Deliverable D2.1

Initial data and metadata harmonisation at domain level to enable fast responses to COVID-19

Project Title (grant agreement No)	Beyond COVID Grant Agreement 101046203			
Project Acronym (EC Call)	BY-COVID			
WP No & Title	WP2: Accessing heterogeneous data across domains and jurisdictions for enabling the downstream processing of COVID-19 and future pandemic episodes data			
WP Leaders	Salvador Capella-Gutierrez (BSC), Alfonso Valencia (BSC), Aastha Mathur (EuroBioImaging), Antje Keppler (EuroBioImaging)			
Deliverable Lead Beneficiary	48 – BBMRI UK			
Contractual delivery date	30/06/2022	**Actual Delivery date**		31/07/2022
Delayed	Rescheduled			
Partner(s) contributing to this deliverable	ELIXIR-NL/VUmc, ELIXIR-NL/Lygature, IACS-ES			
Authors	Tom Giles (BBMRI UK) Phil Quinlan (BBMRI UK) Jeroen Belien (ELIXIR-NL/VUmc) Julia Lischke (ELIXIR-NL/Lygature) Laura Portell-Silva (ELIXIR-ES/BSC) Salvador Capella-Gutierrez (ELIXIR-ES/BSC) Reagon Karki (EU-Openscreen/Frauhofer ITMP) Vasso Kalaitzi (DANS)			

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Contributors	Enrique Bernal-Delgado (IACS-ES), Alfonso Valencia (BSC), Antje Keppler (EuroBioImaging)
Acknowledgements (not grant participants)	
Reviewers	Project Management Board

Log of changes

Date	Mvm	Who	Description
2022-06-28		Tom Giles	Initial commit to Drive
2022-07-19		WP2 Team	WP 2 team review -Final structure and division of writing agreed
2022-07-22		Julia Lischke / Enrique Bernal-Delgado	Section 4 major editing
2022-07-22		Tom Giles	Draft Sent to management board
2022-07-25		Tom Giles	Comments addressed
2022-07-28		Tom Giles / Laura Portell-Silva / Salvador	Rewrite of section 4 to consolidate, sections 5-7 addressed.

Funded by the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Table of contents

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

1. Executive Summary

BY-COVID Work Package (WP) 2 brings together data resources and catalogues across domains, captures data governance and access procedures. It will align metadata descriptions and other relevant semantic information first within domains (e.g., biomolecular and imaging, clinical and health, survey, etc) and in a second stage (in alignment with WP3 developments) expose a reference catalogue with harmonised metadata descriptions across domains.

Deliverable (D) 2.1, part of BY-COVID WP2, will develop the infrastructure required to facilitate access to a broad portfolio of data sources relevant to COVID-19 in preparation for future outbreak responses. Importantly, WP2 will collate, connect, and harmonise data sources within domains and disciplines in preparation of deeper cross-domain harmonisation efforts in WP3 and in connection with pathogen genome data (WP1). WP2 will further collate data governance models, metadata descriptions and access mechanisms ensuring full alignment with WP3 on the acquisition, curation and use of metadata from data sources. Interaction with WP4 will enable the development of tools for processing and harmonising metadata across data sources within the same domain. It also explores the initial infrastructure set-up activities undertaken to provide the capability for the rapid onboarding of resources in the future.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

2. Contribution towards project objectives

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

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

Funded by the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Objective 3 Link FAIR data and metadata on SARS-CoV-2 and COVID-19	1. A platform that links normative pathogen genomes and variant representations to research cohorts and mechanistic studies to understand the biomolecular determinants of variant response on patient susceptibility, and disease pathways.	Yes
	2. An open and extensible metadata framework adopted cross-domain that supports comprehensive indexing of the infectious disease resources based on mappings across resources and research domains.	Yes
	3. A provenance framework for researchers and policy-makers that enables trust in results and credit to data submitters, workflow contributors and participant resources.	Yes
Objective 4 Develop digital tools and data analytics for pandemic and outbreak preparedness, including tracking genomics variations of SARS-CoV-2 and identifying new variants of concern	1. Broad uptake of viral *Data Hubs* across Europe deliver an order-of-magnitude increase in open viral variant detection and sharing.	No
	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.	No
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.	No
	4. To empower national centres to build capacity and train platform users and data providers (e.g., from life, social or health sciences), and with experts from across partner institutions collaborating to create training materials for the identified gaps, and to exchange experiences and knowledge.	No

Funded by
the European Union

3. Scope and methodology

BeYond-COVID (BY-COVID) aims to link comprehensive data and metadata on SARS-CoV-2 and other infectious diseases across scientific, medical, public health and policy domains. The project will mobilise existing data resources, put together in a catalogue created by WP2, and marshal them for research, connect and expose them via the COVID-19 Data Portal, and drive use and analysis by connecting workflows, national portals and analysis environments.

D2.1 focuses on harmonising the resources, as described in the List of Resources identified in Milestone (M) 2.1, to the data and metadata standards outlined in WP3 as well as providing best practice guidelines for adding new resources and laying the infrastructure foundations to enable COVID-19 data access and transfer. It also feeds into the standards for pipelines, data curation and quality assurance, including semantic interoperability defined by WP4 and the continuous update of the use-cases from WP5. It forms a defined subset of Task (T) 2.5 “Harmonising data sources within the same domain across jurisdictions.” as it seeks to harmonise data sources within the same domain, including non-patient related resources, human/patient biomolecular resources, human/patient clinical and health data resources and socio-economics resources.

As many of these datasets are under restricted access control, the initial harmonisation has to be done at the metadata level with additional focus on capturing the data governance models, available metadata descriptions and other elements to enable semantic interoperability in accordance with the standards defined by WP3 (see D3.1) and 4. As part of this deliverable, we also present a set of guidelines on how to incorporate future resources as required in response to new outbreaks. The work outlined in this report also acts as a precursor to T2.6 “Establishing the systems to enable data discovery and training”, which seeks to enable data discovery at origin using existing technologies, e.g., GA4GH Beacon, and extending them, if necessary.

4. Description of work accomplished

4.1 Metadata and Data harmonisation

BY-COVID will consolidate the technical basis for cross-domain discoverability and interoperability, and set out the requirements for interoperability of data (best practices, community associated standards, harmonisation and management of metadata, sample identifiers/persistent identifiers, vocabularies). Particular attention is paid to sensitive data needs and indexing of both data and metadata to preserve trust and privacy. The COVID-19 Data Portal will act as a one-stop shop for congregating all COVID-19 related data including the viral Data Hubs as well as national instances of the COVID-19 Data Portal and platforms holding other data types such as social science data, public health data, or epidemiology data in a truly multidisciplinary effort.

A comprehensive review of metadata standards was carried out by D3.1. The team responsible for the delivery identified that the EOSC architecture and Interoperability Framework is the most interoperable metadata standard for adoption by a wide range of data sources represented by the BY-COVID project. The common set of metadata elements are presented in D3.1 Appendix 1. It includes crosswalks to BioSchemas and DataCatalog Vocabulary (DCAT), but the common set can be expressed in various metadata standards/formats.

