The developed “sustainable and scalable operating model for transnational life-science data management support” is adopted into national ELIXIR Node	<b>No</b>
<b>Objective 2:</b> Strengthen Europe’s data management capacity through a comprehensive training programme delivered throughout the European Research Area ( <b>WP2, WP6</b> )	
<b>Key Result 2.1:</b> A comprehensive ELIXIR Training and Capacity building programme in Data Management, directed at both data managers and ELIXIR users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	<b>No</b>
<b>Key Result 2.2:</b> Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	<b>No</b>
<b>Key Result 2.3:</b> A substantial cohort of data managers, Node coordinators and researchers with specific data management skills, business planning and knowledge of transnational operations across the ELIXIR Nodes	<b>No</b>
<b>Objective 3:</b> Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit ( <b>WP2, WP3, WP5</b> )	
<b>Key Result 3.1:</b> Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to integration with analysis platforms and making the data publicly available according to international standards.	<b>No</b>
<b>Key Result 3.2:</b> Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	<b>No</b>
<b>Key Result 3.3:</b> Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	<b>No</b>
<b>Key Result 3.4:</b> Enable ‘FAIR at source’ practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	<b>No</b>

<b>Objective 4:</b> Align national investments to drive local impact and global influence of ELIXIR (WP4,WP6)	
<b>Key Result 4.1:</b> Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	<b>No</b>
<b>Key Result 4.2:</b> Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	<b>No</b>
<b>Key Result 4.3:</b> Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	<b>No</b>
<b>Key Result 4.4:</b> Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	<b>No</b>
<b>Key Result 4.5:</b> Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	<b>No</b>
<b>Objectives - WP9</b> - Mobilisation of SARS-CoV-2 variant surveillance data tracking services and tools	
<b>09.1</b> Coordinate nascent and established national data hubs focusing on brokering services to define and foster common best practices	<b>Yes</b>
<b>09.2</b> Mobilise open SARS-CoV-2 genome data into the COVID-19 Data Platform ( <a href="https://www.covid19dataportal.org">https://www.covid19dataportal.org</a> ) from individual data hubs.	<b>Yes</b>
<b>09.3</b> Catalyse agreement on SARS-CoV-2 data standards for variants and lineages.	<b>No</b>
<b>09.4</b> Drive development of SARS-CoV-2 variant analysis tools..	<b>Yes</b>

### 3. Introduction

This deliverable relates to tasks 9.1 and 9.2 on strengthening data brokering at the regional/national SARS-CoV-2 data hubs and establishing a network of experts to define common best practices. It is strongly connected to WP8 and related to WP1 in that it establishes a network of data experts in the specific field of SARS-CoV-2 genomic surveillance, and also to WP2 for training and capacity building.

In order to achieve these aims, we first established a network of technical managers of SARS-CoV-2 Data Hubs within ELIXIR. This was important to ensure that technical people were joining the meetings to be able to share expertise and experience. During the first meetings, common needs and training gaps were identified, notably on the data broker role, data standards to use and legal aspects



of data sharing. Dedicated workshops were organised to address each of these and code/expertise were also shared between countries to support each other in establishing or consolidating their local hub. This led to substantial genomic data mobilisation and open data sharing on the EU Covid-19 Data Portal. As we progressed, dedicated task forces covering various topics were established to encourage contributions from all nodes.

An important issue that emerged was that of sustainability beyond SARS-CoV-2, as several hubs were starting to consider expanding their data hub to other pathogens and/or antimicrobial resistance (AMR). The apparent lack of resources and mid-term visibility prompted us to set up a survey to build a case on regional/national data hubs and get an overview of the available financial resources for mid- and long-term sustainability. As we discussed sustainability, the concept of maturity level of a Data Hub was further investigated and discussed with WP7, as it provided a means for each hub to know where it stands, to identify gaps and improve, and to target funding requests based on maturity development needs. A maturity model would also serve as a tool for capacity building. In parallel, the notion that sensitive data hosted by the data hubs was still siloed in each country fostered the idea of federating the data hubs to enable centralised data queries within appropriate and well-regulated data access mechanisms, similar to the Federated European Genome-Phenome Archive. Centralising data collection at the regional/national level would also serve data quality and standardisation.

Altogether, our activities have contributed to establishing an active network of SARS-CoV-2 Data Hubs technical managers, with a common vision to expand and federate them into One Health Surveillance Platforms serving both research and surveillance.

## 4. Description of work accomplished

### 4.1 Set up an expert network of regional/national SARS-CoV-2 Data Hubs

#### 4.1.1 Regular presentations and exchanges between countries, task-forces meetings

We organised regular monthly meetings where countries were also invited to present their activities around data mobilisation and brokering. Altogether, eleven countries have presented their active contribution in setting up a regional/national data hub, curating/analysing SARS-CoV-2 genomic data and brokering it to the EU Covid-19 Data Portal.

- 30 November 2021: Norway (Nils Willassen), Switzerland (Aitana Neves)
- 12 January 2022: Italy (Matteo Chiara, Federico Zambelli), Luxembourg (Nene Djenaba Barry, Jacek Lebioda)
- 2 February 2022: Belgium (Miguel Roncoroni, Bert Droesbeke)
- 3 March 2022: France (Jacques van Helden, Samuel Keuchkerian)
- 31 March 2022: Spain (Isabel Cuesta, Sara Monzón)
- 2 June 2022: Denmark
- 11 Oct 2022: Estonia (Diana Pilvar/Heleri Inno)
- 18 April 2023: Austria (Gunter Maier)

- 9 May 2023: Czech Republic (Robert Pergl)

The slides of the presentations are available in our rolling minutes:

<https://docs.google.com/document/d/15FRcGdQPif5uIGfBLLIramYlo2rm7i5o2MStF97DHlg/edit?usp=sharing>

These presentations generally triggered many questions and interactions among WP members interested in technical details and in understanding how each country addressed common challenges. It was also an opportunity for countries to continue exchanging informal knowledge and expertise bilaterally after the presentations.

