Deliverable D8.1 Access to EU tools and project data

Project Title (Grant agreement no.):	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
Project Acronym (EC Call):	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
WP No & Title:	WP8 ELIXIR-CONVERGE European COVID-19 Data Platform		
WP leader(s):	Guy Cochrane (EMBL-EBI)		
Deliverable Lead Beneficiary:	1 - EMBL-EBI		
Contractual delivery date:	30/11/2020	**Actual delivery date:**	02/12/2020
Delayed:	No		
Partner(s) contributing to this deliverable:	EMBL-EBI		
Authors: Guy Cochrane (EMBL-EBI), Jeena Rajan (EMBL-EBI), Amonida Zadissa (EMBL-EBI) **Contributors:** EMBL-EBI Service Team Leaders **Acknowledgments (not grant participants):** *COVID-19-relevant data, literature and other resource providers*			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
30/09/2020	0v1	Guy Cochrane (EMBL-EBI)	Initial version
25/11/2020	0v2	Guy Cochrane (EMBL-EBI)	Sent to PMU after incorporating internal WP feedback
25/11/2020	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission
02/12/2020	0v4	Guy Cochrane (EMBL-EBI)	MB comments addressed
02/12/2020	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

Table of contents

1. Executive Summary

Rich and integrated biomolecular data are essential to support the world's scientific research into COVID-19. These data underlie our biological understanding (viral biology, the infection processes and patient response), our ability to identify and track COVID-19 (epidemiology and viral spread) and our capacity to intervene (through public health, therapeutic and vaccine approaches).

An early priority of the European COVID-19 Data Platform has been to bring together relevant biomolecular data and the scholarly literature describing these data along with tools, standards and other resources to support operations upon the data.

In this deliverable, we report on content-building activities within the Platform that serve, through the COVID-19 Data Portal, to make discoverable and accessible biomolecular data, literature and collections (tools, standards and related resources) relating to COVID-19 (https://www.covid19dataportal.org/).

Our content building has resulted to date in over 420,000 biomolecular data records, almost 110,000 literature records and 57 linked collections covering tools, standards and other resources. Data continue to grow in number and diversity.

This content has seen heavy exposure to users and two use cases around drug development and bioinformatics illustrate impact.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Our work builds upon and demonstrates the important foundations that are provided by well-managed biomolecular data and shows the potential of these data rapidly to be deployed to support urgent scientific challenges.

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:**
Objective 1: Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities **(WP1, WP5**)	
Key Result 1.1: 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)	**No**
Key Result 1. 2: Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	**No**
Key Result 1.3: The catalogue of successful national business models incorporated into national strategies	**No**
Key Result 1.4: The developed “sustainable and scalable operating model for transnational life-science data management support” is adopted into national ELIXIR Node	**No**
Objective 2: Strengthen Europe’s data management capacity through a comprehensive training programme delivered throughout the European Research Area **(WP2, WP6)**	
Key Result 2.1: 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.	**No**
Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**
Key Result 2.3: 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	**No**

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Objective 3: Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit **(WP2, WP3, WP5**)	
Key Result 3.1: 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.	**No**
Key Result 3.2: Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	**No**
Key Result 3.3: Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	**No**
Key Result 3.4: Enable 'FAIR at source' practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	**No**
Objective 4: Align national investments to drive local impact and global influence of ELIXIR **(WP4,WP6**)	
Key Result 4.1: Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	**No**
Key Result 4.2: Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	**No**
Key Result 4.3: Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	**No**
Key Result 4.4: Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	**No**
Key Result 4.5: Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	**No**
Objectives - WP8 - ELIXIR-CONVERGE European COVID-19 Data Platform	
O8.1 Data management support for EU projects (Task 8.1)	**Yes**
O8.2 Mobilisation of analysis upon SARS-CoV-2 sequence data (Task8.2)	**No**
O8.3 Enhanced access to data, tools and support (Task 8.3)	**No**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

