PREPRINT: Umbrella Data Management Plans to integrate FAIR data : lessons from the ISIDORe and BY-COVID consortia for pandemic preparedness

PREPRINT submitted: 15/12/2022 in Data Science Journal (DSJ) Special issue title: [Data Management Planning across Disciplines and Infrastructures]()

Authors: Romain David, Audrey Richard, Claire Connellan, Katharina B Lauer, Maria Luisa Chiusano, Carole Goble, Martin Houde, Isabel Kemmer, Antje Keppler, Philippe Lieutaud, Christian Ohmann, Maria Panagiotopoulou, Sara Raza Khan, Arina Rybina, Stian Soiland-Reyes, Charlotte Wit, Rudolf Wittner, Rafael Andrade Buono, Sarah Arnaud Marsh, Pauline Audergon, Dylan Bonfils, Jose-Maria Carazo, Remi Charrel, Frederik Coppens, Wolfgang Fecke, Claudia Filippone, Eva Garcia Alvarez, Sheraz Gul, Henning Hermjakob, Katja Herzog, Petr Holub, Lukasz Kozera, Allyson L. Lister, José López-Coronado, Bénédicte Madon, Kurt Majcen, William Martin, Wolfgang Müller, Elli Papadopoulou, Christine M.A. Prat, Paolo Romano, Susanna-Assunta Sansone, Gary Saunders, Niklas Blomberg, Jonathan Ewbank, BY-COVID project community, ISIDORe project community.
Contact author: Romain David romain.david@erinha.eu

Author informations:

- Dr Romain David, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, romain.david@erinha.eu, 0000-0003-4073-7456
- Dr Audrey S. Richard, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, audrey.richard@erinha.eu, 0000-0002-0207-0139
- Ms Claire Connellan, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, claire.connellan@erinha.eu, 0000-0001-6449-8512
- Dr Katharina B Lauer, ELIXIR Hub, Wellcome Genome Campus, Hinxton, CB10 1SD, UK, katharina.lauer@elixir-europe.org, 0000-0002-4347-7525
- Pr Maria Luisa Chiusano, Dept. of Agricultural Sciences, University of Naples Federico II, and Stazione Zoologica Anton Dorhn, Naples, Italy, chiusano@unina.it 0000-0002-6296-7132
- Pr Carole Goble, The University of Manchester, Oxford Road, Manchester, M13 9PL, carole.goble@manchester.ac.uk, 0000-0003-1219-2137
- Dr Martin Houde, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, martin.houde@erinha.eu, 0000-0001-8630-204X
- Dr Isabel Kemmer, Euro-BioImaging ERIC Bio-Hub, European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany, isabel.kemmer@eurobioimaging.eu, 0000-0002-8799-4671
- Dr Antje Keppler, Euro-BioImaging ERIC Bio-Hub, European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany, antje.keppler@eurobioimaging.eu, 0000-0003-4358-2269
- Mr Philippe Lieutaud, European Virus Archive - GLOBAL (EVA-GLOBAL), unité des virus émergents - [UVE - UMR190] - Aix Marseille Université, IRD 190, INSERM U1207, France, Philippe.Lieutaud@univ-amu.fr 0000-0002-5080-3456
- Pr Christian Ohmann, European Clinical Research Infrastructure Network (ECRIN), Kaiserswerther Str. 70, 40477 Düsseldorf, Germany,christianohmann@outlook.de, 0000-0002-5919-1003
- Dr Maria Panagiotopoulou, European Clinical Research Infrastructure Network (ECRIN), 5-7 rue Watt 75013 Paris, France, maria.panagiotopoulou@ecrin.org, 0000-0002-4221-7254
- Dr Sara Raza Khan, European Clinical Research Infrastructure Network (ECRIN), 5-7 rue Watt 75013 Paris, France, sara.raza-khan@ecrin.org, 0000-0003-1587-6171

