BY-COVID - D3.1 - Metadata standards.

Henning Hermjakob, Mari Kleemola, Katja Moilanen, Markus Tuominen, Susanna-Assunta Sansone, Allyson L Lister, Romain David, Maria Panagiotopoulou, Christian Ohmann, Jeroen Belien, et al.

Deliverable D3.1

Metadata standards. Documentation on metadata standards for inclusion of resources in data portal

Project Title (grant agreement No)	Beyond COVID Grant Agreement 101046203		
Project Acronym (EC Call)	BY-COVID		
WP No & Title	WP3: COVID-19 integration platform		
WP Leaders	Henning Hermjakob (EMBL-EBI) Mari Kleemola (CESSDA/TAU-FSD)		
Deliverable Lead Beneficiary	48 - CESSDA ERIC		
Contractual delivery date	31/03/2022	**Actual Delivery date**	29/03/2022
Delayed	No		
Partner(s) contributing to this deliverable	CESSDA/TAU-FSD, UOXF, ECRIN, ELIXIR-NL/VUmc, ELIXIR-NL/Lygature, ELIXIR-UK/UNIMAN		
Authors	Henning Hermjakob (EMBL-EBI) Mari Kleemola (CESSDA/TAU-FSD) Katja Moilanen (CESSDA/TAU-FSD) Susanna-Assunta Sansone (UOXF) Allyson Lister (UOXF) Romain David (ERINHA) Maria Panagiotopoulou (ECRIN) Christian Ohmann (ECRIN) Jeroen Belien (ELIXIR-NL/VUmc) Julia Lischke (ELIXIR-NL/Lygature) Nick Juty (ELIXIR-UK/UNIMAN) Stian Soiland-Reyes (ELIXIR-UK/UNIMAN)		
Contributors	Morris Swertz (UMCG/BBMRI)		
Acknowledgements			

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.

(not grant participants)	
Reviewers	Project Management Board

Log of changes

Date	Mvm	Who	Description
2021-12-21		Mari Kleemola (CESSDA/TAU-FSD)	Initial structure based on workshop 7.12.2021
2022-02-23		Task 3. 1 team	Final structure and division of writing agreed, section 3 edited
2022-02-03		Susanna-Assunta Sansone (UOXF); Allyson Lister (UOXF)	Sections on FAIRsharing added and draft reviewed
2022-03-09		Henning Hermjakob (EMBL-EBI)	Section 4 major editing
2022-03-10		Stian Soiland-Reyes (UNIMAN), Nick Juty (UNIMAN), Mari Kleemola (CESSDA/TAU-FSD), Katja Moilanen (CESSDA/TAU-FSD)	Section 4.3 additions, Section 5 finalised, commenting all text
2022-03-11		Mari Kleemola (CESSDA/TAU-FSD), Henning Hermjakob (EMBL-EBI)	Final draft for peer review
2022-03-28		Mari Kleemola (CESSDA/TAU-FSD)	Comments addressed

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.

Table of contents

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.

1. Executive Summary

BY-COVID Work Package 3 is focussed on services for the discovery and integration of COVID-19 data by delivering a flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources. This will enable the linking of FAIR data and metadata on SARS-CoV-2 and COVID-19, on other related viruses and diseases, and on socio-economic consequences, across research fields, from omics, clinical, and epidemiological research, to social sciences and humanities.

In a series of work package meetings and a workshop, with participation from all other work packages, we have surveyed community metadata standards used by (potential) BY-COVID-19 portal resources (4.1), defined a flexible, three tiered approach to metadata indexing in the COVID-19 portal (section 4.2), derived common metadata attributes for record level discovery (4.3), and established a workflow with FAIRsharing for resource level metadata capture and exchange (4.4).

This work establishes the basis for the further development of the COVID-19 Portal metadata discovery, and provides a path for integration of metadata from multi-domain partners in BY-COVID, as well as our ISIDORe sibling project, and relevant external resources. To ensure smooth integration of partner provided metadata, we will run a technical workshop open to all partners, discussing workflows, metadata attributes and formats, and support tools.

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	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.	No

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.

from outbreak research	3. Research infrastructures on-target training so that users can exploit the platform	No
	4. Engagement so that stakeholders (RI, national centres, policy makers, intergovernmental organisations, funders and end-users) incorporate FAIR and open data in infectious disease guidelines and forward planning.	Yes
Objective 2 Mobilise and expose viral and human infectious disease data from national centres	1.A comprehensive registry of available data with established procedures to collate data governance models, metadata descriptions and access mechanisms in a pandemic scenario.	Yes
	2. Mechanisms for the initial discovery across data sources based on available metadata at the reference collection.	Yes
	3. Demonstrated transnational linking of real-world data from national surveillance, healthcare, registries and social science data that allow the assessment of variants to serve the research needs of epidemiology and public health.	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
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

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.

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

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

3. Scope and methodology

BeYond-COVID (BY-COVID) aims to provide comprehensive open data on SARS-CoV-2 and other infectious diseases across scientific, medical, public health and policy domains. The project will mobilise existing data resources (i.e catalogues) and marshal the resources for research, connect and expose the data and resources via the COVID-19 Data Portal, and drive use and analysis by connecting workflows, national portals and analysis environments. Running for three years, the project brings together 53 partners from 19 countries.

This deliverable is part of BY-COVID Work Package 3 that is focussed on services for the discovery, integration and citation of COVID-19 data by delivering a flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources. This will enable the linking of FAIR[1] data and metadata on SARS-CoV-2 and COVID-19, on other related viruses and diseases, and on socio-economic consequences, across research fields, from omics, clinical, and epidemiological research, to social sciences and humanities. This has the potential to accelerate infectious disease research, surveillance and outbreak investigation.

Inter-domain metadata mapping (Task 3.1) will support data discovery, access, and analysis across fields from molecular biology to social sciences. The harmonised metadata will be discoverable through the central index (Task 3.2) and web portal (Task 3.3), but also be openly accessible for third party applications through web services. The project also addresses analysis transparency, sharing and trusted exchange to support reproducibility (Task 4.3) and attribution and credit for data submitters and workflow developers (Task 3.5). The forthcoming report *D3.2 Tiered indexing system* will describe the implementation of a cloud-based, high performance, scalable indexing system.

WP3 has a close connection to WP2 that mobilises data resources and WP5 that provides the driving use-cases for development. WP4 addresses connectivity to workflows and quality assurance whereas WP6 and WP7 ensure that the portal is aligned with the stakeholder landscape and that people are trained and informed. WP2 will work with WP3 partner UOXF (University of Oxford) to collect all the information about the data resources, as well as the type of (meta)data standards they use and create a dedicated Collection in FAIRsharing. UOXF will also contribute to the policies and rules for participation and inclusion of new data resources, as part of WP6. BY-COVID components from mobilising to analysing and work package touchpoints are presented in Figure 1.

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

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.