3. Introduction

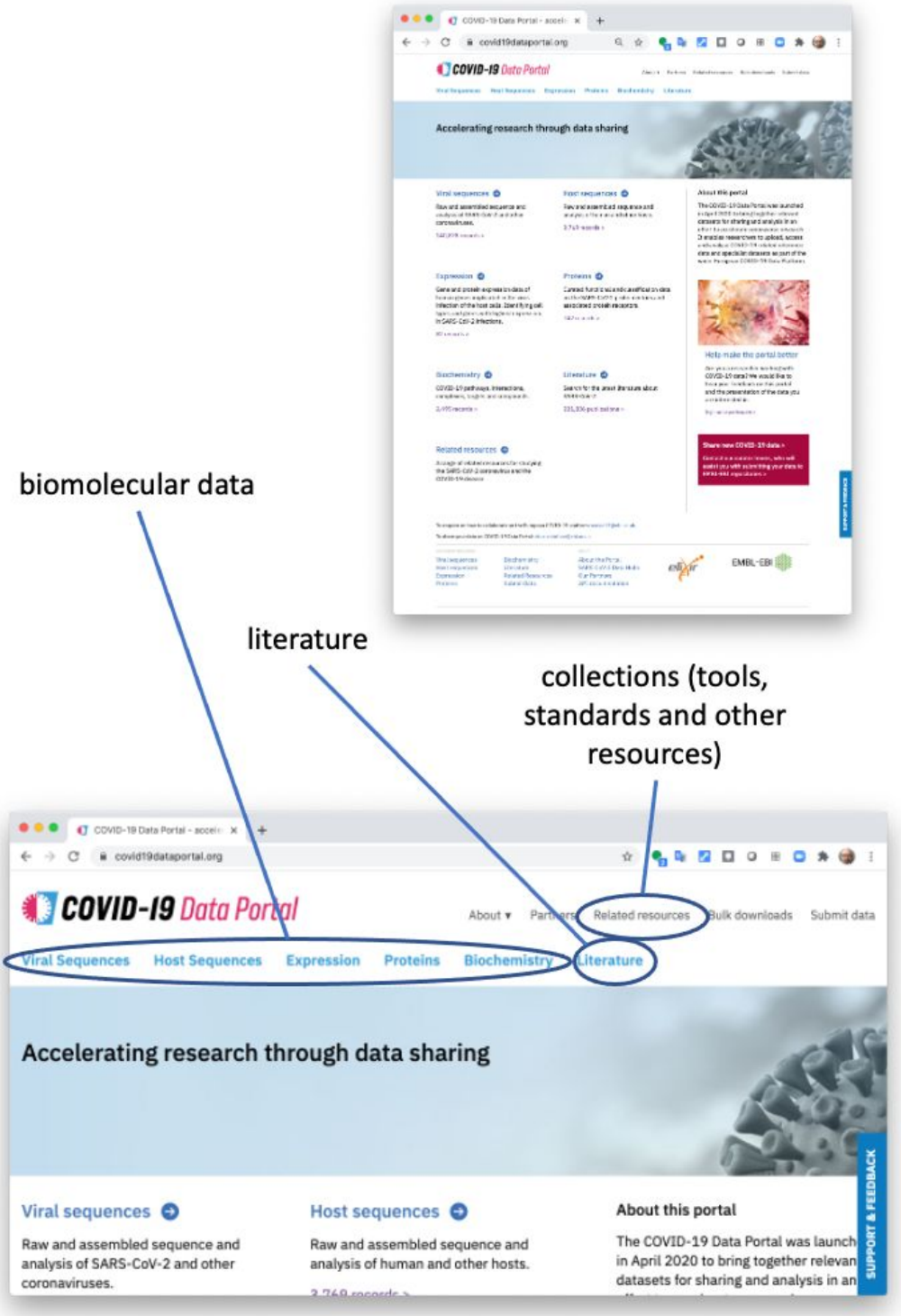
Rich and integrated biomolecular data are essential to support the world's scientific research into COVID-19. These data underlie our biological understanding (viral biology, the infection processes and patient response), our ability to identify and track COVID-19 (epidemiology and viral spread) and our capacity to intervene (through public health, therapeutic and vaccine approaches).

An early priority of the European COVID-19 Data Platform has been to bring together relevant biomolecular data and the scholarly literature describing these data along with tools, standards and other resources to support operations upon the data.

In this deliverable, we report on content-building activities within the Platform that serve, through the COVID-19 Data Portal, to make discoverable and accessible biomolecular data, literature and collections (tools, standards and related resources) relating to COVID-19 (https://www.covid19dataportal.org/; see Figure I). Spanning, for example, brokered SARS-CoV-2 sequence data (relates to ELIXIR-CONVERGE WP1) and human COVID-19 data (ELIXIR-CONVERGE WP7), our work builds upon and demonstrates the important foundations that are provided by well-managed biomolecular data and shows the potential of these data rapidly to be deployed to support urgent scientific challenges.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Figure I: COVID-19 Data Portal, showing left hand menu (biomolecular data and literature) and right hand "Related resources" link to tools, standards and related resources



This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

4. Description of work accomplished

4.1 Biomolecular data

In the life sciences, best practice dictates that biomolecular data generated by projects are deposited in appropriate data resources. The breadth of ELIXIR's data resources, including the ELIXIR Core Data Resources and the ELIXIR Deposition Databases, is such that the appropriate data resources for most biomolecular data are typically already within the ELIXIR family or, if not directly, are connected through data exchange collaborations such as the worldwide PDB or International Nucleotide Sequence Database Collaboration.