D3.1 presented a flexible, tiered system for metadata integration. For tier 1, a limited number of key resources, including the CESSDA Data Catalogue, part of T2.4 on “Relevant socio-economics data sources for infectious disease outbreaks”, will be deeply indexed, capturing granular, record level identifiers and detailed metadata. For these resources, the indexing strategy will support complex interoperability tasks implemented in WP2 and WP5. In tier 2, a broader range of resources will be indexed with a focus on record level discoverability, with coarse-grained metadata, but limited and iteratively refined metadata harmonisation. This tier will support deep discovery of relevant datasets from large resources in the multi-disciplinary COVID-19 space. In tier 3, additional resources will be included in the COVID-19 Data Portal at the resource level only, supporting high level discovery of relevant resources, but delegating record level searches to the relevant resources themselves. Tier 3 will allow us to support discovery of a broad range of relevant resources, while avoiding complex metadata harmonisation challenges for a number of resources beyond the feasible scope of the project.

M2.1 stratified the List of Resources into 4 primary domains of data resources: 1) Non-patient related, 2) Human/patient bio-molecular, 3) Human/patient clinical health, and 4) Socio-economic data. Because of the diversity of the data sources between

Funded by
the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

domains, initial harmonisation efforts have been focused on ensuring tier 3 discoverability of the resources in the COVID-19 Data Portal and ensuring FAIR activities are being undertaken by the data partners to ensure interoperability as we move towards deliverable D2.2.

4.1.1. Non-patient data

The non-patient data resources part of WP2 include diverse data emerging from Structural, Bioimaging and Bioactivity studies. Owing to its non-sensitive nature, such data is openly available through dedicated repositories that employ different metadata standards and controlled vocabularies. This is summarised in the M2.1 List of resources and the BY-COVID data resource collection in FAIRsharing.

Entity	Ontology/Dictionary	Metadata
Human Proteins	HUGO	Uniprot, ChEMBL
Assays	ChEMBL	IC50 vals, Type, Organism, pChEMBL
Chemicals	ChEMBL, PubChem	CAS, SMILES, Physicochemical props.
Pathways	Reactome, WikiPathway	Pathway IDs
Mechanism of action / Process	Gene Ontology	GO components
Disease Indications	MeSH, OMIM	EFO ids, synonyms

Table 1*: Overview of ontologies and identifiers used by data resources associated with T2.1.*

The majority of the non-patient data is open access and already has programmatic access mechanisms in place to access the resource, e.g. via REST APIs. While openness of the data allows easy access to the data itself, the diversity of methods and technologies producing the data makes metadata harmonisation more demanding. Hence the first harmonisation efforts in this domain are being attempted for data from the sub-domains.

The non-patient data harmonised so far include active chemicals tested against SARS-CoV-2, from EU-OPENSCREEN partner institutes, and associated target proteins, mechanism of actions, Gene Ontology components (biological processes and molecular function), and pathways (Reactome). Additionally, thanks to a collaboration between Fraunhofer ITMP and Harald Schwalbe's NMR group in Frankfurt, putative active fragments in a multi-centric NMR screening experiment against expressed SARS-CoV-2 proteins were also successfully harmonised. Lastly, the COVID-19 disease maps built in BY-COVID WP5 have also been integrated and aligned. In this task of data harmonisation, we have followed the recommended standard ontologies and annotation thereby ensuring

FAIRness of overall data. In this process, we have converted different identifiers of chemicals such as CAS number and PubChem compound identifier (cid) to ChEMBl id for uniformness. Similarly, protein entities from ChEMBL and UniProt were converted to HUGO gene symbols. The next steps will consider integration of structural and bioimaging data from other WP2 partners.

4.1.2 Human/patient bio-molecular resources

EMBL-EBI has acted swiftly in response to COVID-19 by launching new initiatives and repurposing existing facilities to provide a range of direct research and support services as part of the BY-COVID project. To support responsible data sharing for cohort studies in line with FAIR principles, i.e., making data as open as possible but as closed as necessary, several resources are currently available or are being implemented as part of the European COVID-19 Data Portal.

There are two primary tools being developed that relate to the Human/patient bio-molecular resources.

1. The Cohort Browser will become an integral part of COVID-19 Data Portal and primary entry point into cohort data. It lists discovery metadata of COVID-19-related cohort studies and indicates which data types are available. Wherever possible, basic aggregate data are also included. Search and filtering functionality will allow the users to locate datasets of interest.
2. The BioSamples database at EMBL-EBI is the central institutional repository for sample metadata storage and connection to EMBL-EBI archives and other resources. The primary function of the BioSamples database is to link different data types on participant level. This is based upon established research infrastructures including the ELIXIR core data and deposition resources.

Within the portal, cohort information is displayed in the EBI pathogens browser. The European Genome-phenome Archive (EGA) uses a custom metadata model to maintain the information it holds around samples and data. Whilst these resources were already registered as part of the core repositories used to provide content to the Host Sequences section, they were not listed in the Cohort Browser. Work is currently underway to transform the metadata for these collections so that they are represented across both resources. However, effective linkage between these resources will be required to ensure usability. The EGA genomics data is a part of the core EMBL-EBI Nucleotide Archive and is fully available via the EGA Rest API and all the samples relating to this project are registered on the Host Sequences section of the COVID-19 Data Portal.

M2.1 also identified several other human/patient bio-molecular resources. The Dutch COVID-19 Data Support Programme metadata contains all mandatory fields of the DCAT standard, some recommended DCAT fields and some domain and funder specific fields. So, interoperability with other catalogues and collections, using DCAT including the XML scheme outlined by D3.1 was already ensured. The metadata is collected using the CEDAR Workbench utilising the CEDAR metadata templates and thus is machine readable and will be automatically harvested by the COVID-19 Data portals Mendix platform. All metadata templates created and consolidated were agreed in various M4M (metadata for machine) workshops with experts from the Dutch clinical fields are publically available. The used ontologies are adapted or published on BioPortal. Thus, this process follows open science recommendations ensuring the generation of FAIR data and enhancing scientific reproducibility. All resources on the Dutch COVID-19 Portal are directly available via an API in both JSON-LD and RDF format. A dedicated team of domain experts and data specialists from both the GO-FAIR foundation and Health-RI have been working together with the data stewards and scientists of each funded project towards a layered approach of making the projects' metadata FAIR. Work on the metadata templates for catalogue (if needed by project), dataset and distribution are still work in progress. The resulting ontology on generic terms and ontology on content have been published at BioPortal for re-use by others. Next, Health-RI, the NL funders (ZonMw and NFU) are working together with the GO-FAIR Foundation and the Dutch University Medical Centres to make the metadata also accessible via a FAIR Data Point. They are also working on a Proof of Concept for an integrated approach towards improved access to healthcare data for both primary and secondary use. For secondary use within research this will be based on solutions like OMOP on FHIR (see also next paragraph clinical and population health data).