- Dr Arina Rybina, Euro-BioImaging ERIC Bio-Hub, European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany, arina.rybina@eurobioimaging.eu, 0000-0002-5609-5710
- Dr Stian Soiland-Reyes, Department of Computer Science, The University of Manchester, Manchester, M13 9PL, UK & Informatics Institute, University of Amsterdam, LAB42, Science Park 900, 1098 XH Amsterdam, NL soiland-reyes@manchester.ac.uk, 0000-0001-9842-9718
- Dr Charlotte Wit, EU-OPENSCREEN, Robert-Rössle-Straße 10, Building 87, 13125 Berlin, Germany, charlotte.wit@eu-openscreen.eu, 0000-0002-5461-8354
- Dr Rudolf Wittner, Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Neue Stiftingtalstrasse 2, 8010 Graz, Austria, rudolf.wittner@bbmri-eric.eu, 0000-0002-0003-2024
- Dr Rafael Andrade Buono VIB-UGent Center for Plant Systems Biology, Ghent, Belgium, rafael.buono@psb.vib-ugent.be, 0000-0002-6675-3836
- Dr. Sarah Arnaud Marsh, Institut Pasteur (IP), Genetics and Genomics of Insect Vectors Unit (Infravec) & Department of Anthropologie and Ecology of Disease Unit (Sonar-Global), 25-28 rue du Docteur Roux, 75724 Paris Cedex 15, sarah.arnaud@pasteur.fr 0000-0003-4194-043X
- Dr Pauline Audergon, Instruct-ERIC, Oxford House, Parkway Court, John Smith Drive, Oxford, OX4 2JY, UK. pauline@instruct-eric.org, 0000-0002-1624-7388
- Mr Dylan Bonfils, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, dylan.bonfils@erinha.eu, 0000-0002-2734-3420
- Pr Jose-Maria Carazo, CNB-CSIC, Instruct-ES, Campus Universidad Autónoma, 28049, Madrid, Spain, carazo@cnb.csic.es 0000-0003-0788-8447
- Pr Remi Charrel, European Virus Archive - GLOBAL (EVA-GLOBAL), unité des virus émergents - [UVE - UMR 190] - Aix Marseille Université, IRD 190, INSERM U1207, Marseille, France, remi.charrel@univ-amu.fr, 0000-0002-7675-8251
- Dr Frederik Coppens, VIB-UGent Center for Plant Systems Biology, Ghent, Belgium frederik.coppens@psb.vib-ugent.be, 0000-0001-6565-5145
- Dr Wolfgang Fecke, EU-OPENSCREEN, Robert-Rössle-Straße 10, Building 87, 13125 Berlin, Germany, wolfgang.fecke@eu-openscreen.eu, 0000-0003-1917-0037
- Dr Claudia Filippone, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, claudia.filippone@erinha.eu, 0000-0001-5483-7656
- Dr Eva Garcia Alvarez, BBMRI-ERIC, Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Neue Stiftingtalstrasse 2, 8010 Graz, Austria, eva.garcia-alvarez@bbmri-eric.eu 0000-0002-3522-5088
- Dr Sheraz Gul, Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Schnackenburgallee 114, 22525 Hamburg, Germany Fraunhofer; Cluster of Excellence for Immune-Mediated Diseases CIMD, Schnackenburgallee 114, 22525 Hamburg, Germany. Sheraz.Gul@itmp.fraunhofer.de, 0000-0003-2543-1643
- Mr Henning Hermjakob. European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK, hhe@ebi.ac.uk, 0000-0001-8479-0262
- Dr Katja Herzog, EU-OPENSCREEN, Robert-Rössle-Straße 10, Building 87, 13125 Berlin, Germany, katja.herzog@eu-openscreen.eu, 0000-0002-1389-8118
- Dr Petr Holub, BBMRI-ERIC, Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Neue Stiftingtalstrasse 2, 8010 Graz, Austria, petr.holub@bbmri-eric.eu, 0000-0002-5358-616X
- Dr Lukasz Kozera, BBMRI-ERIC, Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Neue Stiftingtalstrasse 2, 8010 Graz, Austria, lukasz.kozera@bbmri-eric.eu 0000-0002-9015-4518
- Dr. Allyson L. Lister, Data Readiness Group, Oxford e-Research Centre, University of Oxford, UK, allyson.lister@oerc.ox.ac.uk 0000-0002-7702-4495
- Dr. José López-Coronado, CECT-University of Valencia, Agustín Escardino 9,46980 Paterna (Valencia) Spain, jmlopez@cect.org 0000-0001-7442-8091
- Dr Bénédicte Madon, La Rochelle Université, Littoral, Environnement et Sociétés (LIENSs), UMR 7266 CNRS – La Rochelle University, 2 rue Olympe de Gouges, 17000 La Rochelle, France bcg.madon@gmail.com , 0000-0001-8608-3895
- Mr Kurt Majcen, BBMRI-ERIC, Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Neue Stiftingtalstrasse 2, 8010 Graz, Austria, kurt.majcen@bbmri-eric.eu, 0000-0002-4041-4887

- Dr William Martin, European Vaccine Initiative,Vossstrasse 2, Geb. 4040, 69115 Heidelberg, Germany, william.martin@euvaccine.eu 0000-0003-4839-1752
- Dr Wolfgang Müller, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany. wolfgang.mueller@h-its.org 0000-0002-4980-3512
- Dr Elli Papadopoulou, ATHENA Research Center / OpenAIRE, Artemidos 6 & Epidavrou, 15125, Athens, Greece, elli.p@athenarc.gr, 0000-0002-0893-8509
- Dr Christine M.A. Prat, European Virus Archive - GLOBAL (EVA-GLOBAL), unité des virus émergents - [UVE - UMR190] - Aix Marseille Université, IRD 190, INSERM U1207, Marseille, France, christine.prat@univ-amu.fr, 0000-0002-6963-8673
- Dr Paolo Romano, IRCCS Ospedale Policlinico San Martino, Largo Rosanna Benzi 10, 16132 Genova, Italy, paolo.romano@hsanmartino.it , 0000-0003-4694-3883
- Prof. Susanna-Assunta Sansone, Data Readiness Group, Oxford e-Research Centre, University of Oxford, sa.sansone@gmail.com, UK 0000-0001-5306-5690
- Dr Gary Saunders, EATRIS-ERIC, De Boelelaan 1118, 1081 HZ Amsterdam, The Netherlands, garysaunders@eatris.eu, 0000-0002-7468-0008
- Dr Niklas Blomberg, ELIXIR Hub, Wellcome Genome Campus, Hinxton, CB10 1SD, UK, niklas.blomberg@elixir-europe.org, 0000-0003-4155-5910
- Dr Jonathan Ewbank, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), 98 rue du Trône B-1050 Bruxelles, Belgium, jonathan.ewbank@erinha.eu, 0000-0002-1257-6862

Abstract

The Horizon Europe project ISIDORe is dedicated to pandemic preparedness and responsiveness research. It brings together 17 Research Infrastructures (RIs) and networks to provide a broad range of services to infectious disease researchers. An efficient and structured treatment of data is central to ISIDORe's aim to furnish seamless access to its multidisciplinary catalogue of services, and to ensure that users' results are treated FAIRly. ISIDORe therefore requires a data management plan (DMP) covering both access management and research outputs, applicable over a broad range of disciplines, and compatible with the constraints and existing practices of its diverse partners.

We undertook an iterative, step-by-step, process to build a community-approved living document, identifying good practices and processes, on the basis of use cases, presented as proof of concepts. International fora such as the RDA and EOSC, and primarily the BY-COVID project, furnished registries, tools and online data platforms, as well as standards, and the support of data scientists. Together, these elements provide a path for building an umbrella, FAIR-compliant DMP, aligned as fully as possible with FAIR principles, which could also be applied as a framework for data management harmonisation in other large-scale, challenge-driven projects. Finally, we discuss how data management and reuse can be further improved through the writing of realistic DMPs using 'DMP profiles' and, in the future, the contribution of an inter RIs data steward network, to produce a Community of Practice that could be integrated into planned trans-RI competence centres.

Keywords: Data Management Plan, DMP, Research Data Management, RDM, data management quality, research data, pandemic, outbreak, data services harmonisation, data reuse, FAIR principles, data stewards

Introduction

Data Management Plans (DMPs) serve as framework documents in research projects (Smale et al., 2020) and are necessary tools for data reuse across various scientific and engineering fields. DMPs are now either recommended (e.g. by the Organisation for Economic Co-operation and Development [OECD]), or even required for many grant applications, including those of various national funding bodies, such as the French National Research Agency [ANR] and the U.S. National Science Foundation [NSF DMP] and National Institutes of Health [NIH].