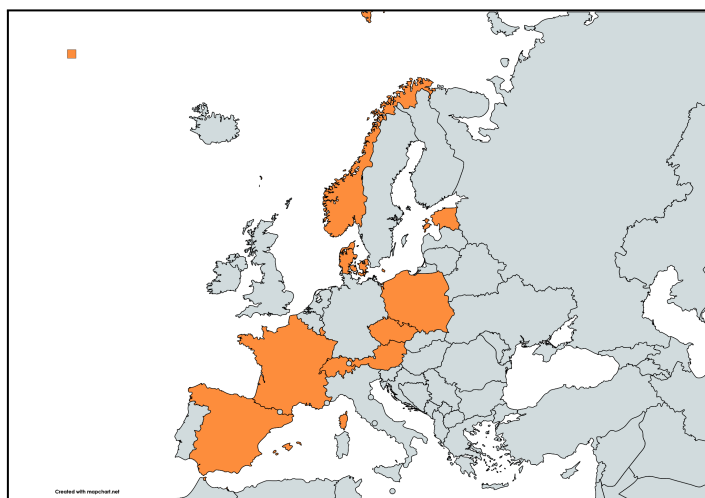
In addition to the monthly meetings, we set up several task forces that actively met to advance dedicated topics (number of members indicated in brackets):

- SARS-CoV-2 analysis pipelines (10)
- Legal aspects of data sharing (3 + SIB Head of Legal & Technology Transfer)
- Data brokering & Maturity model (6)
- Data Standards (8)
- RDMkit: Best practices for viral DM & submission (3)
- Pathogen Data Hubs (9)

#### 4.1.2 Impact and resources survey

Over the Summer 2022, we developed a survey to investigate the resources needed to develop and maintain a data hub and the long term vision for expanding and scaling them. Responses from 11 institutions were collected from August to September 2022 (Figure 1). Our survey covered three topics:

1. Understanding the impact of regional/national data brokering hubs
2. Understanding the resources used during the SARS-CoV-2 pandemic (expressed as person-months (PM))
3. Beyond SARS-CoV-2: resources for maintenance and expansion to other pathogens and antimicrobial resistance



**Figure 1.** Countries participating in the Impact and Resources survey of WP9.



#### 4.1.3 Assessing the maturity level of the Data Hubs

In order to work on the Maturity Model for pathogen data hubs, we decided to build upon the excellent work already performed by WP7 for the FEGA Maturity Model (MM). Over a workshop co-organised with WP1 in Padua in September 2022 that was followed by several meetings of a dedicated WP9 task force, we evaluated and, where needed, proposed re-wordings on all the criteria established by the FEGA MM. We also met with WP7 on several occasions prior to starting and after reviewing all the criteria to address our questions and present our suggestions for changes. WP7 showed interest in taking up some of our re-wordings as well, so this feedback was extremely useful both ways.

#### 4.1.4 Towards federating the Data Hubs

A task force on Pathogen Data Hubs was established to assess the current gaps and needs in the existing research and surveillance ecosystems and thereby propose a new model to federate pathogen data.

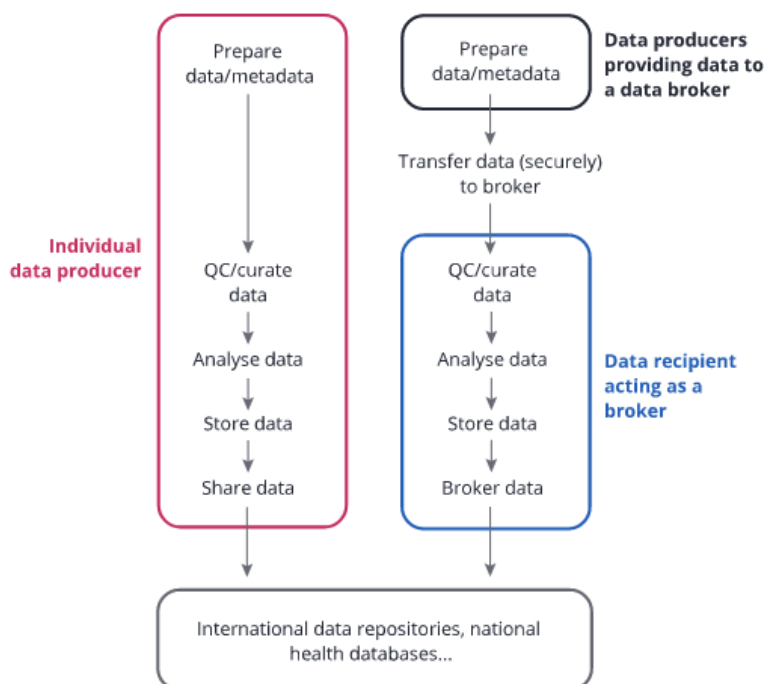
In a nutshell, during the pandemic, several SARS-CoV-2 data hubs were established. These hubs and the vast amount of data sharing during the pandemic demonstrate that many challenges on data quality and data sharing have been successfully tackled, at least to some extent, by several countries. The solutions and developed products now need to be maintained, anchored and further expanded to other data types like wastewater datasets or bacterial datasets linked to antimicrobial resistance and food-borne pathogens within a One Health context. The data hubs also need to be more interfaced to avoid sensitive data silos and ensure that high quality data is available for both research and surveillance. In order to address all these challenges and fill the current gaps within the surveillance and research ecosystems, the task force proposed to establish a Federated network of regional and national One Health Surveillance Platforms (FOSP) with FAIR principles at its core. A central International One Health Surveillance Platform (IOSP) would also be established to coordinate the activities.

After thoroughly discussing the challenges, needs and potential solutions within FOSP/IOSP, the task force agreed to prepare an opinion paper that is currently being finalised (submission planned end of June).

### 4.2 Workshop and documentation on data brokering

The first activity conducted by WP9 was to organise a workshop together with WP8 on technical data brokering to the ENA. The workshop took place on 12 October 2021 and involved +30 participants from 15 countries. In order to better understand the experience of the participants, a short pre-survey was conducted, showing that  $\frac{2}{3}$  of the 10 respondents had started submitting data to the ENA (about  $\frac{1}{3}$  assemblies, and  $\frac{2}{3}$  raw data).