Our early profiling indicated substantial coverage of COVID-19-relevant bimolecular data already within ELIXIR data resources, including from EU projects. In many cases, these resources also had established routes, or initiatives to develop new routes, through which newly generated COVID-19-relevant data would be collected over time. Specific initiatives within the Platform, such as the mobilisation of viral sequence data through the SARS-CoV-2 Data Hubs (VEO and EOSC-Life), data brokering arrangements (ELIXIR-CONVERGE WP1) and support for human data flows (ELIXIR-CONVERGE WP7 and ReCoDID), are continuing to drive at richer COVID-19-related data.

We have maintained a continuous monitoring of COVID-19-relevant data as they appear in ELIXIR Core Data Resources with a view to maintaining a complete and growing data set that can be made available via the COVID-19 Data Portal. Working with different data resource managers, we have guided the labelling of COVID-19-relevant data such as through the addition of relevant flagging attributes. Where possible, this has included rule-based approaches to allow autonomous growth of relevant record sets within the different data resources. We have included into this overall data set records from additional data resources over time. Finally, using the infrastructure provided under EOSC-Life, we have configured the COVID-19 Data Portal to present the data set to users (https://www.covid19dataportal.org/).

4.2 Literature

Scholarly literature provides both an important narrative on our scientific understanding of COVID-19 and the context for, and description of, COVID-19-related biomolecular data. In parallel with our efforts in maintaining the biomolecular data set, we have established search parameters spanning the literature collections served through Europe PMC. Collections include PubMed abstracts, full-text open articles and, added following the emergence of the disease, COVID-19-related pre-print publications. We have presented these for search in the COVID-19 Data Portal (https://www.covid19dataportal.org/literature?db=literature).

4.3 Tools, standards and other resources

Development and adaptation of tools for the management, analysis and visualisation of COVID-19-related biomolecular data have flourished during the crisis. We have curated a set of tools (along with standards and other resources) for presentation in the "Related resources" (https://www.covid19dataportal.org/related-resources) section of the COVID-19 Data Portal.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5. Results

At the time of writing, the European COVID-19 Data Portal presents over 420,000 biomolecular data records, almost 110,000 literature records and 57 resources covering tools, standards and other elements of importance in COVID-19 research. Breakdowns are provided in Tables I and II.

Table I: current content spanning data and literature (available from left hand menu)

Menu item	Data type or section	Number of records
Vira sequences	Raw reads	124,713
	Sequenced samples	96,632
	Sequences	27,551
	Studies	163
	Variants	12,691
	Browser	1
	Genes	22
Host sequences	Association studies	8
	Human reads (consented for full access) / Other species reads	2,718
	Human studies (controlled access)	4
Expression	Gene expression	4
	Gene expression experiments	23
	Protein expression experiments	32
	Raw reads	19
	Single cell gene expression	4
Proteins	Electron microscopy density maps	250
	Electron microscopy public image archive	12
	Protein families	126
	Protein structures	260
	Protein structures - Knowledge Base	7
	Proteins	57
Biochemistry	Complexes	29
	Compound document	8
	Drug targets	390