The resources presented by The University of Nottingham, an associated organisation of BBMRI-ERIC, align with the UKCRC Tissue Directory and Coordination Centre / BBRMI-ERIC MIABIS minimum information about biobank standards. The metadata for these resources have been mapped to the COVID-19 Data portal metadata standard and these data resources have been passed to the portal team for upload to the COVID-19 Data Portal. These resources have already been mapped to OMOP using the CaRROT toolkit described in 4.2.2 and are already discoverable via federated approaches on both the UK Health Data Research Cohort Discovery tool and the BBMRI finder. RO-Crates of these resources are currently being created to track the providence and quality of the data transformation and part of the development of the HUTCH toolkit semantic interoperability with other endpoints being considered (including the GA4GH beacons and any potential federated endpoint developed as part of BY-COVID WP2 or WP4).

4.1.3 Clinical and population health data

Funded by
the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Typical data sources in clinical and population health are Electronic Health Records (EHR), Electronic Medical Records (EMR), Labs data, Imaging data, Claims data, Disease-registries (clinical data), or Population registries and Surveys (population data). These datasets, collected with the primary purpose of providing either care to patients or doing epidemiological surveillance, are rarely collected with a view to be reused for secondary or tertiary purposes.

These data sources / data collections are usually prepared *ad hoc* in the context of clinical projects or in the development of population cohorts. Most follow, at least partially, the FAIR data principles and can only be made via data access requests. In this case, a data controller that collects and maintains the data is the one in charge of cataloguing and access.

At the syntactic level, there are a myriad of metadata standards that could be of use for publication and discovery of this type of data. However, there are some well recognised initiatives on interoperable cataloguing worth considering (see table below)

Acronym	Domain	Information (URL)
DCAT-AP2	Public reporting of data catalogues	https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en
INSPIRE	Health data geo-allocated	https://inspire-geoportal.ec.europa.eu/#
PHIRI	Population health data collections and research studies	https://www.healthinformationportal.eu
CEDAR	Biomedical experiments	https://metadatacenter.org/
ECRIN	Clinical research	https://ecrin.org/tools/clinical-research-metadata-repository
EMA	Observational research in regulation	https://www.ema.europa.eu/en/documents/other/list-metadata-real-world-data-catalogues_en.pdf
HDRUK	Datasets and studies cataloguing	https://web.www.healthdatagateway.org/
BBRMI	Clinical Sample Catalogue	https://negotiator.bbmri-eric.eu/researcher/index.xhtml

***Table 2**: Well-recognized data catalogues for COVID-19 related data sets and collections.*

The most important task in both the data and metadata harmonisation for the Clinical and population health data is ensuring that data is semantically interoperable despite being available through different data sources and/or sites, sometimes from different countries.

Most structured data collected in clinical and public health settings are usually encoded following standardised dictionaries, taxonomies or ontologies. For example, clinical conditions and procedures performed in a hospital admission, or health events in a population registry are usually codified with Disease Classifications as ICD, OPCS, NOMESCO; clinical episodes in primary care are often encoded using the International Primary Care Classifications, while laboratory information is encoded using LOINC. More recently there is a drive for the use of ontologies as SNOMED-CT for clinical episodes or ORDO in the case of rare diseases.

When in secondary use, and the research query implies different data sources, a semantic interoperability layer is usually required. In some cases, the aforementioned ontologies may serve to provide a semantic layer when processing data. However, often a common data model (CDM) is required to ensure interoperability, especially when comparing resources from diverse sources. Common examples of such transport layers include CDISC, OMOP and FHIR. Perhaps the most interesting recent development in this area is that HL7 International and OHDSI have announced a collaboration to provide a single CDM for sharing information in clinical care and observational research environments, which means that the OMOP and FHIR standards will be eventually merging.

4.1.4. Socio-economics

The main relevant data sources selected for socio-economics, with regards to infectious disease outbreaks and the BY-COVID project are the CESSDA Data catalogue and the EUI COVID-19 SSH Data Portal. Secondary relevant data sources have also been identified for potential expansion at a later stage (e.g., European Social Survey, SHARE, Dutch COVID-19 Data Support Programme). The primary and secondary data sources with regards to socio-economics have been identified early on in the project and confirmed at the M2.1 stage. These identified data sources have been registered in the BY-COVID reference catalogue.

With regards to the relevant primary data sources and their access policies, the CESSDA Data Catalogue is a metadata catalogue that does not distribute data and links to original archives where data may be accessed. At the same time, the COVID-19 Data Portal is a registry that, for the time being, does not host data and equally links to the original archives. The CESSDA Data Catalogue is in FAIRsharing and uses the Data Documentation Initiative standard (DDI2.5), while the COVID-19 Data Portal is not in FAIRsharing and uses the internal InvenioRDM JSON Schema, compliant with the DataCite Metadata Schema 4.4.

Both the CESSDA Data Catalogue and the COVID-19 Data Portal provide metadata records using the Dublin Core (DC) Metadata Specification. The harvesting tool implemented in the context of T2.4 has been specifically designed to implement the extended DC metadata scheme through OAI-PMH harvesting mechanisms. The filtered records are fully harmonised and will be exposed on COVID-19 Data Portal after the transformational processes necessary for the final exposure have been applied.

Funded by
the European Union

4.2. Infrastructure to enable COVID-19 data access and transfer

4.2.1. Harvesting tool for the COVID-19 Data Portal

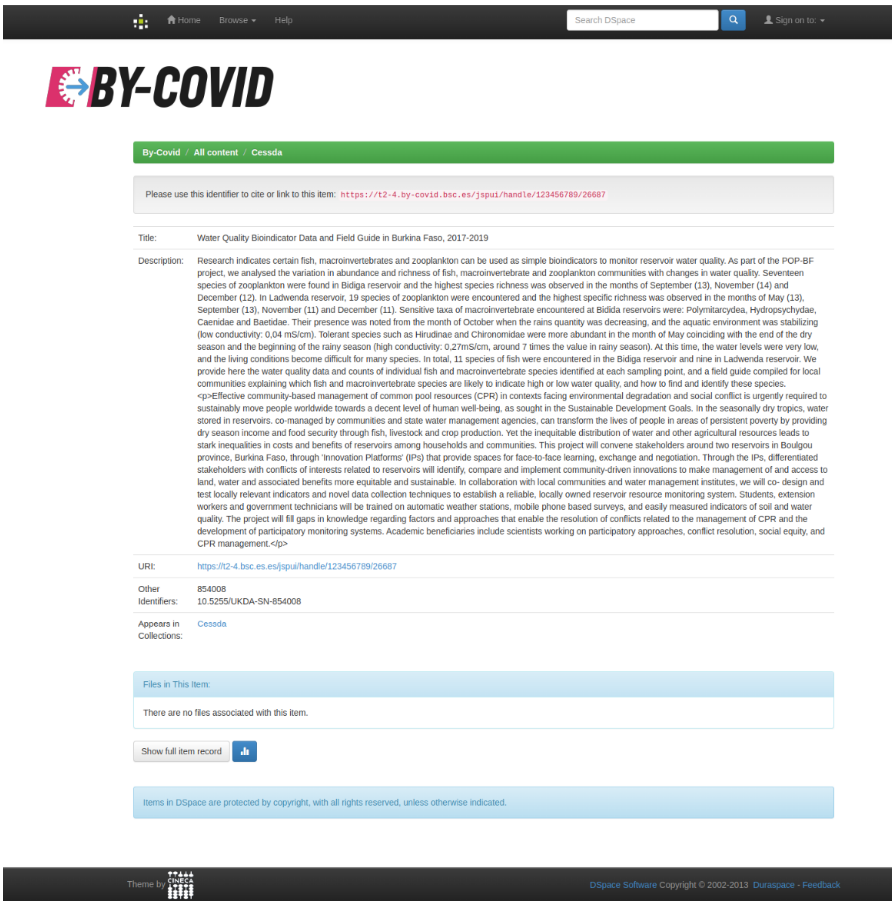
For the needs of T2.4 on “Relevant socio-economics data sources for infectious disease outbreaks”, an advanced harvesting tool has been implemented with the use and extension of the open-source software DSpace. The UI of the harvesting tool can be accessed here: https://t2-4.by-covid.bsc.es/jspui/.