**Figure 2.** Data Brokering Workflow. Individual data producers can process the data, store it, and submit it directly to international repositories or public health databases. Alternatively, in the data brokering model, several data producers can submit their data to a common data recipient. This recipient might be in charge of curating the data, analysing it with common pipelines, storing it, and re-sharing parts of the data to public health databases and international repositories (as agreed with the data providers). The latter service is often referred to as “data brokering” i.e. sharing data on behalf of others within a well defined ethical and legal framework. Note that legal aspects should be considered along all the steps. Image and caption reproduced from RDMkit - [https://rdmkit.elixir-europe.org/data\\_brokering](https://rdmkit.elixir-europe.org/data_brokering).

An online ENA tutorial was developed for the occasion and the session was followed by an intense and very technical Q&A session, also including questions collected at the pre-survey. The answers to the Q&A were included in the [slides of the workshop](#)<sup>1</sup> and distributed to the participants.

After the workshop, a WP9 task force was created to generalise this material into a generic “Data brokering” RDMkit page, explaining the concepts (Figure 2) and main challenges to take into account. One year later, in September 2022, a dedicated Data brokering workshop was also co-organised with WP1 in Padua to continue covering and developing these aspects.

### 4.3 Workshop and documentation on analysis pipelines

Task force meetings on SARS-CoV-2 Analysis Pipelines quickly made clear that knowledge about available analysis pipelines / workflows for SARS-Cov-2 sequencing data analysis was still not widely nor sufficiently spread. As the main contribution of the task force to task 9.4, we decided to compile a list of such pipelines/workflows used by different ELIXIR nodes and partners. This list comprises pipelines that enable analysis flows from raw sequencing data of viral samples to identify molecular variants, but also lineage assignments for those samples. In an online workshop that took place at the

<sup>1</sup><https://docs.google.com/presentation/d/1VG1O45ghTKKpKgym-rJze6eEBnykSuft/edit?usp=sharing&ouid=103740517945101968280&rtpof=true&sd=true>

beginning of May 2023, task force members also compiled an additional list of platforms available for high-level aggregation and visualisation of sequencing-based viral surveillance data produced by the pipelines/workflow from the first list.

Task force members also participated in the BY-COVID [Infectious Diseases Toolkit](#)<sup>2</sup> workshop held in Gent, Belgium in March 2023 to describe major workflows for SARS-CoV-2 sequencing data analysis for the Galaxy platform and their integration into a modular SARS-CoV-2 surveillance system. This work has now resulted in a dedicated [Showcase](#)<sup>3</sup> within the IDTk.

Training on using the Galaxy workflows, on reusing the IDTk Showcase system and on data submission to the ENA from within Galaxy has been delivered through two workshops on “SARS-CoV-2 Data Analysis and Monitoring with Galaxy” (August 9-12 2021, +750 registrations, 82 countries and December 1st, 2021, 180+ participants, 50 countries), and more recently as part of a new [One Health module](#)<sup>4</sup> within the last [global Galaxy training event](#)<sup>5</sup>.

#### 4.4 Workshop on legal aspects related to pathogen data sharing

A two hours online workshop organised by ELIXIR CONVERGE WP9 in collaboration with SciLifeLab took place on 30 March 2022, which was attended by 76 participants from 19 countries. Speakers reported on solutions adopted and developed by project members in the light of GDPR and applicable (local and European) legislation. It showcased legal contexts for data processing and contracts between organisations from several jurisdictions and indicated legal aspects to be considered when genetic, personal and phenotypic patient information is processed, masked and shared across data producers, data brokers and data sharing platforms.

Some key topics and challenges discussed in the workshop:

- lack of knowledge of and / or clarity in legal frameworks; national vs. international law;
- important role of the data broker;
- challenge to handle sensitive data appropriately;
- definitions and applicable rules for non-personal data, pseudonymised personal data, anonymised personal data, metadata;
- rules for ID management, definition of “identifiable”, requirements for “anonymity”;
- different data processing categories, privacy status of processed data;
- regulatory oversight, ethical approval;
- terms and procedures to cover in contracts;
- liability depending on the role ( controller, processor, ...);
- reuse of shared data, licensing;
- data ownership.

It became evident that there is a major need for guidance and clarity on how to operate in a safe and compliant manner when mobilising and sharing such types of data. This is a difficult task requiring a lot of legal expertise, which can impede projects when legal risks cannot be mitigated with certainty, even more so in international projects.

---

<sup>2</sup><https://www.infectious-diseases-toolkit.org/>

<sup>3</sup><https://www.infectious-diseases-toolkit.org/showcase/covid19-galaxy>

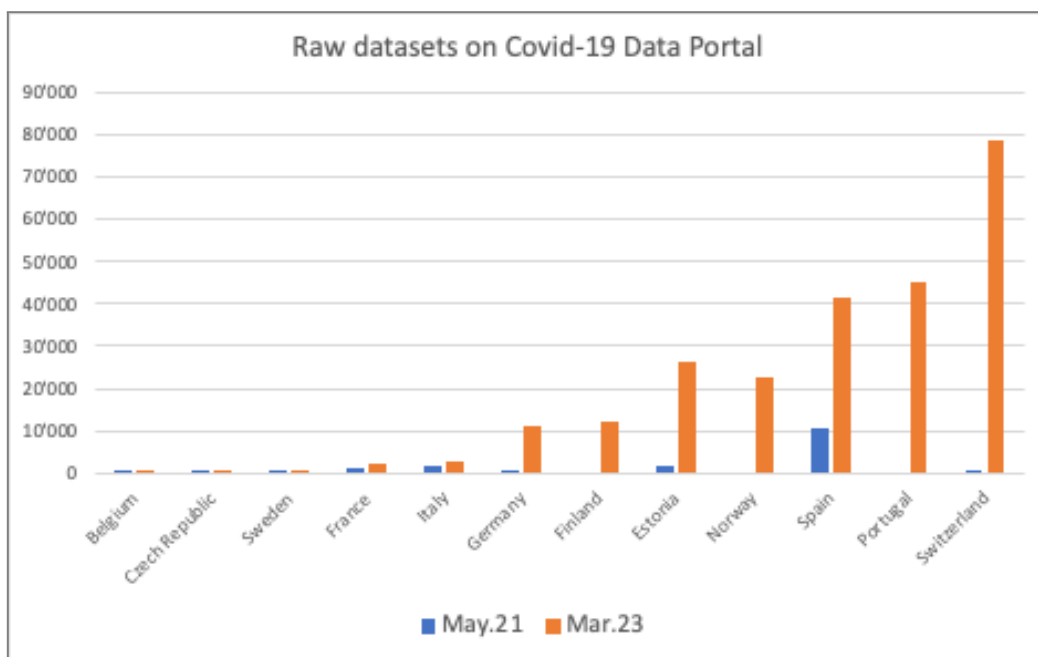
<sup>4</sup><https://gallantries.github.io/video-library/modules/one-health>