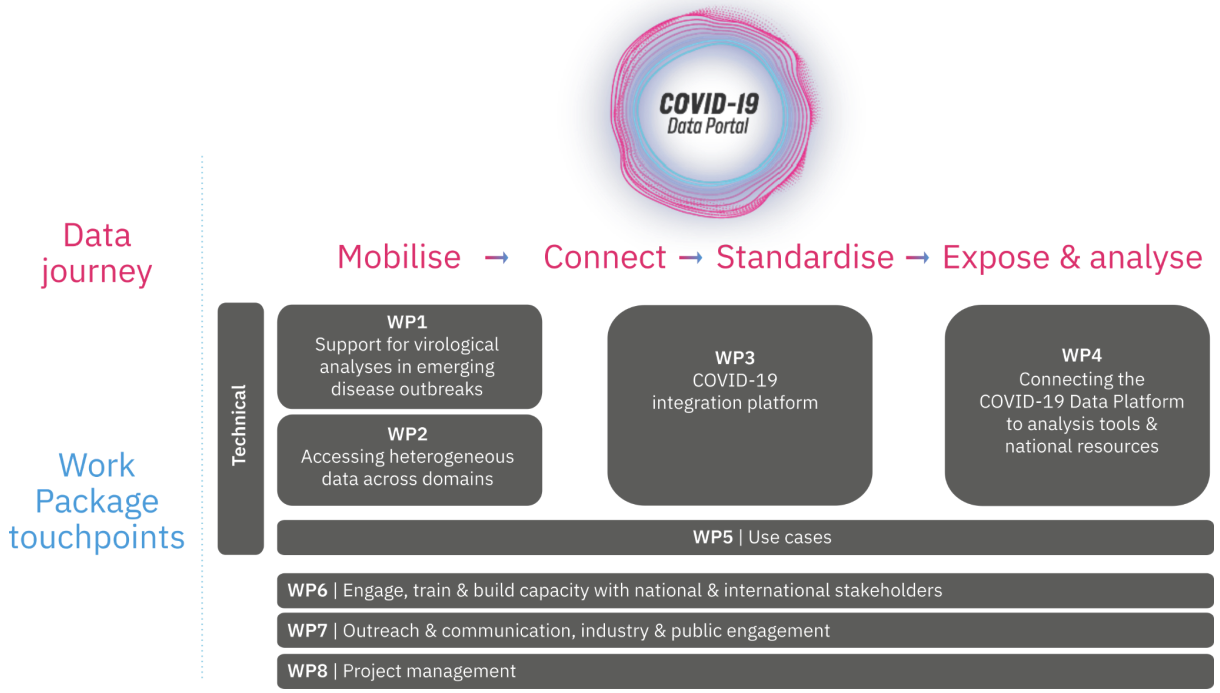

Figure 1. BY-COVID components and iterations

FAIRsharing will also be a key element in the EOSC integration through its collaboration with OpenAIRE, to ensure that the data resources, and the standards they use, are discoverable in the OpenAIRE Research Graph. BY-COVID WP3 members are part of two EOSC Association Task Forces relevant to WP3, FAIR Metrics and Data Quality[2], and Semantic interoperability[3]. This representation will ensure cross-pollination of the work, providing the EOSC Task Forces with real examples and community needs to define the recommendations, and BY-COVID with guidance to follow and implement.

Metadata for the COVID-19 Data Portal will come from a wide range of sources and describe a range of objects from different domains. Chapter 4 describes the relevant metadata landscape and our three-tiered approach and defines the **common metadata elements** for COVID-19 Data Portal that will enable discoverability on record/dataset/study level and describes building blocks for a framework for cross-domain metadata interoperability. The work builds on the experience and skills of the WP3 team, earlier work, desk research, discussions on WP meetings and the results of a BY-COVID metadata workshop held in December 2021.

[2] Sansone and Kleemola are members of the TF FAIR Metrics and Data Quality.
[3] Goble is a member of the TF Semantic Interoperability.

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.

4. Metadata for inclusion of records in the portal

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 will be paid to sensitive data needs and indexing of 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 portal and platforms holding other data types such as social science data, public health data, or epidemiology data in a truly multidisciplinary effort. In the first phase of the project and thus in this report, the focus is on discoverability through the COVID-19 Data Portal.

4.1 Community associated metadata standards used by potential BY-COVID-19 portal resources

The sources from WP2 and WP5 include non-patient related data sources, human/patient bio-molecular data sources, human/patient clinical and health data and socio-economics data sources for infectious disease outbreaks. The different domains and sources recommend and use different metadata standards. Several previous projects and initiatives have charted the metadata standards and practices, and there is active ongoing metadata work. WP3 will build upon previous projects and initiatives and seek synergies with ongoing ones. This chapter describes the metadata landscape briefly.

EOSC Enhance recommendations

As part of the EOSC Enhance project, an extensive landscaping analysis of existing metadata cataloguing efforts, across e-infrastructures and 'clusters' was undertaken by [Goble and Juty 2021]. The report additionally incorporated the outputs of three collaborative workshops (EOSC pilot, EOSC Hub and FAIRsFAIR), and focused on the RIs involved in: ENVRI-FAIR for environmental research; PaNOSC for photon and neutron open science; ESCAPE for astronomy and particle physics; SSHOC for social sciences and humanities and EOSC-Life for life sciences.

The report recognised the diversity of RIs and clusters with respect to the range, data types, infrastructure diversity as well as the maturity of those infrastructures, and the level of adoption of existing or developing standards being used. Overall, it was found that these RIs commonly use generic (DCAT, DC, schema.org) catalogue level, domain-agnostic, metadata standards to provide high level interoperability (tier 3 as defined in section 4.2), which is supplemented by dataset level metadata and domain specific record level information (tier 1, section 4.2).

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.

Importantly, it was recognised that all infrastructures and clusters that were surveyed had significant legacy with respect to catalogues, repositories and existing processes, encompassing data generators, analysts and downloaders, etc. It was recommended that EOSC should build upon those existing mechanisms and systems, and provide lightweight solutions, ensuring low barriers to entry, incorporating some automated mechanisms where appropriate and possible. One means by which integration of Research Infrastructure resources into EOSC could be accomplished is through mapping across generic metadata to a commonly agreed upon 'view' (for instance at the level of DCAT or schema.org), and incorporation of that 'view' into an existing route into EOSC, for instance leveraging those existing processes that are used to populate the OpenAIRE Research Graph. Domain specific metadata could feed into this mechanism through using, for instance, bioschemas, which offers an opinionated view over schema.org, and this can be tailored to suit specific domain data types (see also section 'RDA Metadata Schemas Working Group').