The main purpose of the harvesting tool is the collection, processing, retrieving, searching and browsing of the harvested metadata records. It can be customised to support different metadata schemes and different content providers. The basic protocol used for the harvesting process is OAI-PMH, which is supported by most repositories world-wide.

This resource provides definitions of collections and communities: Communities can maintain an unlimited number of collections in DSpace. Collections can be organised around a topic, or by type of information, e.g., working papers or datasets, or by any other sorting method a community finds useful in organising its digital objects. For the purpose of the BY-COVID project, each collection corresponds to a metadata provider, e.g., CESSDA, EUI, etc. Using this resource millions of metadata records can be aggregated. The metadata processes can be periodically or manually triggered. Users can browse the harvested metadata records using keywords, topics or every other metadata field as well as query available metadata using advanced search functionalities. The search engine for the repository is the SOLR, a highly reliable, scalable and fault tolerant search engine, that provides distributed indexing, replication and load-balanced querying, automated failover and recovery, centralised configuration and more.

The harvesting tool has been extended with a set of endpoints that allow for the search and retrieval of the metadata records programmatically based on specific business needs, including the consumption of the metadata records to the COVID-19 Data Portal. The API is based on Open API principles (RESTful) and provides JSON responses.

The source code of the harvesting tool is currently hosted in a private repository in https://gitlab.com/ but will be made publicly available in due course.



Please use this identifier to cite or link to this item: `https://t2-4.by-covid.bsc.es/jspui/handle/123456789/26687`

Title:	Water Quality Bioindicator Data and Field Guide in Burkina Faso, 2017-2019
Description:	Research indicates certain fish, macroinvertebrates and zooplankton can be used as simple bioindicators to monitor reservoir water quality. As part of the POP-BF project, we analysed the variation in abundance and richness of fish, macroinvertebrate and zooplankton communities with changes in water quality. Seventeen species of zooplankton were found in Bidiga reservoir and the highest species richness was observed in the months of September (13), November (14) and December (12). In Ladwenda reservoir, 19 species of zooplankton were encountered and the highest specific richness was observed in the months of May (13), September (13), November (11) and December (11). Sensitive taxa of macroinvertebrate encountered at Bidida reservoirs were: Polymitarcydea, Hydropsychydae, Caenidae and Baetidae. Their presence was noted from the month of October when the rains quantity was decreasing, and the aquatic environment was stabilizing (low conductivity: 0,04 mS/cm). Tolerant species such as Hirudinae and Chironomidae were more abundant in the month of May coinciding with the end of the dry season and the beginning of the rainy season (high conductivity: 0,27mS/cm, around 7 times the value in rainy season). At this time, the water levels were very low, and the living conditions become difficult for many species. In total, 11 species of fish were encountered in the Bidiga reservoir and nine in Ladwenda reservoir. We provide here the water quality data and counts of individual fish and macroinvertebrate species identified at each sampling point, and a field guide compiled for local communities explaining which fish and macroinvertebrate species are likely to indicate high or low water quality, and how to find and identify these species. <p>Effective community-based management of common pool resources (CPR) in contexts facing environmental degradation and social conflict is urgently required to sustainably move people worldwide towards a decent level of human well-being, as sought in the Sustainable Development Goals. In the seasonally dry tropics, water stored in reservoirs. co-managed by communities and state water management agencies, can transform the lives of people in areas of persistent poverty by providing dry season income and food security through fish, livestock and crop production. Yet the inequitable distribution of water and other agricultural resources leads to stark inequalities in costs and benefits of reservoirs among households and communities. This project will convene stakeholders around two reservoirs in Boulgou province, Burkina Faso, through 'Innovation Platforms' (IPs) that provide spaces for face-to-face learning, exchange and negotiation. Through the IPs, differentiated stakeholders with conflicts of interests related to reservoirs will identify, compare and implement community-driven innovations to make management of and access to land, water and associated benefits more equitable and sustainable. In collaboration with local communities and water management institutes, we will co- design and test locally relevant indicators and novel data collection techniques to establish a reliable, locally owned reservoir resource monitoring system. Students, extension workers and government technicians will be trained on automatic weather stations, mobile phone based surveys, and easily measured indicators of soil and water quality. The project will fill gaps in knowledge regarding factors and approaches that enable the resolution of conflicts related to the management of CPR and the development of participatory monitoring systems. Academic beneficiaries include scientists working on participatory approaches, conflict resolution, social equity, and CPR management.</p>
URI:	https://t2-4.bsc.es.es/jspui/handle/123456789/26687
Other Identifiers:	854008 10.5255/UKDA-SN-854008
Appears in Collections:	Cessda

Files in This Item:

There are no files associated with this item.

Items in DSpace are protected by copyright, with all rights reserved, unless otherwise indicated.

Figure 1*: An illustration of a metadata record from the Harvesting tool, including additional identifiers to gather further information. This result is part of the CESSDA collection. This record is an example visualisation of information gathered by the harvesting tool before generating the XML that is exposed to the COVID-19 Data Portal*.

4.2.2. CaRROT - Convenient and Reusable Rapid OMOP Transformer

As more resources are on-boarded to the COVID-19 portal it is likely that some of them will contain identifiable data which is within the scope of the Data Protection Act 2018. With these resources it is often not possible to move the data in order to harmonise these to common data models for FAIR reuse due to governance constraints. Thus, the responsibility of harmonisation of such data often falls on the data controllers directly.

CaRROT has been developed in response to this constraint. CaRROT is a low risk, GDPR compliant tool that allows remote data mapping by only operating on metadata that is outside of the scope of data governance regulations. This allows teams with specialist expertise in data mapping to assist data partners in undertaking an ETL (Extract, Translate and Load) process to a common data format without ever having access to the underlying data.