<sup>5</sup><https://gallantries.github.io/video-library/events/smorgasbord3/>

## 5. Results

### 5.1 Mobilising SARS-CoV-2 data into the EU Covid-19 Data portal

Over the course of the project, several countries started submitting consensus genomes and raw data to the ENA and hence to the EU Covid-19 Data Portal, with some notable contributions thanks to the work performed within WP9 (technical tutorial, sharing of expertise and code, peer-motivation). The established data brokering pipelines are generally ready to be extended to other pathogens with little effort (change of ENA checklist mostly).



**Figure 3.** Raw data submitted to the ENA prior to WP9 and as of March 2023. We acknowledge that different contributions might explain the increase in submissions, and not only the work performed within WP9.

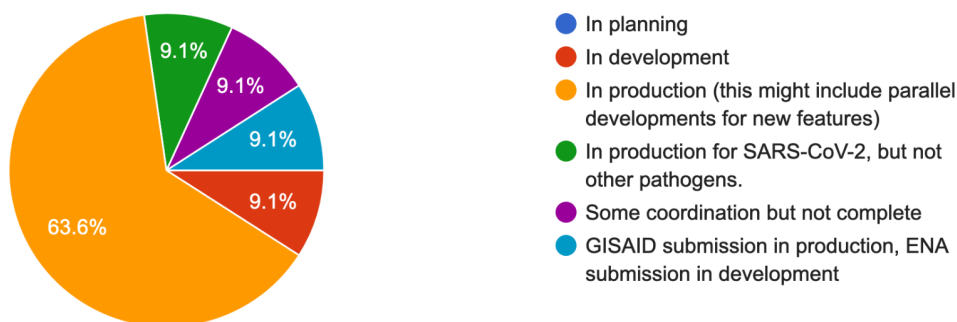
### 5.2 A survey into the contribution of regional/national pathogen data hubs and on the resources needed to develop and maintain them

The results of the survey have been aggregated and compiled into a report that is currently under review among WP9 members. We aim to publish it in F1000 (ELIXIR Gateway).

[https://docs.google.com/document/d/1kv83-OCxUYFprMLW49G7YIOnZn3VUC\\_i4wBo7Xq4-1w/e/dit?usp=sharing](https://docs.google.com/document/d/1kv83-OCxUYFprMLW49G7YIOnZn3VUC_i4wBo7Xq4-1w/e/dit?usp=sharing)

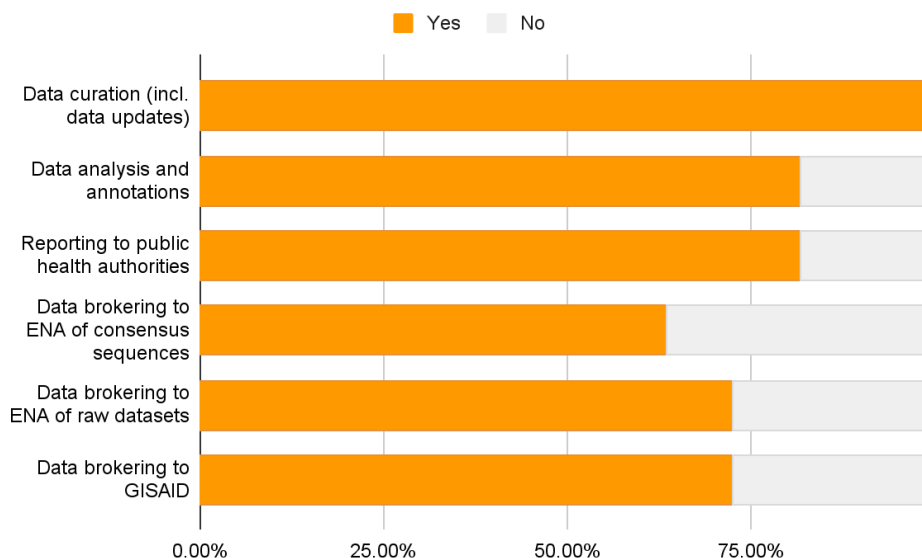
Our survey shows that at least 8 regional/national data hubs (DHs) were in production for SARS-CoV-2 data brokering at the time of this survey (Figure 4).

While here maturity was evaluated with a single criteria, we suggest to now use the developed pathogen DH Maturity Model to better account for differences in DHs maturity and foster capacity building and quality among DHs.