EOSC Guidelines for Research Infrastructures (RIs)

Following on from earlier work (see section 'EOSC Enhance recommendations' above and [Goble and Juty 2021]), the reported recommendations were exemplified through a prototype implementation, and formed the basis on a set of guidelines (see Juty et al. 2021). The major objective was to encourage cross-RI adoption of common metadata interoperability guidelines, through an exemplar implementation. This implementation leverages existing OpenAIRE guidelines for content providers, as well as being congruent with community standards and practices (eg. Schema.org and extensions thereof). In essence, the guidelines cover the following items:

1. Description of a generalised strategy that should be implementable across RIs, leveraging existing, established community practices within those diverse Infrastructures, combined with existing commonly used standard vocabularies (DCAT, etc).
2. Present a prototype implementation, within the Life Science domain, focusing on a number of existing web resources (Protein databases).
3. Each of the (protein) web resources exposed standard metadata for that domain, specifically, being compliant with the Bioschemas 'Protein' profile.
4. Each of the (protein) web resources was registered in a suitable domain registry, in this instance, the EOSC recommended cross-community registry (FAIRsharing).
5. Each of the (protein) web resource records in FAIRsharing provided key information about that resource, including the location of the 'sitemap' and any available API.
6. For each resource, the sitemap was used to navigate the web resource records (individual proteins), using a generic harvesting tool (BMUSE).

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.

7. A mapping file was created, matching properties from Bioschemas Protein profile to properties used by the OpenAIRE research graph (based on the DataCite schema). This conversion was done through software.
8. Ultimately, the OpenAIRE research graph will feed into the EOSC. The OpenAIRE Research Graph aggregates/onboards research product metadata from 2000+ data sources compatible with the OpenAIRE guidelines' profiles, which are based on Dublin Core, DataCite, and JATS metadata schemas

OpenAIRE currently harvests data from a variety of different resources, as defined by the OpenAIRE Service Description Templates (SDT), as well as providing a service to validate level of compliance to specific guidelines. Importantly, OpenAIRE is actively engaging with and onboarding different communities and providers, and developing targeted mappings as required to expand their research graph.

Social sciences and humanities metadata standards

The Social Sciences and Humanities Open Cloud (SSHOC) project charted the metadata standards used in Social Science and Humanities (SSH) domains (see [Broeder et al. 2019]). The SSH metadata landscape is heterogeneous since no one standard can support all use cases. The report recommends two common metadata standards for discovery purposes: Dublin Core, which was the only metadata standard used by all SSH domains; and DataCite, which was recognized as a general enough metadata standard for all SSH domains, but only if persistent identifiers (PIDs), other than DOIs, could also be used. In social sciences, Data Documentation Initiative (DDI) is widely used, but in humanities there is more variation, as shown in Table 1. The CESSDA Metadata Model and the CESSDA Data Catalogue metadata profiles are based on DDI.

Table 1. Recommended metadata standards for social science and humanities [Broeder et al. 2019]

DOMAIN/ERIC	Recommended metadata standards
All	Dublin Core, relaxed DataCite (in addition to DOI, accepts ARK, Handle, PURL, URN and URL as PID)
Social Sciences / CESSDA, ESS, SHARE	DDI Codebook, DDI Lifecycle
Heritage Sciences / E-RIHS	CIDOC-CRM (and its extensions, especially PEM)
Language Sciences / CLARIN	CMDI
Arts and Humanities / DARIAH	CIDOC-CRM (and its extensions), EDM, TEI (teiHeader)

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.

Healthcare and health research data

An EOSC-Life report has concluded that there is currently semantic incompatibility between healthcare and interventional research data standards [Canham et al. 2021]. The ECRIN Metadata Repository allows the discovery of clinical studies and related data objects and is based on the ECRIN Metadata Schema for Clinical Research Objects. In the EHDEN project, two of the most important forms of metadata are information about all the databases available in the network, and the medical studies that are designed, run and published in the network. They propose an adaptation of the widely used schema.org model to advertise this metadata [van Bochove et al. 2020].

Other metadata resources considered here include:

- MIABIS minimum information about biobank[4]
- Real world evidence databanks: Minerva project and specifically metadata schema
- Cohort studies: EUCAN-connect (in collaboration with BBMRI and Maelstrom)
- Genomics Cohorts Knowledge Ontology
- Minimal information for Chemosensitivity assays (MICHA)
- Developments in the creation of the European Health Data Space

Specific considerations for Individual Participant Data (IPD) from clinical trials:

a) It will normally be described at a file / resource level. The resources will usually not be in the COVID-19 portal, but they (for publicly available documents) or their metadata details (for controlled access datasets) will be referenced by the portal.
b) IPD from clinical trials is not likely to be publicly available – access will require application, so application details will need to be incorporated into metadata.
c) It is also not likely to be discoverable directly (no PID or title) but indirectly through people searching on the source study, which will have a title and (usually) a trial registry ID.
d) It may be constrained to in-situ viewing and analysis, rather than download. This fact also needs to be included in metadata where necessary.
e) It is only one component of a collection of digital objects that are necessary to understand the data, e.g. a protocol, analysis plan, a descriptive metadata file.

Many of these considerations also pertain to other individual level data, e.g. social science data.

[4] See: Eklund, N., Andrianarisoa, N. H., van Enckevort, E., Anton, G., Debucquoy, A., Müller, H., Zaharenko, L., Engels, C., Ebert, L., Neumann, M., Geeraert, J., T'Joen, V., Demski, H., Caboux, É., Proynova, R., Parodi, B., Mate, S., van Iperen, E., Merino-Martinez, R., Quinlan, P. R., ... Silander, K. (2020). **Extending the Minimum Information About BIobank Data Sharing Terminology to Describe Samples, Sample Donors, and Events**. *Biopreservation and biobanking*, *18*(3), 155–164. https://doi.org/10.1089/bio.2019.0129

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.

DCAT: Application Profile for data portals in Europe

The DCAT Application Profile for Data Portals in Europe (DCAT-AP) is a specification based on the Data Catalogue Vocabulary (DCAT) developed by W3C. DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web.

FAIRsFAIR: Integration of metadata catalogues

The FAIRsFAIR project addressed the challenges of facilitating cross-disciplinary data discovery and concluded that there is a potential role for using DCAT 2 in supporting findability (F in FAIR) of data at the generic level. However, they note that in order to support other aspects of FAIR, namely accessibility, interoperability and reusability, DCAT must be linked to other standards and tools. (Lambert et al. 2021.)

RDA Metadata Schemas Working Group

Within the Life Sciences, particularly for resources participating in ELIXIR, the 'Bioschemas' approach to improve findability of those resources is being adopted. Bioschemas is an opinionated view which extends schema.org towards a specified domain and associated data types. As such, since it extends the web-embedded metadata of its parent schema.org, leveraging the power and utility of web-based search. Additionally, Bioschemas also provides usage guidelines for metadata properties through the provision of 'profiles' over schema.org types, guiding the user to embed properties correctly. Working in collaboration with Google, bioschemas now has 6 defined types adopted into schema.org directly (eg Gene, Protein).

Bioschemas is not alone in adopting this strategy to make domain-specific information more discoverable; parallel efforts have evolved in other domains, and which are captured under the umbrella activity of 'Research Metadata Schemas', under the auspices of the Research Data Alliance (RDA). This group encompasses participants from a variety of domains, assisting in developing guidelines in a manner analogous to that of bioschemas. For instance science-on-schema (for ESIP[5]), as well as providing guidelines. Another important output from this group was a crosswalk from schema.org properties to corresponding metadata terms in other metadata models ([Wu et al. 2021]). Notable of this crosswalk is the visualisation as a filterable table and Sankey Diagrams (see Figure 2).