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

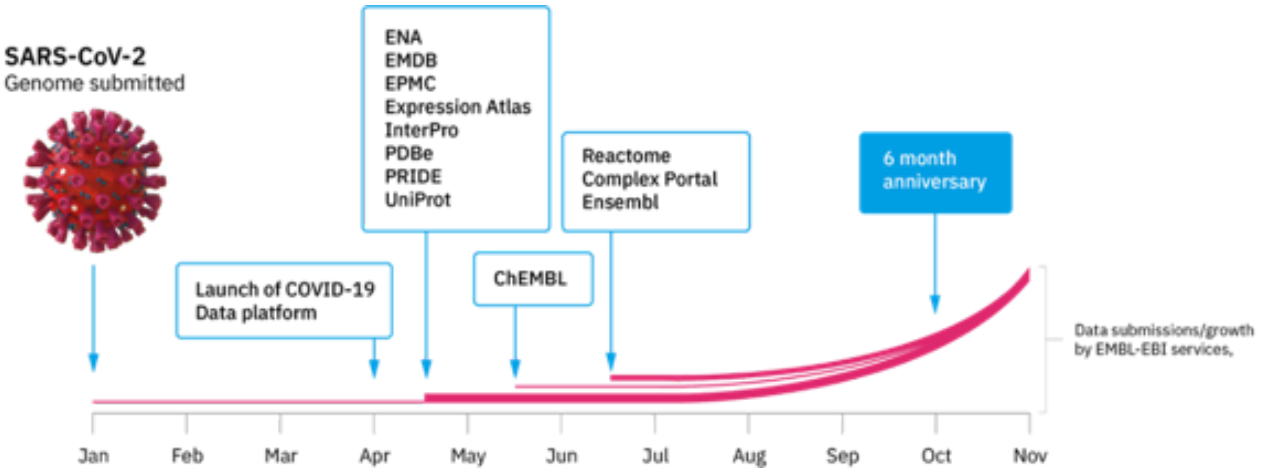
	Interactions	2,052
	Pathways	16
Literature	Literature	109,590

Table II: current content spanning collections of tools, standards and other resources (available under "Related resources"

Tools, standards and other resources	Databases and atlases	30
	Computing support	7
	Standards for data sharing	6
	ELIXIR publications	5
	ELIXIR activity and events	4
	Other European projects	5

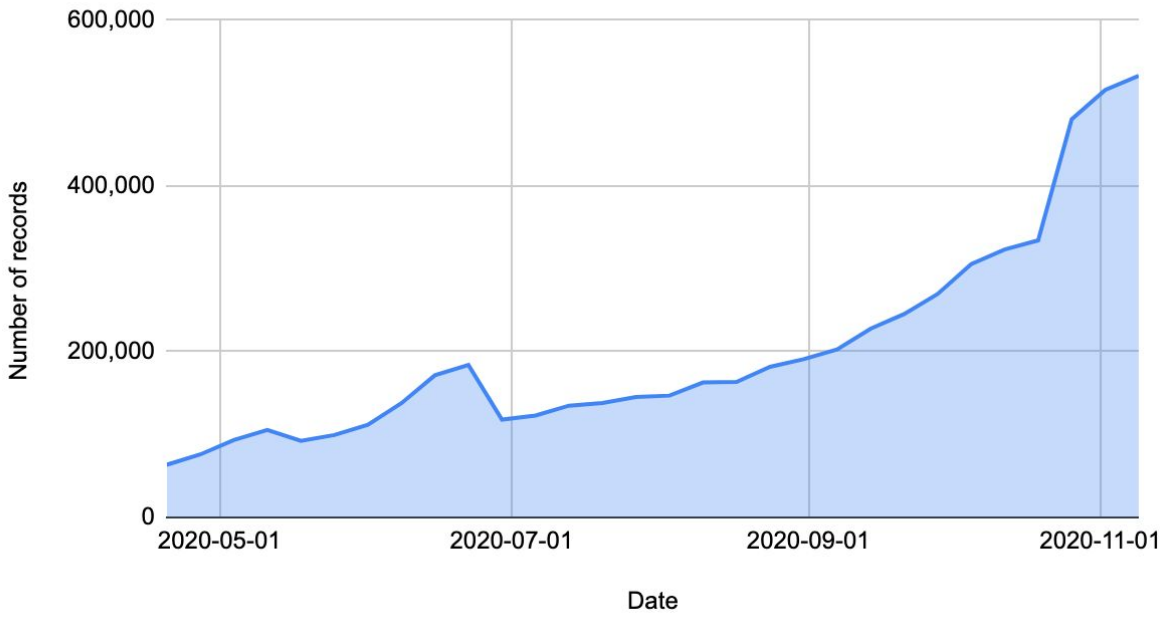
New data resources have been recruited and relevant records included in the data set since March 2020. Resources currently covered are ENA, EMDB, Expression Atlas, InterPro, PDBe, PRIDE, UniProt, ChEMBL, Reactome, Complex Portal, Ensembl, EMPIAR and Open Targets. Figure II shows the addition of data resources over time.