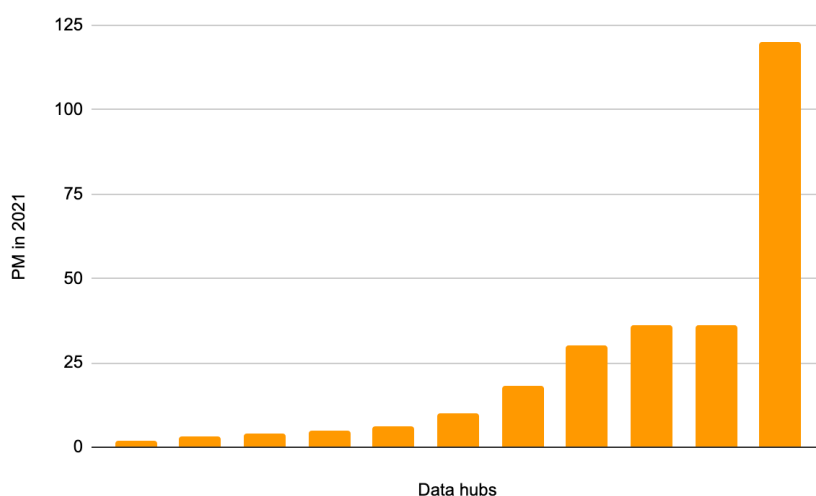
**Figure 4.** Self-evaluated maturity stage of the regional/national DHs. 11 responses.

The performed tasks varied but always included data curation (Figure 5).



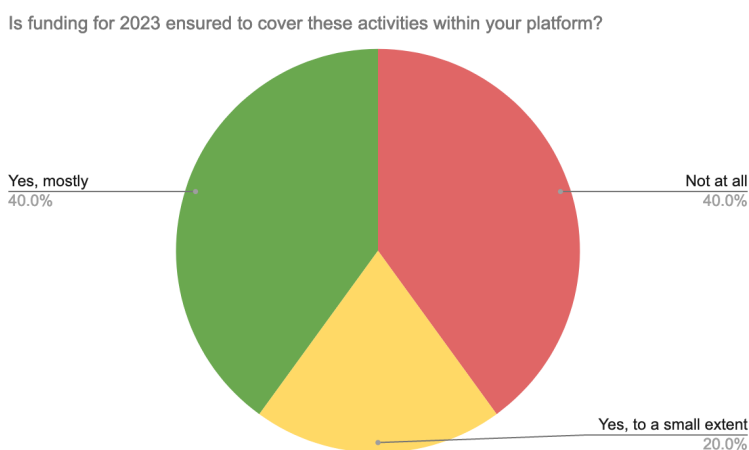
**Figure 5.** Tasks performed by the regional/national DHs. Percentages out of 11 responses.

The median workforce of 10 PM over 2021 (Figure 6) and envisioned 18 PM over 2023 reflect the willingness to expand to other pathogens and applications beyond SARS-CoV-2 while doing so in good conditions with sufficient resources, whereas during the pandemic some DHs might have over-worked with limited resources.



**Figure 6.** Total PM used over 2021 by each DH.

There is however a clear need for more visibility towards short- and mid-term funding, as at the time of the survey, 60% of DHs were still not covered for most of their activities over 2023 (Figure 7). The large panel of funding bodies exemplifies that 50% of the surveillance infrastructures are still covered by research money (institutional money, public and private foundations, Ministry of Research) (Figure 8), a situation that is not sustainable in the long term since infrastructures hardly receive research funding past their pilot phase. While the research community is realising the importance of funding infrastructures with dedicated calls, our survey demonstrated that even established national infrastructures with proven track-record of performance and end-user satisfaction during an urgent context still struggle to find sufficient short-term funding.



**Figure 7.** Planned availability of funding for 2023. 10 respondents.

Overall, our survey has highlighted the need to federate existing DHs to make a stronger case and collaboratively apply to common funding schemes at the European level to establish a major infrastructure for genomic surveillance of pathogens, e.g. into Federated One Health Surveillance Platforms. This would hopefully also encourage national public health authorities to realise that their

country is part of a greater network and that without adequate funding, the belonging to the network might be lost with all its consequences upon major outbreak/pandemic events.

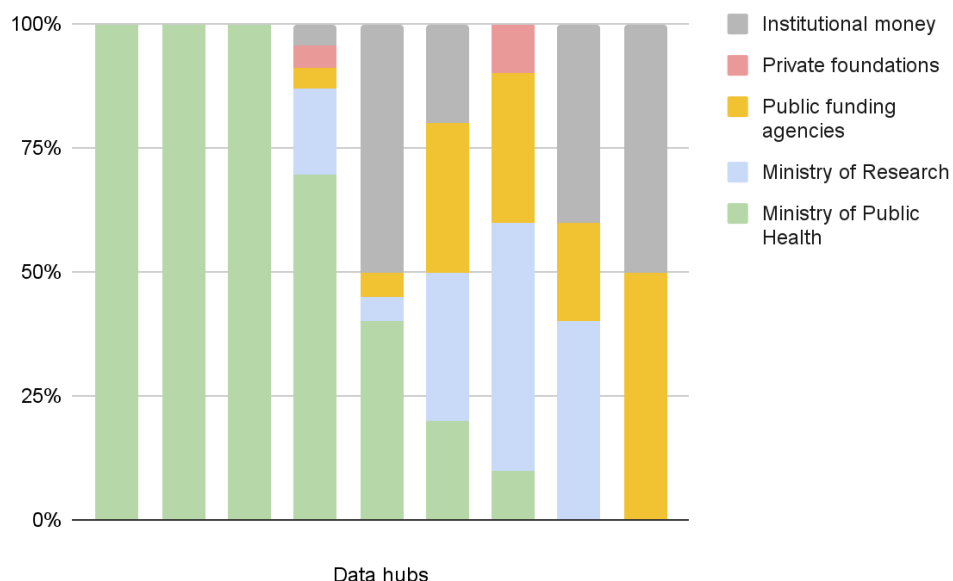


Figure 8. Funding sources over 2021. 9 respondents.

### 5.3 A Maturity Model for Federated One Health Surveillance Platforms

The Maturity Model has been published online. It consists of a Google spreadsheet that is automatically parsed into a Github Page:

<https://elixir-europe.github.io/fosp-maturity-model/>

This Maturity Model can already be consulted by the countries interested in evaluating their maturity and identifying potential areas of improvement. In the near future, we aim to define essential criteria to further support data hubs in prioritising their developments.