[5] Jones, M.B., Richard, S., Vieglais, S., Shepherd,S., Duerr, R., Fils, D., and McGibbney, L.. (2021). **Science-onSchema.org** v1.2.0 (Version 1.2.0). Zenodo. https://doi.org/10.5281/zenodo.4477164

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.

Filter Table Data

author

Standard	Term	Schema.org crosswalk	Parent Schema
ISO-19115-1	identification/citation//identifier/code (use authority or codespace in Identifier to indicate context for this alternate identifier))	sameAs	schema:Thing
ISO-19115-1	identification/citation /citedResponsibleParty/name [role = one of {author, coAuthor, originator, editor}]	creator	Schema:CreativeWork
Dataverse	Author; authorName(M)	creator	schema:CreativeWork
RIF-CS	collection/citationInfo/citationMetadata /contributor OR	creator	schema:CreativeWork

Figure 2. Filtering RDA's schema.org crosswalk on term "author"

EOSC Interoperability Framework metadata schema guidelines and crosswalk

The *EOSC Interoperability Framework* report (Corcho et al. 2021) was produced by the Interoperability Task Force of the EOSC Executive Board, in collaboration with the EOSC Architecture Working Group and related EOSC projects (incl. EOSC Enhance, EOSC-Hub, OpenAIRE Advance and EOSC Future). The report highlights the need for a **minimal metadata** architectural building block. As part of this work, the working group therefore analysed existing metadata models and generated an initial *Crosswalk of most used metadata schemas* [Ojsteršek et al. 2021], as a starting point to build an *EOSC Minimal Metadata Application* profile. This crosswalk maps between what was found as the most commonly used metadata schemes:

- RDA Metadata Interest Group recommendation of the metadata element set
- EOSC Pilot - EDMI metadata set
- Dublin CORE Metadata Terms
- Datacite 4.3 metadata schema
- DCAT 2.0 metadata schema and DCAT 2.0 application profile
- EUDAT B2Find metadata recommendation
- OpenAIRE Guidelines for Data Archives
- OpenAire Guidelines for literature repositories 4.0
- OpenAIRE Guidelines for Other Research Products

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.

- OpenAIRE Guidelines for Software Repository Managers
- OpenAIRE Guidelines for CRIS Managers
- Crossref 4.4.2 metadata XML schema
- Harvard Dataverse metadata schema
- DDI Codebook 2.5 metadata XML schema
- Europeana EDM metadata schema
- Schema.org
- Bioschemas
- W3C PROV Ontology

Common vocabularies examined in this crosswalk include Datacite, Crossref, OpenAIRE, COAR, MARC and Dublin Core. Recommendations from the EOSC IF report for minimum metadata to describe records of FAIR Digital Object include:

- **Mandatory**: Identifier, Creator (0-n), Title, Publisher, Publication Year, Resource Type, Rights/terms of Access (license URL, confidentiality, disclaimers etc), File (file name, download URL, format, size, checksum, version, etc.), Project (where applicable)
- **Recommended**: Subject, Description, Contributor, Date, Language, Alternative Identifier, Related Identifier, Version, Coverage (temporal, spatial), Source
- **Additional recommended properties** (where applicable): Contact, Producer, Production Date, Production Place, Dataset Distribution, Depositor, Deposit Date, Date of Collection, Kind of Data, Series, Software, Data Sources, Origin of Sources, Documentation, Characteristics, Time Method, Frequency, Universe, Unit of Analysis, Standard

Recently, EOSC Future is developing an operational EOSC platform, as a federated *system of systems* approach to connect existing infrastructure. As part of this work to form a "*Minimum Viable EOSC*", EOSC Future has published a draft update of *EOSC architecture and interoperability framework*[6] which is currently open for consultation.

The work on crosswalk for EOSC IF is now largely carried on within the EOSC Task Force for Semantic Interoperability and Task Force for Technical Interoperability of Data and and Data.[7]

[6] L. Florio, M. Van de Sanden, D. Scardaci, M. Williams, O. Appleton, P. Manghi, and K. Jeffery (2021): **EOSC Architecture and Interoperability Framework**, 1.0, European Open Science Cloud, *EOSC Portal*. https://eosc-portal.eu/sites/default/files/EOSC%20Future-WP3-EOSC%20Architecture%20and%20Interoperability%20Framework-2021-12-22%5B17%5D%5B6%5D-2.pdf [10.3.2022]

[7] The EOSC task forces have a solid BY-COVID representation by UNIMAN, BSC, VIB, EMBL-EBI, UOXF, ERINHA, BBMRI-ERIC, CSIC, INSTRUCT, TUNI, UU, DANS, CESSDA and others.

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.

EOSC Future TSP COVID-19 metadata findability and interoperability in EOSC

The EOSC Future Test Science Project (TSP) entitled "COVID-19 metadata findability and interoperability in EOSC (META-COVID)" brings together partners from the life sciences domain (EOSC-Life cluster: ECRIN, BBMRI, EATRIS, EU-Openscreen, EMBL-EBI) and from social sciences and humanities (SSHOC cluster: CESSDA-TAU-FSD, CLARIN-ERIC) with the objective to develop a framework for a metadata model, characterising the research approach and workflow across research infrastructures and to apply it to a use case in COVID-19 research. The framework will be applied to enable cross-domain interoperability and common metadata formats, iteratively improving crosswalk and mappings to support indexing and discoverability for interactive and API use. The first deliverable of the TSP is an inventory of metadata schemas applied across infrastructures and domains and a survey has been conducted to list this information for the participating RIs. The prominent role of ECRIN and CESSDA in both EOSC Future and BY-COVID will facilitate knowledge transfer.

ISIDORE

ISIDORe "Integrated Services for Infectious Diseases Outbreak Research" is a 36-month long European project led by ERINHA and a brother project of BY-COVID. It assembles an unprecedented One Health-driven, integrated portfolio of cutting-edge research services and resources, dedicated to the study of epidemic-prone diseases including SARS-CoV-2 variants. All services will produce data and the challenge is to reuse BY-COVID outputs and recommendations in ISIDORe and to facilitate feeding of existing platforms. In particular, ISIDORe WP3 has the objective of FAIR data management and enhancement of capacities & tools for data curation, access & mobilisation. Another task in WP4 of ISIDORe will focus on data quality and data services quality provided to feed BY-COVID related platforms.

4.2 Three-tiered approach

Exposing and effective connection and linking of different data types requires indexing based on cross-mapped metadata. Indexing and incorporation of metadata into the COVID-19 Data Portal proceeds via a flexible, tiered system for metadata integration (Table 2).

- For tier 1, a limited number of key resources 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

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.

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.

We anticipate that some external resources over the course of the project will migrate from tier 3 upwards, as resources for the detailed metadata harmonisation and technical indexing become available.

To minimise repeated requests to collaborators and "Federation Fatigue", we are co-ordinating requests for metadata with WP2, and centralise requests for resource level (tier 3) metadata in FAIRsharing. 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. Resource metadata will be captured in FAIRsharing and selectively imported to the COVID-19 Data Portal through the API (see section 4.4).