In 2015, Michener detailed ten key items of any research DMP to support future management and possible reuse of data: (i) funders' requirements, (ii) data to be collected, (iii) data organisation, (iv) data documentation, (v) data quality assurance, (vi) data storage and preservation strategy, (vii) project's data policies (licensing, ethical considerations, etc.), (viii) data dissemination, (ix) team members' roles and responsibilities, and (x) budgetary aspects. Templates from many funders align with this model (e.g. Horizon Europe Data Management Plan Template [HE DMP]).

Research data can have enduring value for scientific progress, as scientists use, reuse and combine data sets in new analyses, to test new hypotheses and reach novel conclusions. This is the concept underlying a data life cycle – where research data can often have use beyond its original purpose (Jacobs & Humphrey, 2004). Effective 'data sharing' requires that data is made accessible, with clear conditions for its access and reuse. The related but distinct concept of 'open data' implies that there are no limitations to prevent anyone from obtaining the data. Be it through data repositories or clearly-defined direct requests to the original data owners, data that is not open can still be made accessible as long as the procedures for obtaining it are defined (Hutson, 2022).

While usually relatively straightforward for single investigator-driven projects, DMPs are more difficult to define and maintain in the case of international, multidisciplinary and/or multi-partner research projects (e.g. Stall et al., 2020). Moreover, an effective DMP must not become obsolete after a project is completed (Vines et al., 2014), and needs to address the implementation of Findable, Accessible, Interoperable, Reusable (FAIR) data principles (Wilkinson et al., 2016), whether or not the data is produced by the project itself.

To apply the FAIR principles as broadly as possible, and increase data reuse in a pragmatic way, as proposed by David et al. in 2020, community-approved steps (e.g. preparing, implementing, assessing, etc.) must be shared by all partners of the project, in an iterative and sustainable manner. Moreover, a DMP should undergo revisions during the complete lifetime of a research project, and should ideally be updated to facilitate data reuse and openness (as possible and necessary) after the project's conclusion (Williams, Bagwell, & Nahm Zozus, 2017). Researchers report significant benefits from a well-prepared DMP that outweigh the initial investment in implementing thorough data management practices (Burnette, Williams & Imker, 2016).

Challenges for the ISIDORe Project and its data

Introduction of ISIDORe

[ISIDORe] (Integrated Services for Infectious Disease Outbreak Research) is a three-year project coordinated by [ERINHA], funded in 2022 through the Horizon Europe Research and Innovation Programme. It aims to support researchers through the provision of free of charge access to services, to facilitate rapid research response to infectious disease threats, and improve the EU's readiness to epidemic-prone pathogens.

ISIDORe brings together 17 major European life-science research infrastructures (RIs) and infectious disease networks in a multidisciplinary consortium of 154 partners[1] across 32 countries. Collectively, the consortium offers open access to an integrated portfolio of cutting-edge resources and research services covering fundamental research, such as structural biology and bio-imaging, as well as diagnostic, therapeutic and vaccine preclinical development, clinical research, epidemiology, social sciences and regulatory matters.

International requirements

The international nature of the project entails for the ISIDORe DMP to align with different federal and European requirements. The main resulting challenges include i) implementing the requirements and recommendations from international fora [e.g. [CoData], Research Data Alliance [RDA], World Data System [WDS], [Go-FAIR] or the [EOSC], ii) harmonising data sharing from heterogeneous scientific initiatives and iii) ensuring the FAIRification of all data, including sensitive data, by a legally and ethically compliant approach, ensuring confidentiality and privacy.

Diversity of partners and services

While some of the RIs participating in ISIDORe had already collaborated in the frame of RI cluster projects such as [RI-VIS], [CORBEL] and [EOSC-Life], the ISIDORe project represents a unique challenge, due to the high diversity of participating RIs and networks, in terms of their respective sizes, expertise, legal structures and/or working practices. This required a new approach to coordination and collaboration among the different partners. This also made the elaboration of plans for data management, to favour the fast generation of FAIR policies for data sharing, reuse and interoperability, and open access to existing methods and software, such as on line tools, workflows and registries, particularly challenging.

The diversity of partners and services included in ISIDORe requires its DMP to be broadly interpretable, while preserving the necessary individual specificities, so as not to become meaningless, by preserving two aspects: the general principles that will be used to approach data management, and the specifics for services and use cases.

[1] The Consortium brings together 154 partners under the umbrella of ERINHA, BBMRI, EATRIS, ECRIN, Elixir, EMBRC, EMERGEN, Euro-BioImaging, EU-OPENSCREEN, EVA, Infrafrontier, Infravec, Instruct, MIRRI, Sonar-Global, Transvac, and VetBioNet.

Diversity of partner maturity

Data standardisation and accessibility can be maximised by careful adherence to FAIR, CARE (Carroll et al., 2020), TRUST (Lin et al., 2020) and other established principles in the data production and management processes. Furthermore, to avoid the duplication of effort and a lack of standardisation, it is critical to identify and adapt pre-existing online tools, community-adopted standards and processes, and implement the relevant complementary ones, especially from other multidimensional data science networks. In addition, the generation of new data needs to be carefully considered to avoid the repetition of studies or study groups, such as control groups.

In addition to complying with the various federal laws, ISIDORe's DMP must also meet the (sometimes diverging) recommendations and procedures for handling data of the consortium's partners. While some, especially in Life Science (LS)-RI communities with extensive experience in data handling (e.g. structural biology: [INSTRUCT], sequencing, high-content screening: [EU-OPENSCREEN], biological and biomedical imaging: [Euro-BioImaging]) have a solid set of internal procedures, implemented in close exchange with the relevant repositories (e.g. [PDB], [Uniprot], [Ensembl], [BioImage Archive], [EMPIAR], [IDR], [ECBD], etc.), others still rely primarily on the diligence of individual researchers. The challenge is therefore to accommodate this high heterogeneity. It requires first determining precisely the current state-of-the-art in data management across the RIs and networks, to reconcile differences and avoid duplication of the efforts already in place. The aim is to create an overarching core DMP that encompasses all existing processes, and defines minimal procedures for those partners that do not currently have a fully-developed DMP.