### 5.4 Advocating for Federated and International One Health Surveillance Platforms

The dedicated task force is finalising an opinion paper that summarises the challenges that still exist with the current situation of siloed pathogen data hubs, and where the concepts of Federated and International One Health Surveillance Platforms are presented (Figure 9).

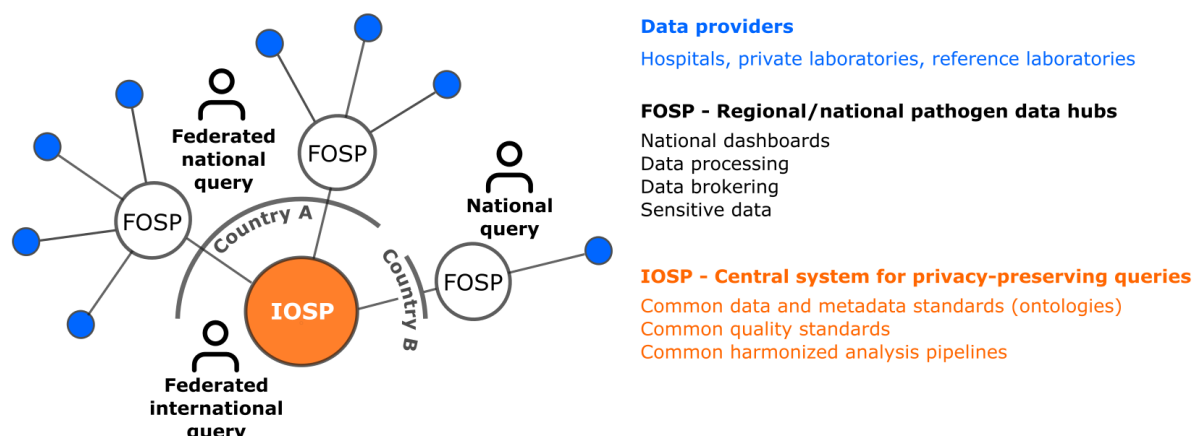


Figure 9. Federated One Health Surveillance Platforms for high quality surveillance data and FAIR data access for research.

In a nutshell, the FOSP/IOSP would serve the following aims:

- **Foster high quality data.** Ensure timely collection of regional/national pathogen molecular data with agreed quality metrics and minimal metadata. Foster the establishment of data curation services within the FOSP node.
- **Fill the gap on Findability of sensitive data for research.** Through the IOSP Portal, enable privacy-preserving queries to be made to find data for research studies. Support FOSP nodes in establishing interfaces with IOSP according to agreed standards (Figure 9).
- **Fill the gap on Accessibility of sensitive data for research.** Set up Data Access Committees within IOSP, within a clear ethical framework. Support FOSP nodes in labelling data with predefined access levels for semi-automated data access.
- **Fill the gap on Interoperability of data for research.** IOSP to contribute to international standards definition, where needed. Support FOSP nodes in adhering to common data standards (as agreed within IOSP).
- **Foster comparable processed data.** Foster open sharing of workflows and benchmarking with common open datasets. Organise pathogen/topic-specific workshops to harmonise analysis pipelines. Define quality labels for processed data generated within workflows successfully evaluated at External Quality Assessments programs.
- **Fill the gap on Reusable data for research.** Define minimal standards for data, metadata description, including provenance reports for processed data.
- **Foster capacity building and improvement cycles.** Establish a Maturity Model with minimal requirements to become a FOSP node. Provide capacity building support through the federated network. Help nascent and mature nodes identify gaps to be addressed on a regular basis through periodic self evaluations of their maturity level.
- **Integrate within the existing surveillance ecosystem.** FOSP nodes could act as main contact points and trusted sources of high quality curated data for both the ECDC and EFSA NGS Systems (Figure 10).
- **Integrate within the existing Open Research Data ecosystem.** Establish common standards (as agreed within IOSP, e.g. on minimal metadata and anonymisation needs) for data brokering services to international open data repositories.
- **Collect, process and share data within a well established legal framework and governance.** Establish a clear and trustworthy governance. Define and oversee all legal documents between IOSP and the FOSP nodes.



- **Equity.** Ensure credit is given to data providers and data processors through metadata requirements and appropriate citation procedures.
- **Access timely data.** Promote that minimal, anonymized data are rapidly openly shared, with embargo periods to be agreed upon.
- **Support pandemic preparedness.** Maintain and scale pathogen data infrastructures of quality.

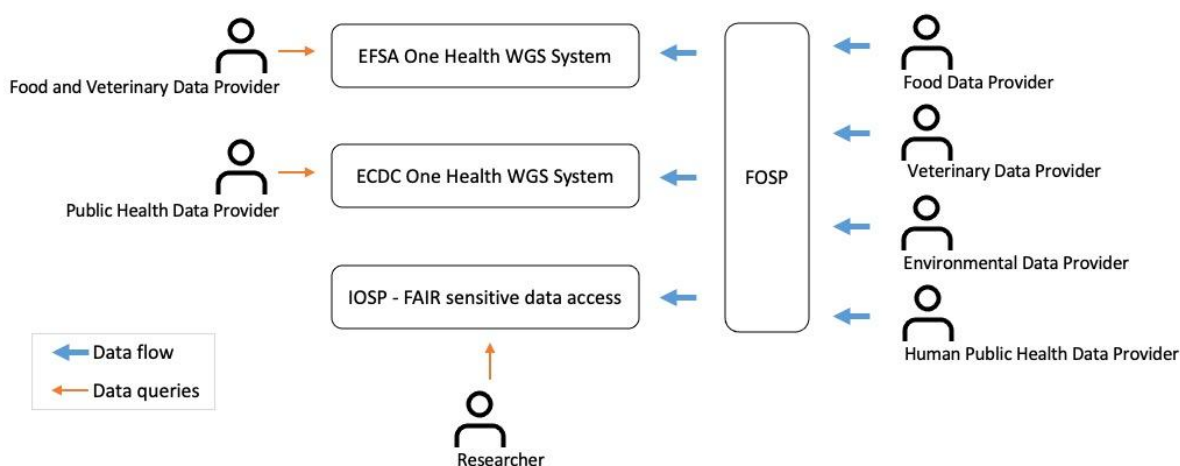


Figure 10. Integration of the FOSP/IOSP infrastructures within the existing surveillance and research ecosystems.