Table 2: Three Tiered Indexing Concept

Tier 1	Deepest indexing available, capturing granular, record level identifiers and metadata, support for interoperability use cases	Aim: Key resources migrate from Tier 3 to Tier 1 over project duration. ↑
Tier 2	Coarse-grained metadata and attributes, focus on record level discoverability	
Tier 3	Focus on resource level discoverability.	

4.3 Common metadata attributes for record level discovery

Record level (tier 2) discoverability is built on the current COVID data portal requirements and the OmicsDI Metadata Format Specification[8]. The portal is based on the EMBL-EBI search system [Park et al 2017], which is highly scalable and provides discoverability for billions of records from more than 100 resources. Each resource will provide their

[8] OmicsDI Data Format Specification: http://blog.omicsdi.org/post/omicsdi-spec/ [18.2.2022]

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.

metadata in an agreed format. In common with most metadata standards, the requested fields are grouped into three categories:

- **Mandatory**: id, name, description, dates (creation, modification, release).
  These elements ensure basic discoverability and support implementation of consistent search strategies across all indexed data objects.
- **Recommended**: These fields are agreed by project partners and serve to provide mainly domain-specific metadata, which is important for discoverability, but different between disciplines, for example DNA sequence metadata versus population study metadata.
- **Additional**: These provide a flexible option to provide additional metadata that may improve discoverability. These fields are listed in the shared attribute table to support harmonisation, but may also be added on an ad-hoc basis by an individual resource. It is expected that the weight of these fields for results ranking will be lower than for mandatory and recommended fields.

Tier 2 indexing will be based on a minimum common metadata set for including sources in the portal. The focus is on attributes that researchers would need for discoverability rather than interoperability. Analysis of existing metadata crosswalks, previous reports (see chapter 4.3) and current portal requirements revealed that the mandatory attributes at this level are similar regardless of discipline or domain. We have developed our common metadata elements to ensure compatibility with the DataCatalog Vocabulary (DCAT) as a key EU recommended metadata standard, and schema.org/BioSchemas as key metadata standards for domain-independent and life science discoverability, respectively.

This common set of elements can be expressed in various metadata standards/formats. To support rapid implementation and ease of metadata validation, we will initially request metadata to be provided by participating resources in XML, and provide a suitable validator. We anticipate that the initial common metadata set will evolve over time, reflecting increasing breadth of included resources, and depth of metadata provision. However, we will aim to minimise changes in mandatory fields, and restrict evolution of the metadata format to the recommended and additional fields, to avoid the need for frequent changes to the format requirements mandatory for all partners. Appendix 1 provides the current (static) status of the common metadata set as of March 10, 2022, while a living spreadsheet[9] provides a continuously updated view.

[9] BY-COVID common metadata elements: https://docs.google.com/spreadsheets/d/1YXGhGHm5ErqmLEOtjOAHehLKPRYSOfBcGWFeoAsx58Y/

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.

4.4 FAIRsharing: collections, links to COVID-19 Data Portal and connections to EOSC

The portal will also include information on additional relevant resources (tier 3), where FAIRsharing[10] will contribute in several ways, as summarised in Figure 3. FAIRsharing assists with the selection, documentation and visualisation of the relevant standards and crosswalks registered and interlinked in FAIRsharing.org by providing a live BY-COVID reference standards FAIRsharing collection (draft). FAIRsharing also acts as the registry of those data resources developed as part of BY-COVID through the dedicated BY-COVID data resources FAIRsharing collection (draft). This collection provides up-to-date metadata for BY-COVID data resources (in WP2) 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; the connection between the two resources will be enabled by a semi-automated solution that uses the FAIRsharing API. Finally, FAIRsharing will be a discovery route in EOSC via its collaboration with OpenAIRE; work is in progress to map the metadata fields from both resources prior to incorporation of FAIRsharing content into the OpenAIRE knowledge graph.

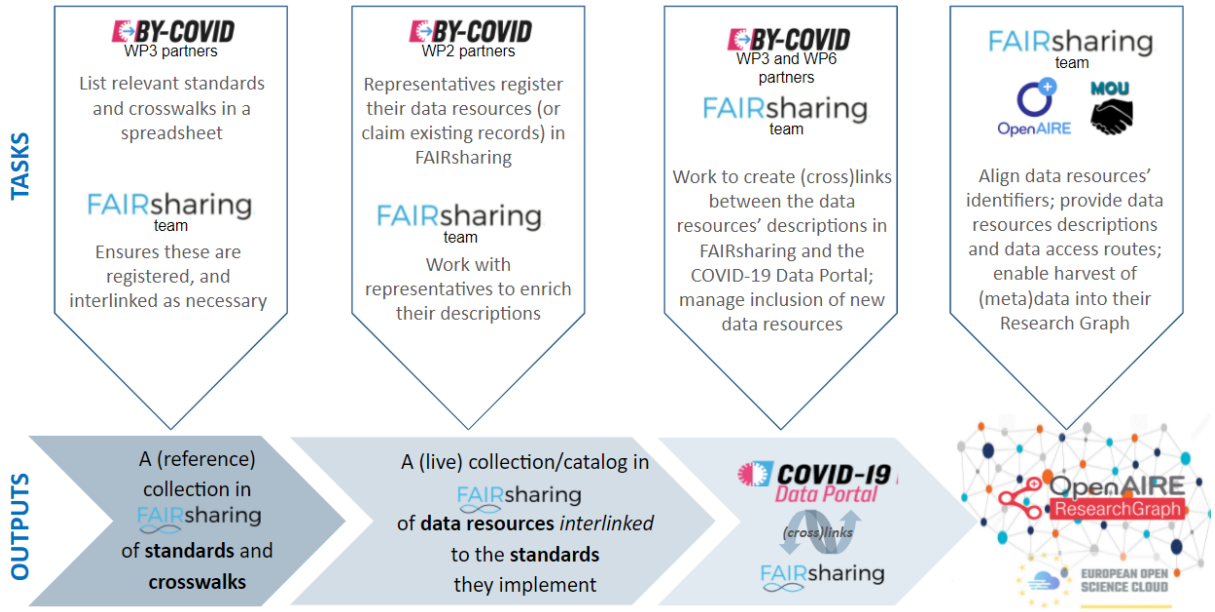

Figure 3. FAIRsharing connectivity workflow

[10] FAIRsharing (https://fairsharing.org)

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.

5. Discussion

This report summarises existing metadata schema mappings and recommendations (section 4.3) that we have considered for BY-COVID for the purpose of registering and indexing records in the COVID Data Portal. As explained in section 3 on methodology, the existing work has informed our approach to identify common metadata attributes (section 4.2). The curated collections of metadata standards will be made available through FAIRsharing (section 4.4).

Common metadata standards are essential in order to ensure FAIR aspects of the portal for programmatic use, but are also important at the point of registration to improve usability for human discovery, e.g. faceted browsing.