Dealing with sensitive data

A further challenge is linked to the type of data that is generated within the ISIDORe project. Certain partners produce data with intrinsic biosecurity and/or biosafety issues. For others, there are concerns regarding the potential of identifying individuals from clinical or social science data.

The ability to conduct meta-analyses of data from cross-disciplinary studies is especially important in emergency situations, such as in a pandemic. A well-designed DMP that also ensures linking and accessing information on already existing data and relevant research findings (such as on pathogens of concern) will provide a roadmap not only on how to handle the data, but also on the establishment of processes to handle the mix of sensitive and non-sensitive metadata, and assess potential risks for data misuse.

ISIDORe Project DMP building method

A survey to define the landscape

To overcome the challenges described above, ISIDORe's strategy included building on existing components. For their comprehensive mapping, a survey about data management tools was launched among the participating RIs and networks. One of the findings of the survey was that DMP tools are often not yet integrated into researchers' project workflows.

Therefore, the ISIDORe DMP must indicate i) where researchers can find guidance for the stewardship and preservation of both data and software, ii) where and how DMP tools and content can be easily and efficiently integrated. For the latter, repository selection, licensing, formats and metadata recommendations, including use of identifiers for building linkages to external systems which provide data interoperability, as defined by the FAIR core principles, must be listed and updated during the project The timing and periodicity of specific actions for team members and leaders have to be addressed and updated (with their respective contributions).

Alignment with existing recommendations

ISIDORe implements the data policy description workflow that was created by the Digital Curation Centre (on behalf of the FAIRsFAIR project) and FAIRsharing (Sansone et al., 2019): it is designed to help with the creation of FAIR-aligned data policies [DCC_FAIRsharing]. Three community-developed data policy description efforts ([FAIRsharing policy record metadata], the FAIRsFAIR FAIR Data Policy Checklist (Davidson et al., 2022) and RDA (Hrynaszkiewicz et al., 2020)) have been aligned. In addition, all checklist fields and RDA-endorsed policy features within its scope are available within FAIRsharing policy records, making them accessible to both humans and machines. This creates a FAIR data policy 'workflow' from a) FAIR Data Policy Checklist to b) deposition of the policy and assignment of a Digital Object Identifier (DOI), to c) submission of that policy into FAIRsharing and other adequate repositories. Following this workflow will ensure the FAIRification of ISIDORe data management policies.

Mapping requirements for the DMP

The ISIDORe DMP will make use of pre-existing data repository concepts and integrate available data management tools and guidelines as much as possible (Figure 1), including [RDMKit], the [FAIR Cookbook] (Rocca-Serra et al., 2022), the Data Stewardship Wizard [DSW] (Pergl et al., 2019), and the Infectious Diseases Toolkit [IDTk]. Despite this relatively esoteric landscape, the DMP should be easy to understand by the different ISIDORe partners, so that they can apply its guidelines properly. The toolbox developed for discovery of digital objects related to sensitive data, harmonised across 6 life-science infrastructures and developed using an iterative process, will serve as an example of how to deal with issues arising across research infrastructures (David et al., 2022).

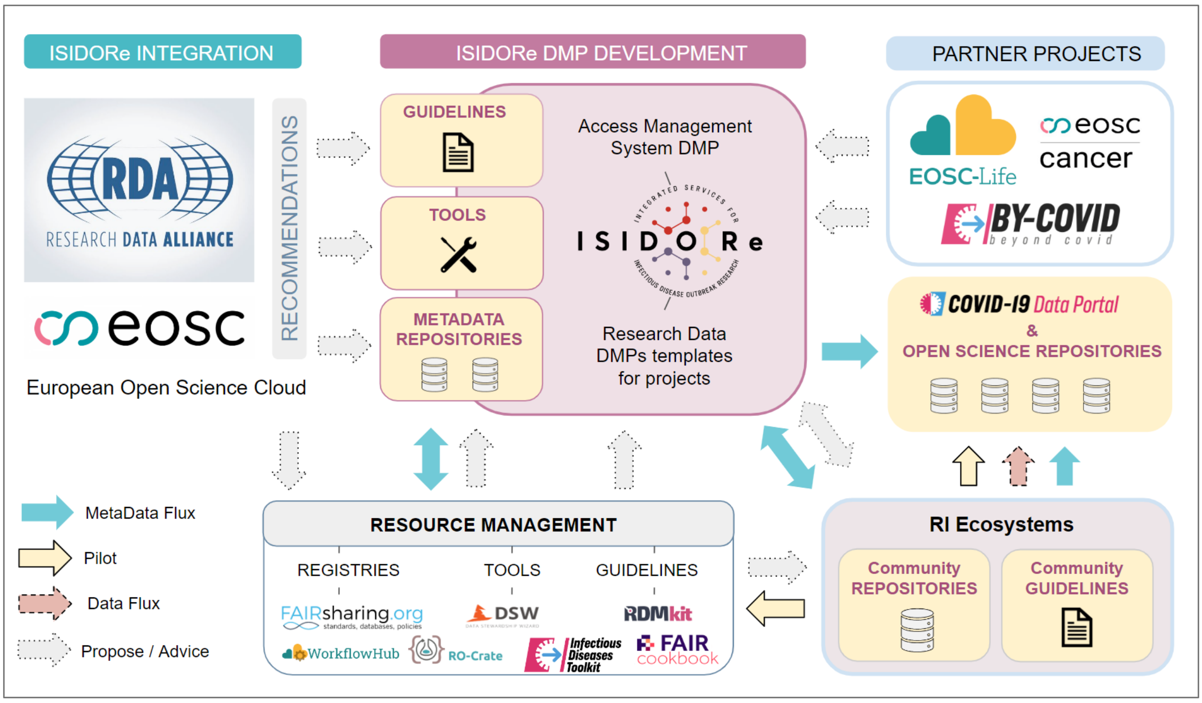

Figure 1: DMP building scheme for the ISIDORe project:

The ISIDORe DMP is being developed by integrating available DMP guidelines, resources and advice provided by participating RI communities, partner EU projects (such as those shown) and the overall recommendations by the RDA and EOSC. The established DMP procedures, FAIR guidelines, data repositories and metadata will be collected through platforms and tools (such as those illustrated). A live representation of the standards and databases used within the DMP will be available as a FAIRsharing collection. While many LS-RI partners already have DMP workflows in place, the goal and challenge of the ISIDORe project will be to set up cross-disciplinary pipelines, onboarding additional communities and adapting DMP procedures to meet the needs of infectious diseases research (data and tools) for the different disciplines in the consortium.