## 5.5 ELIXIR CZ Proof-of-Concept FAIRification of Czech COVID-19 Virus Data

COG-CZ, a group formed by scholars and experts in the Czech Republic, was established with the purpose of scrutinising and monitoring virus variants and mutations during the COVID-19 pandemic. Their findings were regularly made available on the website <https://virus.img.cas.cz>, which offers a comprehensive collection of samples, variants, and lineages. Additionally, they released weekly reports in PDF format.

Although the web portal provides access to all the data and some nice user features, such as graphs and a map, it is not FAIR-compliant. In our proof-of-concept study, we aimed to complement the portal with a FAIR-compliant solution. For that, we utilised "FAIR Data Point" (FDP), a vision of a group of authors of the original paper on FAIR on how (meta)data could be presented on the web using existing standards (<https://www.fairdatapoint.org>). A so-called "Reference Implementation" is available open-source (<https://github.com/FAIRDataTeam/FAIRDataPoint>) following the FDP specifications (<https://specs.fairdatapoint.org>). It is mainly based on the DCAT W3C standard for digital catalogues.

We installed and configured a FDP to accommodate necessary COG-CZ metadata and filled the catalogues from the COG-CZ datasets using scripts. The result is available at <https://ccmi.fit.cvut.cz/fdp-wp9>. The solution consists of a REST web server with API providing metadata in RDF formats and a client web application. While the client web application is useful for browsing and searching the content, the merit lies in the FAIR-compliant machine-actionable REST API also offering SPARQL queries.

While the solution is not production-ready due to technical limitations of the FDP Reference Implementation and the contents has not been fully curated, it provides an example of how virus data can be FAIRified. Fighting the pandemics essentially requires Findability, Accessibility, Interoperability, and Reusability of virus data. The provided PoC shows how this could be technically achieved and is meant mostly as a study material and inspiration for implementing FAIR principles by the repositories providers.

## 5.6 Data brokering documentation

The ENA online tutorial for SARS-CoV-2 data brokering is available here:

[https://ena-covid19-docs.readthedocs.io/en/latest/submission\\_workshop/getting\\_started.html](https://ena-covid19-docs.readthedocs.io/en/latest/submission_workshop/getting_started.html)

The RDMkit material for presenting data brokering also beyond SARS-CoV-2 data is available here:

[https://rdmkit.elixir-europe.org/data\\_brokering](https://rdmkit.elixir-europe.org/data_brokering)

## 5.7 Documentation of Analysis Pipelines and Visualisation Frameworks (WP9 Task Force - SARS-CoV-2 Analysis Pipelines)

A table of SARS-CoV-2 analysis pipelines/workflows and tools used by CONVERGE partners and a table of visualisation frameworks that can be used to present aggregated results of SARS-CoV-2 surveillance efforts are provided in: [📄 Table of SARS-CoV-2 analysis pipelines](#)

# 6. Conclusions

We summarise below the key messages from our work.

### **A network of pathogen data hubs has been successfully established**

This network of pathogen data hubs has been actively contributing to SARS-CoV-2 open data during the pandemic. Starting from semi-automated prototypes, many of these hubs now run in production with the highest quality and security standards. Importantly, those managing these resources have come together within a trustworthy network where expertise, knowledge and code are being shared, and where a common vision for strengthening and federating this network has emerged.

### **This network will serve as a pillar towards establishing the FOSP/IOSP ecosystem**

While many platforms exist that centralise pathogen data for research or surveillance purposes, our network has identified gaps that still need to be addressed to seamlessly integrate all the actors at play. The FOSP/IOSP ecosystem aims to be complementary and integrated to the surveillance NGS systems from ECDC and EFSA, while also supporting research based on FAIR data and increasing data quality overall.

### **Need to invite legal experts to the table**

Our work has tackled challenges faced by pathogen data hubs, bringing together a network of experts and enabling sharing of expertise on topics such as data collection and data brokering. Yet, during our monthly meetings, we realised that regional/national data hubs were little prepared to

handle the legal aspects related to mobilising and brokering pathogen genomic data and its associated (potentially sensitive) metadata.

As illustrated in the legal survey and workshop presentations from various countries, several data hubs successfully (yet often painfully) managed to identify the legal conditions under which pathogenic data and associated metadata could be collected and shared. Most of them managed to establish a legal framework for data collection (17 out of 20 survey respondents) and data sharing (13 out of 15). There are however currently no common legal procedures and documents across countries, these contracts being often bound to confidentiality. Moreover, in the pandemic emergency situation, each data hub had to rapidly find solutions to adapt to EU (e.g. GDPR) and national regulations (e.g. national laws on epidemics/research...), with sometimes different legal interpretations on what is personal and/or special categories (sensitive) of data between countries (e.g. isolation date was considered as component of personal data in at least one country). Finding a common ground for legal issues related to data sharing across Europe (e.g. identification of legal basis, development of data sharing agreement templates) will therefore require substantial additional work, unfortunately out of the scope of this CONVERGE WP9. As a consequence of this complexity, mainly non-sensitive data have been shared to international repositories, while sensitive data might have been shared only with national public health authorities, often remaining siloed and inaccessible to researchers. The legal framework for making sensitive data FAIR still needs to be addressed by most countries.

In summary, our activities have highlighted that those involved in data brokering did not have sufficient legal expertise, while legal experts involved in the pandemic may not always have had sufficient knowledge on the data and metadata types to give specific advice on how to share the data with contextual information related to its human host and keeping epidemiological value. Altogether, we feel that legal experts were missing around the table and should be an integral part of future discussions on the FOSP/IOSP ecosystem. For a sustainable impact these must closely align with the regulation on a European Health Data Space currently being negotiated, which foresees to facilitate the provision of and access to “relevant pathogen genomic data, impacting on human health” (art. 33 draft proposal).