Crosswalks are suitable for finding commonalities, but also mapping across existing metadata schemas; however for this to be useful at a programmatic level (e.g. generating Bioschemas markup from DCAT metadata), the current practice of recording crosswalks as spreadsheets (e.g. EOSC IF's crosswalk) needs improvement towards structured formats.

While BY-COVID needs to curate the existing metadata schema mappings to a subset that applies to the COVID Data Portal and that can be consumed programmatically, this raises a question of sustainability, who will maintain these mappings and FAIRSharing records beyond the BeYond-COVID project? For this reason we see it as important to continue this work in open collaboration with existing initiatives within EOSC task forces and RDA working groups.

Provenance-wise it is also important to note that any mapping between standards and schemas are at risk of losing information, e.g. because of a difference in granularity of terms and structure. A possible mitigation strategy is to also keep the original metadata alongside the translated/mapped metadata, with corresponding provenance trace, e.g. as supported by RO-Crate packaging (Soiland-Reyes et al. 2022). However, for the discoverability focus of the work presented here, a limited conversion loss of information is acceptable.

Discoverability metadata in the COVID Data Portal will be openly and freely available for any user according to FAIR principles, and without any access control. Therefore it is essential that we ensure legal agreements are in place that allow the sharing of metadata that have been submitted to the portal. It is important to distinguish between usage rights of records in a database (which may be restricted, confidential etc.) and the licence of the metadata of the corresponding entry in the Data Portal, which for the dedicated purpose of discoverability naturally must be open.

Given the above, it may still be challenging to the submitter to ascertain the intellectual property rights that govern that metadata, so we may not be able to require the submitter

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.

to re-license their submitted metadata as a public domain licence like CC0; yet such an approach moves the problem to the programmatic users of the Data Portal if they are not legally able to integrate or republish its metadata according to the FAIR principles.

6. Conclusions

The BY-COVID metadata framework will support data producers and metadata sources in choosing the right standards and approaches for their metadata to become findable in the COVID-19 Data Portal and interoperable across resources or domains. Indexing and incorporation of metadata into the COVID-19 Data Portal proceeds via a flexible, tiered system for metadata integration. In the first phase of the project and thus in this report, the focus is on discoverability, particularly on metadata record level discoverability of data (tier 2) and discoverability of relevant resources (tier 3).

Record level (tier 2) discoverability is built on the current COVID data portal requirements and the OmicsDI Data Format Specification and is based on a minimum common denominator for including sources in the portal. The focus is on attributes that researchers would need for searching and discovering data in the first place and the core elements or attributes in this level are similar regardless of discipline or domain. The common set of metadata elements are presented in Appendix 1. It includes crosswalks to BioSchemas and DataCatalog Vocabulary (DCAT), but the common set can be expressed in various metadata standards/formats.

The portal will include information on additional relevant resources (tier 3). 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.

This report also summarises existing metadata schema mappings and recommendations that we have considered for BY-COVID for the purpose of registering and indexing records in the COVID Data Portal. The existing work has informed our approach to identify common metadata attributes. BY-COVID members are part of various communities, projects, initiatives and task forces relevant to WP3 which will ensure cross-pollination of the work.

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.

7. Next steps

The common metadata attributes will form the basis for metadata provided by project partners. We will refine the existing metadata indexing system (D3.2) of the COVID-19 Portal to take these updates into account, and to provide a continuously updated version of the COVID-19 Data Portal (D3.3), as well as strategies for broader pandemic preparedness.

To ensure smooth integration of partner provided metadata, we will run a technical workshop open to all partners, discussing workflows, metadata attributes and formats, and presenting support tools, in particular a custom XML validation tool. All training materials will also be linked from the BY-COVID internal website.

In order to test the integration of resources for discoverability into the portal, we will initially focus on limited test cases for each of the three tiers. While for tier 3 the addition of resource level metadata for a new resource will be straightforward, we have identified CESSDA Data Catalogue as a realistic use case for tier 2 integration, with subsequent migration to level 1. The main tier 1 use cases will be based on requirements from WP2 and WP5.

The work will be continued in open collaboration with existing projects and initiatives within EOSC task forces and RDA working groups.

9. Impact

WP3 is focussed on services for the discovery, integration and citation of COVID-19 data by delivering a flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources. 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. To achieve this, there needs to be cross-discipline uptake on metadata for the tiered indexing system. WP3 work has the potential to upskill data sources in metadata producing and increase common understanding of metadata needs and importance. While there is a diversity in the use of and level of adoption of existing or developing metadata standards within the communities, WP3 will strive to find synergies and common features. Developing and enhancing existing metadata mappings based on use cases will also be a means to integrate resources into EOSC cloud. This report provides the basic building blocks and landscape analysis that further WP3 work will build upon.

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.

10. References

[van Bochove et al. 2020] Kees van Bochove, Emma Vos, Maxim Moinat, Sebastiaan van Sandijk, Tess Korthout, & Peyman Mohtashani. (2020). **EHDEN - D4.5 - Roadmap for interoperability solutions** (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4474373

[Broeder et al. 2019] Broeder, Daan; Trippel, Thorsten; Degl'Innocenti, Emiliano; Giacomi, Roberta; Sanesi, Maurizio; Kleemola, Mari; Moilanen, Katja; Ala-Lahti, Henri; Jordan, Caspar; Alfredsson, Iris; L'Hours, Hervé; Ďurčo, Matej (2019): **D3.1 Report on SSHOC (meta)data interoperability problems** (Version 1.0). SSHOC deliverable. *Zenodo*. https://doi.org/10.5281/zenodo.3569868

[Canham et al. 2021] Canham, Steve, Ohmann, Christian, Boiten, Jan-Willem, Panagiotopoulou, Maria, Hughes, Nigel, David, Romain, Sanchez Pla, Alex, Maxwell, Lauren, Aerts, Jozef, Facile, Rhonda, Griffon, Nicolas, Saunders, Gary, van Bochove, Kees, & Ewbank, Jonathan. (2021). **EOSC-Life Report on data standards for observational and interventional studies, and interoperability between healthcare and research data.** Zenodo. https://doi.org/10.5281/zenodo.5810612

[Corcho et al. 2021] O. Corcho, M. Eriksson, K. Kurowski, M. Ojsteršek, C. Choirat, M. van de Sanden, F. Coppens (2021): **EOSC Interoperability Framework**, Publications Office of the EU, 2021. https://doi.org/10.2777/620649

[Goble and Juty 2021] Carole Goble, & Nick Juty. (2021). **Analysis of existing research data cataloguing efforts towards integrated discovery.** EOSC Enhance deliverable. *Zenodo.* https://doi.org/10.5281/zenodo.4693217

[Juty et al. 2021] Nick Juty, Carole Goble, Alasdair Gray, Paolo Manghi, Allyson Lister, Susanna-Assunta Sansone, Andreas Czerniak (2021): **D4.4 Guidelines for RIs to enable discoverability of research data.** EOSC Enhance deliverable. EOSC Portal. https://eosc-portal.eu/sites/default/files/Final-D4.4%20Guidelines%20for%20RIs%20to%20enable%20discoverability%20of%20research%20data-pdf.pdf [10.3.2022]

[Lambert et al. 2021] Simon Lambert; Ricarda Braukmann; Eva Méndez, Eva; Marina Sánchez; Joy Davidson (2021): **D3.7 Report on integration of metadata catalogues** (Version 1.0 DRAFT). FAIRsFAIR deliverable. *Zenodo.* https://doi.org/10.5281/zenodo.5744913

[Ojsteršek et al. 2021] M. Ojsteršek, O. Corcho, M. Eriksson, K. Kurowski, M. van de Sanden, and C. Frederik (2021): **Crosswalk of most used metadata schemes and guidelines for metadata interoperability**, Zenodo. https://doi.org/10.5281/zenodo.4420116

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.