A current investigation of data sharing statements (DSS), as included within the trial registry entries of COVID-19 related studies, has revealed strong differences in how the request for data sharing details has been interpreted and, when data sharing is described as possible, a huge heterogeneity in the specification of the access procedures (Canham et al., 2022). This implies that the requests for data sharing statements within trial registry data need to be more clearly explained and explicitly structured, and where possible, standardised category systems to indicate the type of access that will be available should be used.

To organise this homogenisation, ISIDORe MetaData mobilisation is based on an Access Management System (AMS). The main aim of the AMS is to administer responses to calls and requests for services, and to collect feedback from users in a standardised and centralised manner. This AMS, currently in development, will allow the management of the catalogue of services, project submissions information with all associated metadata, and create a community-validated registry of projects (both those submitted and accepted). Existing data registries and repositories used by accepted projects will be indexed in the ISIDORe AMS and syndicated through metadata tools and registries (described in the Supplementary Material). Metadata mobilisation will be achieved through existing RI AMSs when they exist (Figure 1); associated syndication processes are described in the ISIDORe DMP.

Furthermore, since the project's contributors are numerous and diverse, the DMP should be flexible to fit the overall needs, whilst retaining minimal common requirements in standards, protocols and rules for metadata, traceability and provenance of data, as well as ease of access. This work on the ISIDORe DMP may introduce elements that could significantly impact the DMPs of individual project's members, making them more general and potentially more ambitious.

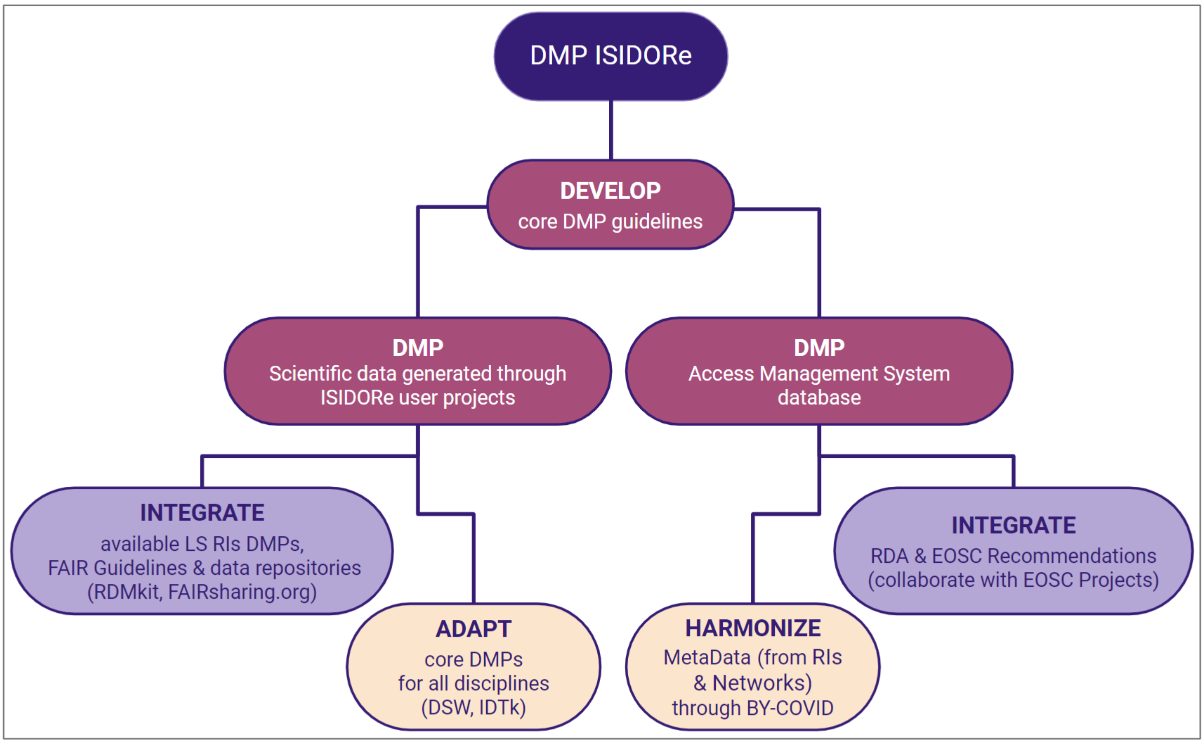

Figure 2 The two sides of ISIDORe DMP approach

ISIDORe requires a data management plan (DMP) covering both Access Management Systems (AMSs) and scientific data generated through ISIDORe user projects. The Scientific Data part integrates available LS RI DMP components and adapts them for all disciplines. The AMS part harmonises management metadata from RIs and networks with the support of BY-COVID and other EOSC/RDA recommendations. Both approaches are applicable over a broad range of disciplines, and must be compatible with the constraints and existing practices of its diverse partners.

The FAIR-compliant DMPs, covering service management and delivery, have generally been designed and implemented at the level of the RIs and networks. This type of DMP is distinct from the DMP covering the production of scientific data (Figure 2). In this case, in many of the distributed RIs and networks that are partners in ISIDORe, the onus to support users in DMP compliance has previously fallen on the individual service providers (e.g. national or regional core facilities).

All data challenges in ISIDORe are supported by a sister project called BeYond-COVID [BY-COVID], the European Open Science Cloud (EOSC) project to store and share pandemic related research data. The BY-COVID project aims to make data and analysis tools related to infectious disease outbreaks more interoperable and accessible, not only to scientists but also to medical staff in hospitals, and policy makers, such as government officials, or indeed any other potential user.