Data partners are required to prepare the data by presenting a pseudonymised view of it as a series of CSV files (with one representing demographic data; all measurements in the metric system; dates and datetimes in the ISO-8601 format). The pseudonymised dataset is profiled by the data holder with the open-source data profiling tool White Rabbit. This generates a scan report that details the tables and fields of the dataset, along with values in each field. This output should be checked by the data holder before it is transferred to the team performing the mapping to remove any data values which could be deemed as confidential or sensitive. A Data Dictionary may be optionally supplied when the field names or values are not self-explanatory. This can provide (a) vocabularies associated to certain fields; and (b) descriptions of values. For example, it would be possible to specify that a column contains one of the existing vocabularies represented within the OMOP CDM (eg SNOMED, ICD9, ICD10 or RxNorm).

A web-based tool built using the Django Python web framework, React JavaScript and PostgreSQL is provided to assist in the generation of rules. A core reason for establishing a central tool for creating rules is so that they can be reused across projects. Upon upload of the scan report, the mapping tool will record all tables, fields, and values within that scan report. This enables users to visually navigate the structure of the dataset for the purposes of manual mapping. This is organised in a hierarchical manner, allowing the user to select a field or value and mark it with the desired target OMOP concept code. Previous similarities can be reused. For example, assuming a field "Gender" is provided with the value "M", and in a previous dataset this has been mapped to the OMOP concept 8507 "Male", then the system automatically applies the previous mapping rule to the new dataset. In this way, the utility of the system increases over time, and the manual effort required to generate mapping rules is reduced, particularly in the case where a new Scan Report is supplied that describes a dataset that has already been previously or partially mapped. At any stage in the mapping process, the user can see a summary view of the mapping rules currently extant, and review, discuss and remove any mapping rules. Once a user is satisfied that they have mapped all of the required fields and values, they can set the Scan Report status as "Mapping Complete". This process makes the mapping rules associated with that Scan Report available for reuse. Work-in-progress is not mined for rules to apply to similar Scan Reports.

Once the mapping process is complete the generated rules are then exported in CSV format (for human readability) and in JSON format (for return to the data partner). Data partners are then required to use these rules to convert the dataset to OMOP. A Python ETL package, available on Pypi is provided to assist with this task. Comprehensive documentation is available to guide the user in setting up and running the toolkit. This can either be used once to convert the dataset, or set to watch a directory for changes to data, triggering a re-mapping of the dataset.

Funded by
the European Union

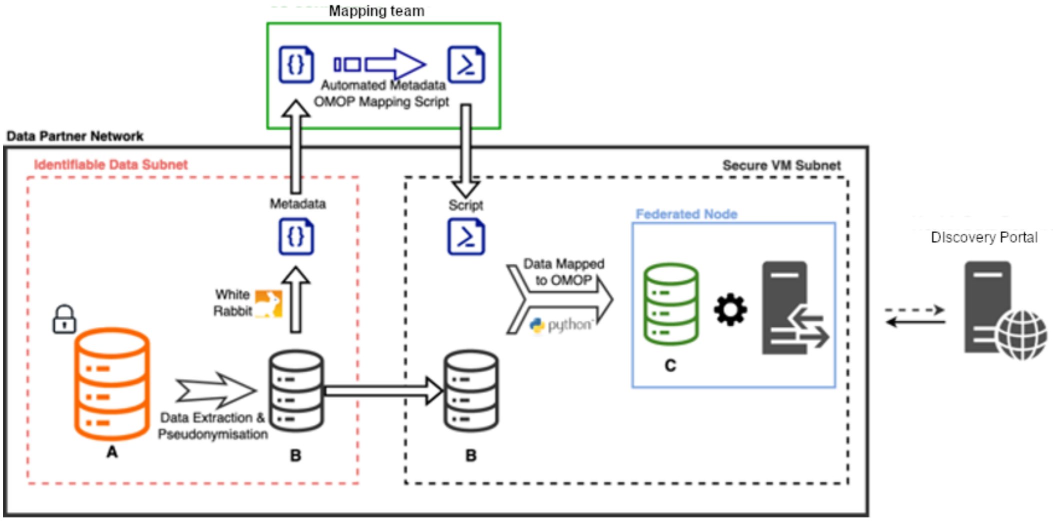


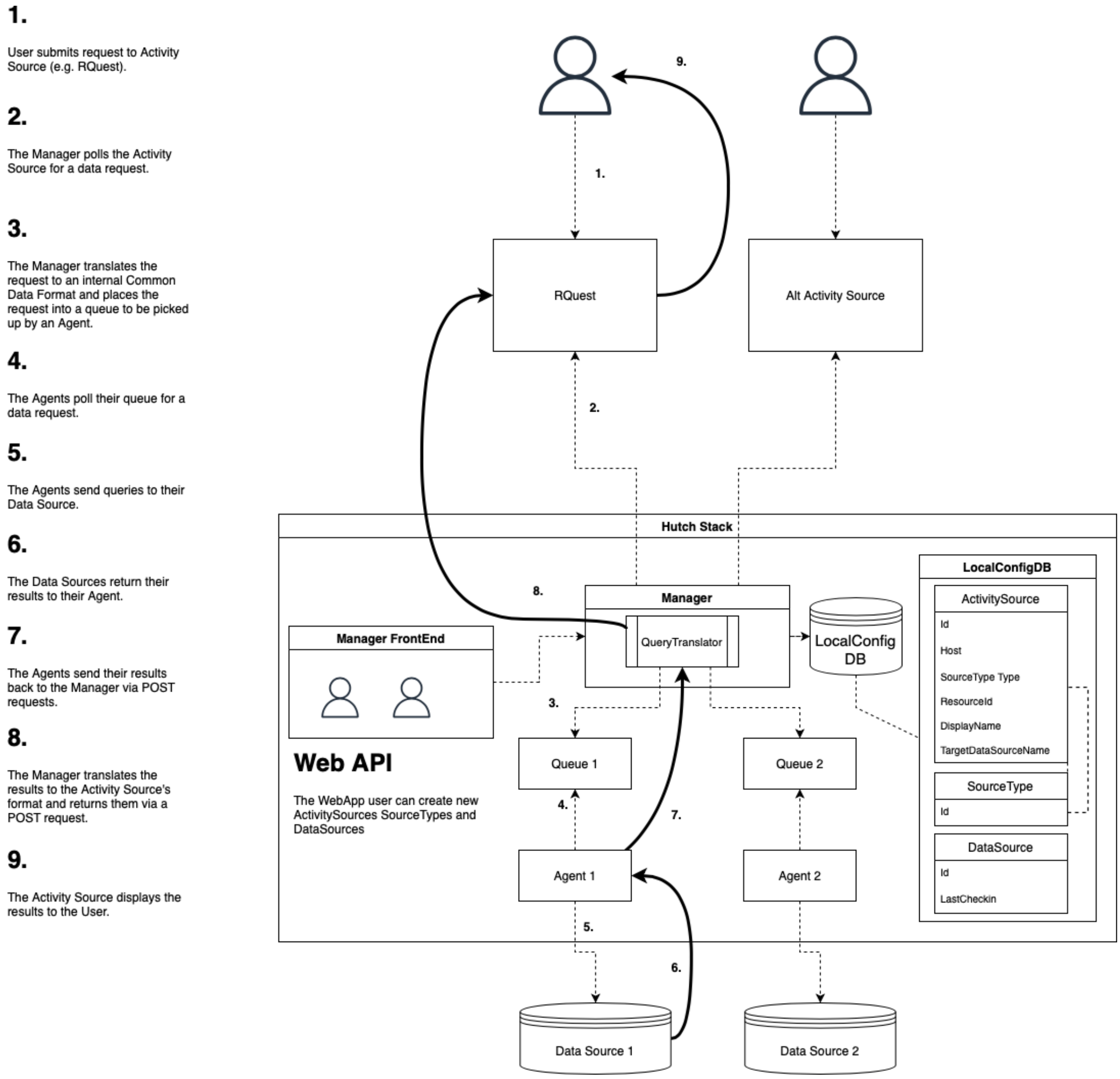
***Figure 2**: An example of a federated architecture using the CaRROT toolkit. Each data partner (dark box) has their identifiable data (A, red dashed box) from which an extraction is made and pseudonymised (B, red dashed box). A metadata extraction is performed with White Rabbit and a mapping script to the OMOP CDM created by the external data team (green box). The pseudonymised data are securely transferred (B) into a hosted secure VM (dashed dark box), mapped to OMOP and from there the data are made queryable by federated mechanisms. Only aggregated, fully anonymous data ever leaves the data partners environment.*