[Park et al 2017] YM Park, S. Squizzato, N. Buso, T. Gur, R. Lopez (2017). **The EBI search engine: EBI search as a service-making biological data accessible for all.** Nucleic Acids Res. https://doi.org/10.1093/nar/gkx359

[Soiland-Reyes at al. 2022] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): **Packaging research artefacts with RO-Crate**. Data Science https://doi.org/10.3233/DS-210053

[Wu et al. 2021] M. Wu, P. Hagan, B. Cecconi, S.M. Richard, C. Verhey, RDA Research Metadata Schemas WG (2021): **A Collection of Crosswalks from Fifteen Research Data Schemas to Schema.org.** Research Data Alliance. https://doi.org/10.15497/RDA00069

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.

Appendix 1. The BY-COVID common metadata set as of March 10, 2022

Note: a living spreadsheet provides a continuously updated view of the metadata set.

M/A/R = (M)andatory or (R)ecommended or (A)dditional
M* = at least one of the date fields must be present

Field	Comment	Example	M/A/R	BioSchemas crosswalk	Data Catalog Vocabulary (DCAT) crosswalk (pdf)
Database Section of XML				Attributes in this section are the crosswalk to BioSchemas.DataCatalogue	Overview of DCAT model
name	Name of the database or provider	<name>PRIDE</name>	M	DataCatalogue:name	dcterms:title
description	A short description of the provider	<description>The proteomics identification database is an EBI resource for Proteomics</description>	R	DataCatalog:description	dct:description
release	The tag for the database release to which the data belongs.	<release>Release-May-2016</release>	A	[]	
release_date	The date of the database release to which the data belongs.	<release_date>2015-05-13</release_date>	R	DataCatalog:dateCreated	dct:issued
entry_count	The number of entries in the XML file. This field is used for validation purposes.	<entry_count>2</entry_count>	R	[]	dcat:dataset
keywords	Keywords or tags used to describe this content. Multiple entries in a keywords list are	<keywords> C2c12 myotubes,Filamin c,Filip1</keywords>	M	DataCatalog:keywords	dcat:keyword

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.

	typically delimited by commas.				
url	Home page of the resource.	<url>https://www.ebi.ac.uk/pride</url>	M	DataCatalog:url	foaf:homepage
search_url	URL which can be used as a prefix to show a specific record, defined by an identifier.	<search_url>https://www.ebi.ac.uk/pride/archive/assays/</search_url>	A	[]	
Each Entry Section of XML				Attributes in this section are the crosswalk to BioSchemas.DataSet	
id	Original and UNIQUE identifier across the repository, database or provider	<entry id="PXD000001"></entry>	M	Dataset:identifier	dct:identifier
name	Name, title of the dataset, can be considered as the title of the publication	<name>TMT spikes</name>	M	Dataset:name	dct:title
description	A short description or abstract of the dataset. It can be considered similar to a "publication abstract"	<description>Expected reporter ion ratios: Erwinia peptides</description>	M	Dataset:description	dct:description
date	Date of publication of the dataset	<date type="**publication**" value="2014-09-22">	M*	[]	dct:issued
date	Date of initial creation of dataset submission in the database	<date type="**creation**" value="2014-09-22">	M*	[]	
date	Date of successful submission to the database	<date type="**submission**" value="2014-09-22">	M*	[]	
date	Date of the latest update to the dataset	<date type="**updated**" value="2014-09-22">	M*	[]	
Each Entry Section of XML - Additional Fields				Attributes in this section are the crosswalk to BioSchemas.DataSet	
data_protocol	Description of the software, pipeline and tools used to process the data	<field name="data_protocol"></field>	R	[]	
sample_protocol	Description of the sampling	<field name="sample_protocol"></field>	R	[]	

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.

repository	The name of the repository or provider	<field name="repository">PRIDE</field>	M	Used in Dataset.includedInDataCatalog	dct:publisher
species	Organism(s) studied in the experiment that generated the data (Free Text)	<field name="species">Homo sapiens</field>	A	[]	
disease	Disease(s) studied in the experiment that generated the data (Free Text)	<field name="disease">Lung carcinoma</field>	A	[]	
tissue	Tissue(s) studied in the experiment that generated the data (Free Text)	<field name="tissue">Lung</field>	A	[]	
cell_type	Cell type(s) studied in the experiment that generated the data (Free Text)	<field name="cell_type">brain cortex glial cell</field>	A	[]	
full_dataset_link	The original link of the dataset in the provider's web service, it should be a universal URL that can be used to find the original data	<field name="full_dataset_link"> http://www.ebi.ac.uk/pride/archive/projects/PRD000123 </field>	M	Dataset.url	dcat:landingPage
submitter	Name of the person who submitted the data into the original repository	<field name="submitter">Yasset Perez-Riverol</field>	A	Used in Dataset.creator	dcterms:creator
submitter_mail	Submitter's contact email	<field name="submitter_mail">yperez@ebi.ac.uk</field>	A	Ditto	dcat:contactPoint
submitter_affiliation	Submitter's affiliation, institution, department, etc.	<field name="submitter_affiliation">European Bioinformatics Institute</field>	A	Ditto	
instrument_platform	Instrument used to analyze the experiment's samples	<field name="instrument_platform">LTQ Orbitrap</field>	R	[]	
technology_type	Technique of instrumental analysis used in the experiment	<field name="technology_type">Tandem MS/MS</field>	A	Dataset.measurementTechnique	

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.

modification	Post-translational modifications; used mainly in Proteomics experiments	<field name="modification">Oxidation</field>	A	[]	
submitter_keywords	Keywords describing the dataset further, in this case added by the submitter of the data	<field name="submitter_keywords">ProteoGenomics</field>	A	Dataset.keywords	dcat:keyword
quantification_method	Free text describing the quantitative method used in the data analysis	<field name="quantification_method">SILAC</field>	A	[]	
submission_type	In ProteomeXChange this field is used to classify the type of submission	<field name="submission_type">COMPLETE</field>	A	[]	
software	Software(s) used for data analysis	<field name="software">Trans-Proteomics Pipeline</field>	A	[]	
publication	Free text describing the publications, citation, title	Some resources do not provide information about the PubMed article in which the dataset was published. In such cases OmicsDI provides a mechanism to add the publication as free text by means of the field <field name="publication"/>.	A	[]	
dataset_file	This a URL of an individual data file in the dataset	<field name="dataset_file"> ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2010/07/PRD000123/PRIDE_Exp_Complete_Ac_9777.xml.gz </field>	A	Used in Dataset.distribution	
Each Entry Section of XML - Cross-references					
ref	Cross-references of that entry to other repositories	<ref dbkey="19770167" dbname="pubmed">	R	Used in Dataset.citation	
Please provide suggestions for new fields below					