These two projects are working together: BY-COVID provides support, tools, data discovery and dissemination strategies to enable data from user projects implemented by ISIDORe to be as FAIR as possible, while ISIDORe promotes the adoption of FAIR (meta)data standards by partners via its network of data stewards, and tries to ensure that data generated by users is disseminated into the public domain as widely as possible, preferably via dedicated sites such as the [COVID-19 Data Portal]. As the digital data generated from each research project needs to be actively managed over time to ensure availability and useability, depositing data resources with a trusted digital archive can ensure that they are curated and handled according to good practices in digital preservation [ICPSR_DMP]. Through the COVID-19 data portal, BY-COVID provides a platform for data discovery across a broad range of disciplines, with the flexibility to add any relevant data resources proposed by ISIDORe through a lightweight metadata format jointly developed by both projects (Hermjakob et al., 2022). The collaboration between the two projects will lay the groundwork for Europe's future pandemic preparedness with FAIR data pipelines enabling rapid response to outbreak situations and a broad framework for answering research questions across research disciplines and country borders.

Furthermore, BY-COVID provides standardised data management and analysis methods (partially based on the Galaxy Project), protocols, and training to ensure FAIR is an integral part of the data pipeline. Building on existing infrastructure, data repositories are explorable through a single platform (COVID-19 data platform and in the future EBI pathogen portal) allowing users to deposit, access and analyse pathogen data more easily. FAIRsharing is the BY-COVID data source catalogue of choice. It describes relationships among the data sources (databases, knowledge bases, repositories) and assists with the selection, documentation and visualisation of standards. The use of [RO-crate] (Eguinoa et al., 2022), based on FAIR Digital Object (FDO) principles, guarantees efficient data sharing, including data and metadata compliance using community-adopted standards such as Bioschemas and Common Workflow Language. Workflow repositories such as [WorkflowHub] that include the use of persistent and unique identifiers (e.g. DOIs, [cool URIs]) ensure findability and discoverability of research computational workflows and pipelines. Data discovery is further enabled by the BY-COVID data discovery network, based on the principles of the Beacon Network [employing an Application Programming Interface (API) that allows for instance discovery of genomic, phenotypic and clinical data].

Building a bridge between partners' data

As ISIDORe aims to provide free access to a large portfolio of scientific services across disciplines,the implementation of the majority of user projects will generate experimental data. Efficient integration of common DMP principles will require advising researchers at very early stages of their projects and guiding them through FAIR principles and the available tools and resources that fit their projects' data needs as well as open science requirements. The current DMP strategy in Euro-BioImaging, for example, offers incoming users a consultation with the RI's FAIR Data steward before data is acquired. On the one hand, this raises awareness of the available guidelines, repositories and tools, and on the other hand, it brings data generation onto the DMP track. It also fosters timely adoption of the DMP with regards data provenance and life cycle management requirements, as well as readiness for data integration into open science frameworks after data generation and analysis are completed. Given the potential impact of ISIDORe, one additional key is to establish a directed cross-consortium exchange between individual FAIR data stewards from the different RIs, to support the harmonisation of DMP principles and procedures.

One solution to the problem of a comprehensive, broadly applicable yet specific DMP is to divide it into sections, including a general, overarching one, describing overall data management principles (so called “core DMP”), and RI/field-specific sections. This may be achieved through several curated sets of questions (knowledge models) created in the DSW and pre-filtered by the different ISIDORe partners, depending on the domain to which the data belong and how data is to be disseminated. The generation of RI-specific curated catalogues will greatly benefit from, and be accelerated by, existing collections of domain-specific and recognized resources and repositories that may already be available for RIs (such as those collected through FAIRsharing, RDMKit and IDTk), and through the existing collaboration of RIs with data and workflow repositories. Indeed, these may have already paved the way for easy and fast submission of data.

Projects that require the provision of services by two or more RIs, and whose datasets must therefore comply with the requirements and recommendations of those RIs, present a special challenge. Again, the introduction of a two-tier DMP will help to overcome this, although the participating RIs will still need to reach a consensus on the recommendations to be given to users regarding their specific data constraints. One can imagine that this will also be facilitated by the variability of knowledge models within the DSW. Ideally, the RI or network (preferably via its data steward) should be automatically alerted of possible conflicting recommendations or overlapping issues, and thus ISIDORe will be able to provide users with a curated and coherent catalogue.

Building alignment between consortium partners on the core DMP principles, tools and implementation of available resources to develop harmonised DMP workflows can be fostered by consolidating a consortium-wide data steward team/committee. Data steward representatives from consortium RIs and Networks are to be invited to participate in shaping and curating ISIDORe’s DMP, in collaboration with partner projects and initiatives such as those developed through [BY-COVID], [EOSC-Life] and [EOSC4Cancer] and taking in account the procedural details of their respective DMPs (García Álvarez, Mayrhofer et Holub, 2021, Blomberg, Sarntivijai and Mayrhofer, 2020). This will allow the creation of sustainable exchange and cross-community validated solutions on the RI landscape.

Sensitive data within ISIDORe DMP

Based on the experience of the [CORBEL] project, with its multi-stakeholder task force, sensitive-data experts previously examined major issues associated with sharing of individual sensitive data. They developed a consensus document on providing access to such individual participant data, using a broad interdisciplinary approach. This included 10 principles and 50 recommendations (Ohmann et al., 2017), representing the fundamental requirements of any framework used for the sharing of clinical research data. The document, to be included in ISIDORe’s DMP, covers the following main areas: making data sharing a reality (e.g. cultural change, academic incentives, funding), consent for data sharing, protection of trial participants (e.g. de-identification), data standards, rights, types and management of access (e.g. data request and access models), data management and repositories, discoverability, and metadata. Its adoption helps promote and support data sharing and reuse among researchers, adequately inform trial participants, and protect their rights, and provides effective and efficient systems for preparing, storing and accessing data.