4.2.3. HUTCH - Heterogeneous Data Connector for Healthcare

A number of the resources identified in M2.1 are restricted by governance and cannot be harmonised centrally. For these resources a federated approach will be required to make the data discoverable for reuse in research. HUTCH is being developed from the perspective of an infrastructure layer to make secure healthcare data that, due to governance constraints, cannot be shared without approval, discoverable.

The focus of this tool is to provide a resource that can be deployed at either institutional or research group level and be interoperable with the current shifting landscape of data standards. It will be capable of pooling a range of API (Outbound) to look for queries. Users can view and configure Activity Sources (i.e., A collection on a discovery tool), Data Sources, i.e., OMOP on PSQL, and Source Types, e.g., BC|Platforms RQUEST (figure 3). Activity Sources interact with Agents via the Manager. The Manager regularly polls Activity Sources for queries via an outbound only connection. Queries are processed via a translational layer to a schema.org based internal schema. Using the stored Activity Source, Data Source and Source Type information the manager creates jobs to be processed by the Agents and presents these on a queue. Agents pull queries from the queue, process them and return results back to a separate Manager queue, where they are converted back to the original format of the Activity Source via the translational layer and thus returned to the Activity Source (see figure 3). Agents also periodically send a check-in signal to the manager, which the user can see in the UI confirming the health status of the underlying data resources.

A manager UI is provided to allow for the effective management of connections between different databases and external query portals in a secure manner (fully centralised audit tracking via logs, user managed obfuscation control / control of connected portals on a per database basis).



1.

User submits request to Activity Source (e.g. RQuest).

2.

The Manager polls the Activity Source for a data request.

3.

The Manager translates the request to an internal Common Data Format and places the request into a queue to be picked up by an Agent.

4.

The Agents poll their queue for a data request.

5.

The Agents send queries to their Data Source.

6.

The Data Sources return their results to their Agent.

7.

The Agents send their results back to the Manager via POST requests.

8.

The Manager translates the results to the Activity Source's format and returns them via a POST request.

9.

The Activity Source displays the results to the User.

***Figure 3**: A Schematic Diagram illustrating the standard flow of data within HUTCH. Each component is part of a modular stack allowing the resource to scale up to match an institution's infrastructure requirement. Stack information: React frontend, .NET backend API, PGSQL DB, RabbitMQ queue, and Python agents. Currently deployed in Azure. The HUTCH MVP will be query language agnostic, this means that queries from Activity sources in OMOP format will be processed only on OMOP Data Sources by OMOP agents and metadata queries will only be processed in RO-Crates format by RO-Crates Agents. Cross linkage of query and metadata languages may be explored in the future, but this is not on the current roadmap.*

Within HUTCH interactions between the components of the system are logged (to check that things like obfuscation are functioning effectively) and to maintain a paper trail of all interactions with Activity Sources for audit. The internal query language is being built on schema.org to ensure interoperability with a variety of Data Sources, i.e., FHIR on

Funded by
the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Databricks, and Source Types, i.e., GA4GH beacons. Using RO-Crates the structure, history, authorship, and providence of the underlying data resources will also be captured and made discoverable programmatically. This will allow us to return (on request from an Activity Source) detailed information about the Data Sources that could either be used to automatically populate metadata catalogues or to build a framework for federated analytic queries.

4.2.4. BY-COVID Knowledge Graph: A comprehensive network representation of chemical and biological worlds of COVID-19

As required by WP2 and WP3, there is a need to harmonise different types of data such as structural data, bioimaging data, bioactive molecules and COVID-19 biology. Taking this into consideration, we have generated a BY-COVID Knowledge Graph by integrating and harmonising entities with standard ontologies and metadata annotations (Figure 4). The BY-COVID Knowledge Graph is represented in the form of triples (source-relation-target) using an open source Python (ver 3.10) package called Pybel. Its framework also provides a library of functions for querying and analysis of the BY-COVID Knowledge Graph.