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.

funder	3 metadata fields about funder are needed for OpenAIRE: the name of the funder, identifier, and the type of the identifier see e.g. https://guidelines.openaire.eu/en/latest/data/use_of_datacite.html#funding-information and https://guidelines.openaire.eu/en/latest/data/field_contributor.html		A	[]	
PID of the publication that is related to the dataset	OpenAIRE requires the PID of the publication	<publication PID="10.1234/foobar" PIDtype="DOI"> Title of the publication/citation of the publication </publication>	A	Used in Dataset.citation	
Type of the publication's PID	OpenAIRE requires the type of publication's PID		A	[]	
Unit of analysis/observation (for social sciences) or just a general field not specifically for social science	Use of the DDI Analysis Unit vocabulary is recommended for social sciences	<analysis_unit vocabulary="DDI Analysis Unit">Individual</analysis_unit> <analysis_unit vocabulary="DDI Analysis Unit">Family</analysis_unit> <analysis_unit>Homo sapiens</analysis_unit>	A		
dataset distribution	This property links the Dataset to an available Distribution.	dcat:distribution	A		
spatial/ geographical coverage	This property refers to a geographic region that is covered by the Dataset.	dct:spatial DDI (social sciences): <geogCover>Finland</geogCover>	A		
temporal coverage	This property refers to a temporal period that the Dataset covers.	dct:temporal DDI (social sciences): <timePrd date="2017-01-10" event="start"/><timePrd date="2018-05-05" event="end"/>	A		
spatial resolution	This property refers to the minimum spatial separation	dcat:spatialResolutionInMeters	A		

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.

	resolvable in a dataset, measured in meters.				
temporal resolution	This property refers to the minimum time period resolvable in the dataset.	dcat:temporalResolution	A		
substance	I suggest to insert a new row with field substance of which tissue is one of the potential values (other values are for example blood, saliva). an additional question could then be based on provided answer on substance like tissue and then provide the anatomical origin of the tissue?		A		
Research activity id (RAID)	Most datasets and other digital objects in clinical research are linked to and found by searching on the source reserch (trial, study). Many studies, including almost all clinical trials, are registered as a research activity.	Clinicaltrials.gov ID (e.g. NCT05189457), Eudra CT number (e.g. 2021-002445-15).	A		
Data Access type	(Often but not always linked to managed access) the type of access possible, especially when it is something other then free download of material. More than one access type may apply.	Public download, public API access, public on screen access only, download after self-attestation, download restricted to authenticated users, download after approval by DAC, in situ access for analysis after approval by DAC, etc.	A		
Managed Access flag	Set if a resource is available but not under open access, e.g. it may require prior membership of a defined group, pre-requisites for access such as ethical aproval of the planned	Boolean flag	A		

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.

	secondary use, or consideration of the request by a data access committee.				
Managed Access Details	if managed access…URL of a page providing details on the access procedure and requirements	https://yoda.yale.edu/how-request-data	A		
Sensitive Data Type	if managed access…the type of sensitive data involved, or equivalently the reason for the managed access	Pseudonymised individual health data, human image data, human genomic data, data falling under Nagoya protocol, commercially sensitive data, etc. (see: category 1 in the categorisation system for sensitive data, version 3, https://zenodo.org/record/5507324#.YgT1Ld_MI2w)	A		
Associated consent	if managed access for human data…the consent provided by the data donors for use, including any specific restrictions on use	Based on DUO classification: No consent, No restriction, General research use, Health / biomedical research, Disease specific research, Non-commercial research only, With geographical restrictions, Specific research types (e.g. genetics only), No algorithm development use	A		
Branch of science / discipline	Describes the dataset's branch of science. Useful for the end-users, when the catalogue is multidisciplinary. A list of allowed values is needed.	DataCite example: <subjects> <subject xml:lang="en-US" schemeURI="http://dewey.info/" subjectScheme="dewey" classificationCode="000">computer science</subject> </subjects>	A		
Type of the data	Type of the data classifies the dataset and makes it easier for the end-users to find the type of data they are interested in. Some kind of list of allowed values is quite likely needed; in		A		

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.

	the present portal data types seem to be e.g. “viral sequences”, “host sequences”, “expression”, “proteins”.				
universe	Defines the population (sampling protocol/procedure then explains how the sample has been formulated). High-level information of how the objects to be researched are selected, but in lower abstraction level than analysis unit.	example 1: <universe clusion="I">Finnish population aged 18 or over</universe> <universe clusion="E">Åland Islands</universe> example 2: <universe clusion="I">Finnish women who were born before the year 1924 and who had experienced the Winter War and the Continuation War</universe> example 3: <universe clusion="I">Managers of municipal health and social service offices, select staff at the Social Insurance Institution of Finland (Kela), and social workers in social assistance and adult social work</universe> <universe clusion="E">the Åland Islands</universe> <universe>brain cortex glial cell</universe>	A		
Date(s) of the data collection	Date or range of dates when data was collected.	<collDate date="1987-01-01" event="start"></collDate> <collDate date="1988-12-31" event="end"></collDate>	A		
Licence	This property refers to the licence under which the Catalogue/Dataset/Dataservice/ Datadistribution can be used or reused.		A		dct:license

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.

URL to original dataset description	URL to a specific record/entry in the original provider's catalogue/web page (in the case of FSD, CESSDA will probably push our metadata records to BY COVID data portal and this suggested metadata field would include URL to the original dataset description in our catalogue)	http://urn.fi/urn:nbn:fi:fsd:T-FSD3532	A		
Mode of collection	Method used to collect data e.g. "Laboratory experiment", "Face-to-face nterview", "Self-administered questionnaire: Web-based (CAWI)". In Social Sciences it would be recommended to use DDI Mode of Collection vocabulary: https://vocabularies.cessda.eu/vocabulary/ModeOfCollection?lang=en	Example 1: <collection_mode vocabulary="DDI Mode of Collection">Laboratory experiment</collection_mode> Example 2 <collection_mode vocabulary="DDI Mode of Collection">Self-administered questionnaire</collection_mode>	A		
Type of the research instrument	The type of data collection instrument used e.g. "Technical instrument(s)", "Questionnaire", "Data collection guidelines: Observation guide" In Social Sciences, it would be recommended to use DDI Type of Instrument vocabulary: https://vocabularies.cessda.eu/vocabulary/TypeOfInstrument?lang=en	Example 1: <research_instrument vocabulary="DDI Type of Instrument">Technical instrument(s)</research_instrument> Example 2: <research_instrument vocabulary="DDI Type of Instrument">Structured questionnaire</research_instrument>	A		

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.