Figure II: data and literature resources shown from COVID-19 Data Portal up to October 2020; most recent additions EMPIAR and Open Targets have been added since the creation of the Figure



This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Figure III: data and literature growth in the COVID-19 Data Portal



6. Conclusions

We have identified from ELIXIR Core Data Resources and other sources sets of COVID-19-relevant records relating to biomolecular data, scholarly literature and tools, standards and other resources. We have worked with resource managers to label records with relevance to COVID-19, often through automated rule-based methods, to allow these to exist as a dynamic and growing data set. We have configured the COVID-19 Data Portal to show these records in a searchable and integrated way through the left hand data menu headings and the "Related resources" item. Numbers of resources included, along with the records from each of the resources, have grown rapidly over time.

7. Impact

The content made available through the COVID-19 Data Portal has been accessed over 3 million times by so far around some 100,000 distinct users.

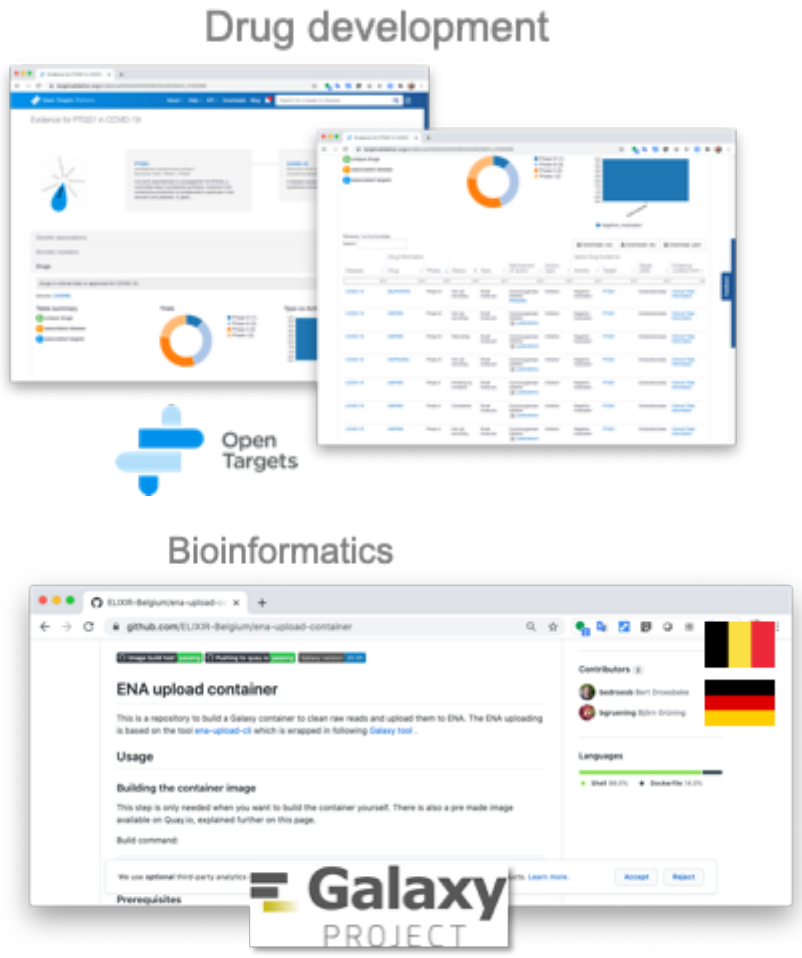
Two use cases illustrate the impact of this work (see also Figure IV):

- Use of target and compound data available from the COVID-19 Data Platform by the Open Targets group (https://covid19.opentargets.org/): data were accessed and curation/analytical processes applied to provide a prioritisation list of targets annotated with compounds and details of the status of trials with these compounds; given the relevance of this resource to the drug discovery/development community, we have since integrated this data resource into the COVID-19 Data Portal.
- Use of the Platform to support specialist viral sequence analysis and other workflows by the Galaxy Community (https://elixir-europe.org/news/ENA-new-tool-COVID-19-data): a suite of tools have been developed and made available through Galaxy to its user community to

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

read and write viral sequence data to/from the European COVID-19 Data Platform; these tools are being used in addition by ELIXIR-Belgium to support national viral sequence data flows.

Figure IV: drug development and bioinformatics use cases



8. Next Steps

Ongoing work will continue in the following areas:

- Continued tracking and identification of data resources of COVID-19 relevance
- Continued tracking of data records to assure discoverability from the COVID-19 Data Portal
- Enhanced indexing schemes to allow greater data breadth, including resources hosted beyond EMBL-EBI
- Enhanced indexing schemes for tools, standards and other resources to provide greater utility from the COVID-19 Data Portal

9. Deviation from Description of Action

Not applicable.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.