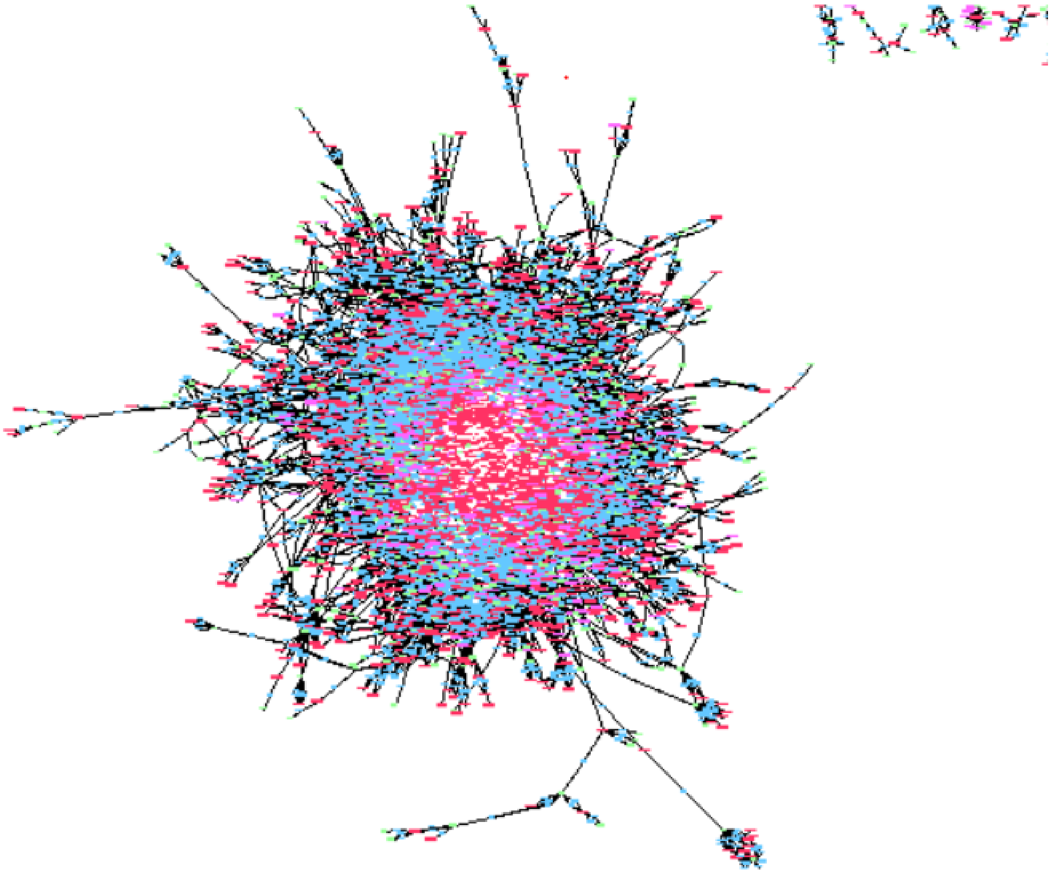


Figure 4: A snapshot of the BY-COVID Knowledge Graph.

Funded by
the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

The chemicals and targets used in partner institutes have been used as a primer to create the BY-COVID Knowledge Graph. In this regard, we have used the data mentioned in (4.1.1. Non-patient data). To this extend, we have extended the BY-COVID Knowledge Graph using the ChEMBL and UniProt API to fetch additional information about assays, Gene Ontoloogy components, diseases, pathways and mechanism of actions (Figure 5). So far, the BY-COVID Knowledge Graph comprises 15000+ and 85,000+ nodes and relationships, respectively.

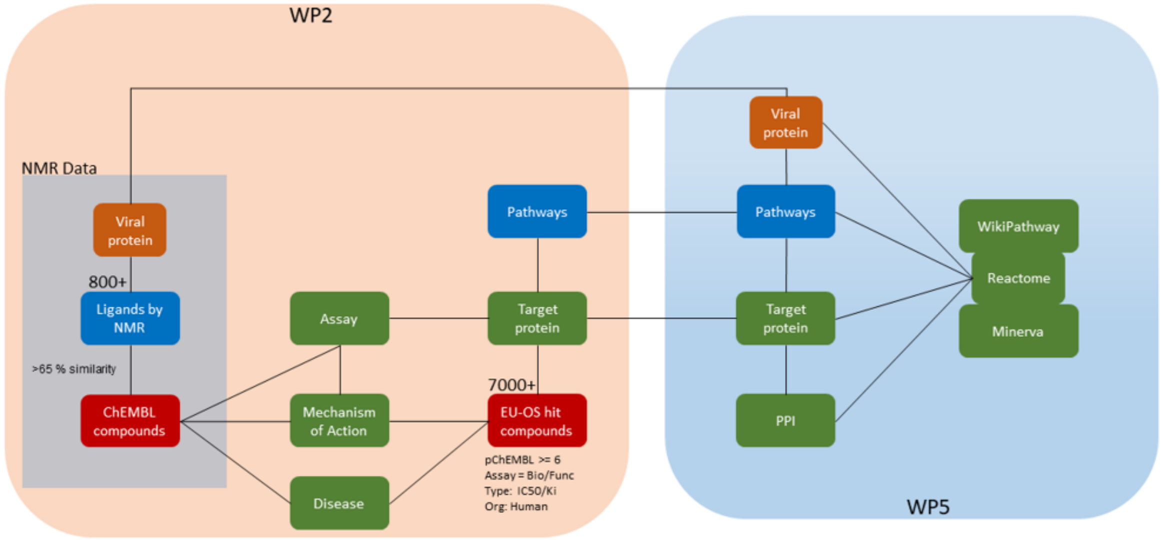


Figure 5: A representation of BY-COVID Knowledge Graph pipeline, including additional sources of information.

The underlying framework of the BY-COVID Knowledge Graph is built in such a way that it can be readily integrated/compared with other Knowledge Graphs and can be exported to several compatible formats such as JSON, CSV, SQL, GraphML and Neo4J as per the need of the users.

5. Discussion

This report summarises the activities undertaken by the existing data and metadata controllers identified in M2.1 (section 4.1) to ensure that their data and metadata is FAIR and align with the proposed standards in D3.1, which aims to facilitate the records indexing in the COVID-19 Data Portal. It also discusses the infrastructure tools that are being developed to enable the rapid onboarding of resources and records using different tier levels (section 4.2).

The activities undertaken to deliver M2.1 have made it apparent that a one schema fits all model cannot work across the diversity of the data resources represented across the BY-COVID project. The diversity of data and metadata across WP2 means that data and metadata harmonisation must vary, not only across the domains but also within subdomains. In some of the domains significant steps have already been taken to harmonise the metadata and underlying data resources. For example, the non-patient data resources part of WP2 due to its non-sensitive nature is already widely available via programmatic mechanisms and dedicated repositories. Thus, it has been possible to generate the BY-COVID Knowledge Graph described in 4.2.4 to integrate and harmonise the different types of data directly. Similarly with the socio-economic data described in 4.1.4, it has been possible to create the harvesting tool described in 4.2.1 to collate metadata from existing public repositories, including records from central catalogues at CESSDA and EUI, into a single resource that can be subsumed by the COVID-19 Data Portal once integration work is completed. As illustrated in 4.2.1, the ongoing work is focused on generating a tier 2 file following the described standard in D3.1.

Within the human/patient biomolecular, and clinical and population health datasets progress has been somewhat more limited, EMBL-EBI has been actively onboarding collaborating institutes to the COVID-19 Data Portal and the metadata of the BBMRI associated datasets from the University of Nottingham (PANTHER and ACE) have been converted to the DCAT derived XML scheme outlined in D3.1 and passed to the portal team for ingress. Other data partners, including the Dutch COVID-19 Data Portal resources are also in the process of onboarding harmonised resources via the Mendix platform.

To further support the onboarding of related resources we have developed an example of best practices on how to transform a human/patient resource to the portal schema. We have created a Zenodo repository for the PANTHER dataset (with the actual data redacted due to GDPR related compliance) to illustrate how a dataset in a non-compliant format can be effectively transformed. In this case from the MIABIS / HDR-UK metadata format to the format required for acceptance into COVID-19 Data Portal. We also included as part of this package a RO-Crate describing the providence of the data and the conversions, not only of the metadata but also the raw data into the OMOP CDM using the CaRROT toolkit described in 4.2.2. The hope is that this resource can provide a framework for onboarding other data partners with resources that do not already conform to the standards adopted across the BY-COVID project.