### **Funding of FOSP nodes however remains uncertain, weakening the FOSP/IOSP ecosystem**

Another important challenge faced by many members of our network has been mid- and long-term funding, ensuring sustainability of the developed resources beyond Covid-19. Our survey has shown that data hub managers generally have only short-term funding perspectives and that as key infrastructures, a stronger case should be made to encourage public health authorities to financially support these resources. While international funding might be sought to develop the IOSP and tools for the FOSPs, nodes also have substantial maintenance, development and curation needs that should be covered by national entities. Without strong nodes anchored in a region or country, the whole FOSP/IOSP ecosystem will remain weak and uncertain.

### **Need to identify funding sources for establishing the FOSP/IOSP ecosystem and to support individual FOSP nodes**

In order to develop the IOSP, funding opportunities are envisioned within Horizon Europe or NIH funding schemes. While funding might also be available for establishing new FOSP nodes based on common tools and resources, the question of the running costs in production remains to be solved. Business models should be further investigated, where each country will likely find its own solution, whether funded through research grants, by public health authorities or by the data providers themselves.

## 7. Impact

As shown in Figure 3, the likely impact of our work on data mobilisation and open data sharing has been significant, in particular as regards raw genomic data that were otherwise generally not being shared on other repositories.

The maturity level model and the proof-of-concept of the FAIRification of a national pathogen data hub also pose the foundations towards establishing and expanding/improving production-grade One Health Surveillance Platforms that comply with FAIR principles.

Our achievements and willingness to continue the networking activities demonstrate that the regional/national pathogen data brokering model works and is key (i) to ensure common best standards on e.g. data anonymization, (ii) to ensure higher data quality thanks to curation and standards and (iii) to ensure that data is not siloed by using e.g. pseudonymised identifiers that can be used to link data to other datasets where authorised. Importantly, where gaps have been identified in the international research and surveillance ecosystems, our preliminary work on the FOSP/IOSP infrastructure may serve global and national efforts in developing pandemic preparedness programs.

## 8. Next Steps

ELIXIR CONVERGE WP9 has enabled creating a supra-national network of pathogen genomic data hubs that started sharing and working on common best practices. In the future, we envision that pathogen genomic data hubs expand their activities beyond SARS-CoV-2 to embrace One Health surveillance of pathogens across human, veterinary, food and environment compartments, also focusing on e.g. antimicrobial resistance and food-borne pathogens. While we have established the contours of this Federation of One Health Surveillance Platforms, future work should focus on the actual implementation of the proposed FOSP/IOSP concept. In particular, this will require funding to enable additional work from existing and novel task forces, listed below.

**Governance task force:** this task force will define the governance bodies of the FOSP/IOSP ecosystem with the aim of ensuring equity and trustworthiness.

**Legal and ethical task force:** This task force will describe the ethical and legal frameworks that the FOSP/IOSP ecosystem operates in and maintain a map of the landscape of regulatory authorities and regulations that must be considered when interfacing between healthcare and research infrastructures. A priority should be to survey the legal landscape for secondary use in the EHDS, European health and environmental surveillance systems, and the European Reference Laboratories responsible for certifying medical devices for clinical use (including bioinformatics solutions).

**Maturity model task force:** this task force will be in charge of finalising, maintaining and expanding the draft FOSP maturity model, also identifying essential indicators to be checked for a new data hub to enter the FOSP ecosystem in production. It may start by testing the model on a few existing pathogen data hubs to identify missing/superfluous indicators, or indicators that need rewording for clarity. As the FOSP/IOSP ecosystem evolves, the task force will ensure that the maturity model is up-to-date. It should also support FOSP nodes in filling the maturity model and suggest areas of improvement.

**Data Access task force:** this task force will define the operational functioning of Data Access Committees that will be in charge of evaluation data access requests. This would notably include definition of data access tags for datasets, as well as recommending SOPs for FOSP nodes to handle data requests semi-automatically and in line with international standards.

**Surveillance task force:** this task force will aim to establish tight contacts with EFSA and ECDC to enable FOSP nodes to be recognised as data brokers for their region/nation, complying with their respective data formats and standards.

**Infrastructure task force:** this task force will develop (if needed) and provide tools and infrastructure to support interfaces between FOSP nodes and IOSP (e.g. API for finding data and accessing it; secure data transfer mechanisms; secure environment for hosting sensitive data within a data hub; ...).

**Data standards and quality task force:** this task force ensures data quality within the FOSP ecosystem. It should notably describe the metadata and data to be collected by FOSP nodes, agree on compulsory minimal sets of data per use case (e.g. for different pathogens), propose controlled vocabularies and ontologies to be used, define quality tags for raw data and processed data, propose datasets to benchmark and validate algorithms and pipelines, and propose analytical tools to be integrated at the FOSP node or centrally at IOSP where relevant.

**FAIR task force:** this transversal task force ensures that interfaces and standards put in place comply with latest FAIR objectives and make continuous suggestions for improvements. It also supports FOSP nodes with open data sharing of anonymised datasets, e.g. to ENA.

At this stage however, the funding of these future activities remains unclear. An essential next step will therefore be to start identifying mid- and long-term funding sources to support our ambitious goals and the challenging workload ahead. Potential sources include future open calls under e.g. EU4Health<sup>6</sup>, JPIAMR<sup>7</sup>, as well as starting an ELIXIR Community on Federated Pathogen Data. We have also started discussing with BY-COVID to try to identify potential sources of bridge funding.

## 9. Deviation from Description of Action

Not applicable.

---

<sup>6</sup> [https://hadea.ec.europa.eu/news/2023-work-programme-eu4health-out-2022-11-22\\_en](https://hadea.ec.europa.eu/news/2023-work-programme-eu4health-out-2022-11-22_en)

<sup>7</sup> <https://www.ipiamr.eu/calls/>