Access Management Systems as a glue

The main objective of the ISIDORe Access Management System (AMS) is to administer responses to calls for proposals and requests for free access to the project's services, as well as to collect feedback from users in a standardised and centralised manner. Some of the ISIDORe partners already have a pre-existing AMS, key to the facilitation of good data management in access provision within ISIDORe. Therefore, ISIDORe has conceived a central ISIDORe AMS system as a decoupled 'headless' component that will feed and be fed by the partners' pre-existing AMSs. For that purpose, minimal interoperability elements with the central ISIDORe AMS have been defined which the partners' AMSs will have to implement in order to provide the central system with up-to-date, formatted content regarding the details of the services proposed and the status of the requests for services that they process.

For partners that do not have a pre-existing AMS or cannot ensure the transmission of the set of minimally-defined data, a new basic AMS system that matches the minimal interoperability criteria with the central AMS will be implemented. This could either be developed from scratch or be based on elements from one of the 2 pre-existing systems in use by partner RIs, [ARIA], and the [EVAg] portal. EU-OPENSCREEN, Euro-BioImaging, MIRRI, EMBRC and Instruct, use the ARIA AMS. They also use it for managing user requests submitted in response to ISIDORe's calls for proposals. ARIA can handle many aspects of RI operations, from proposal submission, and the evaluation process, to keeping a record of users' on-site visits and their feedback. Such a system enables all application and access provision data associated with a proposal to be collected and findable and is compliant to the General Data Protection Regulation [GDPR]. Additional developments are planned for ARIA to allow for the output of a proposal, such as publications and datasets, to also be findable via the AMS. This would be a powerful tool to allow data traceability. EVA-GLOBAL's AMS is currently based on the Drupal open-source Content Management System (CMS) and also implements GDPR compliance. Similar to ARIA, the EVA portal manages many aspects of RI operations and service provision. Currently used by the EVA and Infravec networks, it has the advantage of being user-friendly and has demonstrated great adaptability in the context of rapid responses to epidemic and pandemic situations (including the Zika epidemic, COVID-19 pandemic and most recently, the 2022 Mpox outbreak). This flexibility would be a great asset, especially considering ISIDORe's expected role in rapid research response.

Discussion

Exploiting the ISIDORe framework

The ISIDORe project is a unique opportunity for the 17 participating RIs and networks to work together to establish a new approach for service provision, to improve pandemic preparedness and accelerate research in outbreak response. The project needs to reconcile the partners' disparate philosophies, processes and practices, as well as ensuring that they all embrace common contemporary standards, while implementing best practices. The aim is to establish a working consensus for this broad, multifaceted and complex community. The task is made more challenging since ISIDORe has been designed to collaborate closely with its sibling data-focused project, BY-COVID.

The iterative processing of the DMP

Our goal here is to describe the joint efforts of ISIDORe and BY-COVID towards the establishment of a complete and harmonised DMP, through a living document, allowing contributions from all stakeholders, prior to finalisation of a definitive and functional version. As DMPs require upgrading in an iterative fashion by all partners involved in their creation, the use of a web-based DMP tool (e.g, DSW, EOSC [DMPOnline]) that allows users to contribute in parallel, manages versioning and provides guidance in building the document, is paramount during the complete maturation cycle of the project's DMP.

How to exploit pre-existing efforts

The ISIDORe DMP has to deal with two facets of the project, the RI/network-centric access provision, and the service provider level data generation at each project scale. Thus, it cannot simply be a merging of institutional-level DMPs, but it needs to be designed through a dynamic agreement among all partners, based on the best current protocols to manage and exchange data in the respective fields. The partners with more advanced best practices act as drivers, laying the groundwork to define the content of the DMP, while the active contribution of all the other partners ensures that all needs and perspectives are appropriately considered and integrated in the set-up of a general proof of concept, with successive refinements and adaptations to the respective specificities and requirements.

Stakeholders involved need to develop equitable and sustainable financial models for data sharing, to ensure the long-term resourcing of data preparation and storage, as well as the request and sharing process. These tasks are not easily predictable and thus not easily linked to initial funding. The discussion regarding sustainable business models for data infrastructures is ongoing, and it is difficult to identify a preferred model yet. A particular problem is that, while many established national and international data repositories have core streams of income from research funders, these sources of income are usually short-lived and may be vulnerable to change in priorities or in responsibilities (Ohmann et al., 2017).

Diversity and specificities integration

To integrate FAIR, CARE and TRUST principles, ISIDORe necessarily adapted basic recommendations to this interdisciplinary, international project. It additionally needed to take into consideration the fact that the individual partners in the project could have conflicting priorities, or had their own prior implementation of FAIR data sharing. This requires an inclusive harmonisation for data management practices, as a basis for future global meta analyses.

The adoption of the FAIR principles has encouraged researchers to comply with data and metadata standards. This need for 'FAIRification' has spurred several international efforts (cf. description in Supplementary Material), and members of both the ISIDORe and the

BY-COVID consortia are either directly involved in these efforts, or engaged in the exploitation of these approaches in the context of their respective projects.
The interoperability principle in FAIRification can only come into play when the semantics of the content are well-defined across heterogeneous data sources (Holub et al., 2018). Ontologies are one of the semantic tools that are frequently used to support interoperability (Haque et al., 2022). The ISIDORe project, therefore, might benefit from the development of an ontology project as a guide to the elaboration of the DMP, especially for the section concerning the implementation of FAIR principles and open data (distinct from the [OBO academy's] domain ontologies).Such an ontology project would aim at clarifying the relationships between the entities (in the ontological sense) and metadata, as well as the selection of data standards and data types. Reusing project partners' existing metadata, and data standards, might simply be done by merging the information from each partner infrastructure in ISIDORe and agreeing on the most consistent one in case of equivalent candidates. A better approach would, however, be to evaluate for each data/metadata standard, if a 'linked data', 'linked open data' or even better, 'FAIR linked open (meta)data' standard exists that could be used (Frey and Hellmann, 2021). Using such new linked (meta)data standards would contribute to the semantic web and thus highly increase the potential findability of ISIDORe data, as well as its capacity to be interconnected, the ability to give it additional meaning and the potential of its syntactic power in data mining processes. Wikidata entities and their associated metadata could be good FAIR-linked open metadata candidates for ISIDORe data, as they are an essential part of the current semantic web, validated and maintained by a large community and could be expanded by the ISIDORe community with consistent novel inputs (Waagmeester et al., 2020). Thus, in an ideal world, by using linked data as metadata, we might enable ISIDORe data to be linked to the semantic web.