As mentioned in D3.1, FAIRsharing will complement the COVID-19 Data Portal by acting as a catalogue of data sources, describing their characteristics, including access terms and protocols, and the standards used at the source to represent the data. FAIRsharing acts as the registry of the data resources developed as part of BY-COVID through the dedicated FAIRsharing collection. Although quite complete, this collection is labeled as “in progress” as new resources are expected to be added as well as additional information, especially for the clinical data sources. The collection provides up-to-date metadata for the BY-COVID data resources and how they relate to the broader ecosystem of standards, databases and policies. FAIRsharing and the COVID-19 Data Portal are collaborating on the establishment of cross links between the FAIRsharing records describing data resources and the references to those resources within the COVID-19 Data Portal. This collaboration also links across to the work undertaken in WP4 around the use of RO-Crates that is looking to provide a standardised framework to describe not only the existence of data but also the context, provenance and methods used to create it. This format, which utilises schema.org vocabularies, is an effective way of describing data resources and their associated contextual information (people, organisations, software and equipment).

Whilst the FAIRsharing approach can be widely adopted across many of the resources discussed, its application is limited for the patient related resources from T2.3 that are covered by GDPR. With many clinical resources, for example, it is not possible to include an extensive description of the resource directly. The Infectious Disease Toolkit (IDTk) that is currently being developed by WP4 will be a key system in ensuring that these resources are effectively represented. By defining their particularities in a domain dependent manner, the IDTk will also allow the data controllers affiliated with WP2 to add the missing pieces in the description of resources when necessary aside from showing the information already in FAIRsharing. Thus, the resources will be totally described in one single portal. Furthermore, for the record level information from these resources to become discoverable for reuse in research, a federated mechanism will need to be implemented. The development of the HUTCH connector discussed in 4.2.3 aligns with these aims, as does the primary use case from WP5. However, more work and discussion is required around the details of the federated mechanisms for data access and mobilization.

Initial work has already started in the context of two workshops - one virtual and another one face-to-face during the ELIXIR All Hands meeting in June 2022 - aiming to understand what the challenges are and the potential alternatives as well as how to best align with the needs of the use-cases in terms of required data. An essential aspect is how to increase the automation level regarding data access using programatic means as well as how to elaborate a progressive metadata catalogue to capture existing information at clinical resources. Such an effort would accelerate the process of understanding what data is available and whether it is worthy requesting access or not as this is a time-consuming process.

6. Conclusions

The BY-COVID WP2 will support data producers and metadata sources in enabling their data and metadata to become findable in the COVID-19 Data Portal and interoperable across resources and domains. In the first phase of the project and, thus, in this report, the focus is on discoverability, particularly on metadata and collection level discoverability of data and relevant resources.

This report highlights the areas where focus and collaboration are required. Whilst FAIRsharing will complement the COVID-19 Data Portal by acting as the catalogue of data sources, describing their characteristics, including access terms and protocols, and the standards used at the source to represent the data, its is apparent that several of the data resources present cannot fit within this model. For these resources their inclusion in the BY-COVID ecosystem will be done via the Infectious Diseases Toolkit (IDTk) in WP4 and will depend on how the use cases from WP5 are developed, e.g., which data sources are finalized utilized and what is required to facilitate data access and mobilization, whenever possible.

This will enable the linking of FAIR data and metadata on SARS-CoV-2 and COVID-19, other infectious diseases and related data, and ultimately increase the potential for collaboration and exploitation of data.

7. Next steps

The activities outlined in D2.1 have laid the foundation for the work that needs to be completed in D2.2 on *Data Access and Transfer across research domains and jurisdictions* and D2.3 on *Enabling data discovery at source using beacon-like mechanisms*.

A lot of the work expected for D2.2 is already underway as evidenced by the BY-COVID Knowledge Graph discussed in section 4.2.4. Further work will be aligned by the needs of the use-cases in WP5, especially for the clinical data resources, where federated analysis

Funded by
the European Union

mechanisms developed by WP4 will be needed as data is unlikely to move from their original source.

D2.3 seeks to address the issues of enabling data discovery at source via federated mechanisms. The continued development of open source connectors that can sit within a partner's secure data environment (such as HUTCH - 4.2.3) and tools to effectively transform data into FAIR transport formats (such as CaRROT - 4.2.2). WP4 is developing a client-server federated approach, PHIRI, which is based on an *ad hoc* CDM (common to all the sites) that contains the interoperability layer to allow the deployment of an analytical pipeline on premise, containing the data quality assessment scripts. In the baseline use case in WP5, a Common Data Model using synthetic data is provided to test the ability to link data sources from multiple sites, and mobilise such data to respond to a query on the real life effectiveness of the COVID vaccines. This initial work provides an exemplary model of governance, provenance and data quality. It also aligns with the work around provenance currently being undertaken by WP4, thus close cross collaboration between WP4 and WP2 will be essential to ensure the harmonised development of these systems is effective in real world user-cases.

BY-COVID should also promote the creation of these interoperable transport layers using CDM ontologies such as FHIR and OMOP for new human patient biomolecular and clinical / population data resources, as this aligns resources with the FAIR principles of interoperability and accessibility for effective re-use in research. This links into the work currently being undertaken in WP4 to developing a semantic interoperability model based on RO-Crates to describe the datasets, their provenance, and the processes taken to generate FAIR transfer formats. Alongside the co-development of tools and resources, WP2 and WP4 should also seek to develop a series of documents that outline the best practice for how a data controller can ensure that their resource is aligned to the FAIR standards adopted by the BY-COVID programme. The Infectious Diseases toolkit (IDtk) seems the best resource to capture the initial results of this effort, and serve as the basis for further iterations.

The majority of these activities undertaken by partners in D2.1 have been in response to the outputs of D3.1. When the refined COVID-19 Data Portal indexing system (D3.2) is published, we will take these changes into account and provide a continuously updated version of the BY-COVID WP2 Data Resources into the already agreed mechanisms. At this stage, it should become far clearer which of the data resources that are available can be progressed to the tier 2 discoverability status outlined in D3.1. Furthermore the List of Resources identified in M2.1 is not static. As more resources are added, these will also need to be addressed by the WP2 team to ensure continued growth, harmonisation and discoverability of data.

Funded by
the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

8. Impact

WP2 is focussed on services for the discovery, integration and citation of COVID-19 related data to bring together resources and catalogues across domains to enable fast responses to COVID-19 and future outbreaks. By developing and enhancing existing metadata mappings for localised resources, D2.1 has provided a route by which future resources can be integrated into the COVID-19 Data Portal. It is also developing a range of tools to facilitate these processes that hold their own intrinsic values.

There is a diversity in the use of and level of adoption of existing or developing metadata standards within the various domains and communities represented by the BY-COVID project. The work undertaken across WP2 also has the potential to upskill data sources in how to best catalogue their metadata and make their data FAIR for effective reuse in research. Moreover, it serves as a trigger to seek alignment within and among different scientific domains. Such efforts should allow us to be better prepared for future outbreaks.