Further steps and perspectives

The ISIDORe project has a laudable mission to define and implement best practices in collaborative data management, with a fundamental objective to promote their adoption by the community in the most effective way. ISIDORe has the potential to lead the way with innovative aspects relating to data controllers, data processors, future proofing data sets (both data type and format) regarding storage, processing and transfer. Additional aspects to be considered will include data security, GDPR compliance, quality control, data anonymization, analysis and traceability of data as well as communication and knowledge management.

Indeed, the heart of the work will consist of data mobilisation planning, to convince all partners of the importance of replicating methods for which there is already an established implementation, or at least a proof of concept, all the while taking into account the human dimension (FAIR literacy, and the availability of qualified personnel for support, training and the assessment of results [Mons, B, 2015]). The quality criteria that are identified at each step will have to be constantly revisited. In the end, dissemination, support for DMP application, training and (self)assessment must be FAIR and re-used, and could potentially feed into RDA recommendations. An ultimate challenge for the ISIDORe community could be the implementation of machine actionable DMPs (Miksa, Walk & Neish, 2020) that adhere to all DMP Common Standards [RDAMSC]. But for that to happen, as for most research projects, FAIR DMP Literacy provided by a well organised network of data stewards must first be improved.

Acknowledgements

This work is mainly a product of ISIDORe (Integrated Services for Infectious Disease Outbreak Research) funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº101046133. It is also a product of the BY-COVID Project funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº101046203. Complementary support was provided through the EOSC-Life European program under grant agreement Nº824087. Special acknowledgement to AM, DD, KE, MT, PG, SA, for filling the survey without being an author of this paper.

Competing Interests

The authors have no competing interests to declare.

CRediT authorship contribution statement

- Conceptualization: ARi, CC, JE, NB, RD
- Original Draft Preparation: CC, JE, RD
- Writing Paper: ARi, AK, ARy, CC, CG, CO, CW, IK, KL, JE, MLC, MH, MP, NB, PL, RD, RW, SRK, SS
- Reviewing & editing paper: ARi, AK, ARy, CC, CG, CO, CW, IK, KL, JE, MLC, MH, MP, NB, PL, RD, RW, SRK, SS, RA, SAM, PA, DB, JMC, RC, FC, WF, CF, EGA, SG, HH, KH, PH, LK, AL, JLC, BM, KM, WMa, WMü, EP, CP, PR, SAS, GS
- Figure conception: ARy, RD - Figure contributions: ARy, RB, SA, RD
- Survey conception: ARi, CC, JE, RD
- Survey completion (within the authors): GS, KH, KL, KM, PA, PL, RD, SR, WMa

Data Availability statement

CC-By licence - Data contact:RD. Survey & survey data will be made available after anonymisation with the ISIDORe DMP deliverable

References

Bishop, B, Gunderman, H, Davis, R, et al., 2020, 'Data Curation Profiling to Assess Data Management Training Needs and Practices to Inform a Toolkit', *Data Science Journal*, 19(1), p.4. DOI: http://doi.org/10.5334/dsj-2020-004

Blomberg, N, Sarntivijai, S, Mayrhofer, MTh., 2020. 'EOSC-Life Data Management Plan', *Zenodo*. DOI: https://doi.org/10.5281/zenodo.4010442

Burnette, M H & Williams, S C & Imker, H J, 2016, 'From Plan to Action: Successful Data Management Plan Implementation in a Multidisciplinary Project', *Journal of eScience Librarianship* 5(1): 6. DOI: https://doi.org/10.7191/jeslib.2016.1101

Canham, S, Felder, G, Ohmann, C & Panagiotopoulou, M, 2022 'Identification of COVID-19 clinical studies intending to share individual participant data for secondary use: Protocol for a pilot study', *Zenodo*. DOI: https://doi.org/10.5281/zenodo.7064624

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., et al., 2020. 'The CARE Principles for Indigenous Data Governance'. *Data Science Journal* 19, 43. DOI: https://doi.org/10.5334/dsj-2020-043

David, R, Mabile, L., Specht, A. et al. RDA – SHAring Reward and Credit (SHARC) Interest Group R.D.A., 2020, 'FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles', *Data Science Journal*, 19(1), p.32. DOI: http://doi.org/10.5334/dsj-2020-032

David, R, Ohmann, C, Boiten, JW et al., 2022, 'An iterative and interdisciplinary categorisation process towards FAIRer digital resources for sensitive life-sciences' *data. Sci Rep* 12, 20989. DOI: https://doi.org/10.1038/s41598-022-25278-z

Davidson, J., Grootveld, M., Verburg, M.,et al., 2022, 'FAIR-enabling Data Policy Checklist', *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6225775

Eguinoa, I, Suchánek, M, Knaisl, V, et al., 2022, 'BioHackEU22 Report: Enhancing Research Data Management in Galaxy and Data Stewardship Wizard by utilising RO-Crates.' *BiorXiv* (in prep)

Frey, J, & Hellmann, S, 2021, 'FAIR Linked Data - Towards a Linked Data Backbone for Users and Machines', in *Companion Proceedings of the Web Conference 431–435 (Association for Computing Machinery, 2021)*. DOI: https://doi.org/10.1145/3442442.3451364

García Álvarez, E, Mayrhofer, M, Holub, P, 2021, 'BY-COVID - D8.2 - Data Management Plan', *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6884816

Giraldo, O, Alves, R, Bampalikis, D, et al., 2022, 'A FAIRification roadmap for ELIXIR Software Management Plans', 1st International Conference on FAIR Digital Objects (FDO 2022) *Research Ideas and Outcomes* 8, e94608. DOI: https://doi.org/10.3897/rio.8.e94608

Goble, C, Soiland-Reyes, S, Bacall, F, et al., 2021, 'Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory'. *Zenodo*, DOI: https://doi.org/10.5281/zenodo.4605654.

Goble, C, Bacall, F, Soiland-Reyes, S, et al., 2022, 'WorkflowHub – a FAIR registry for workflows'. *F1000Research,* 11. DOI: https://doi.org/10.7490/f1000research.1118984.1

Haque, AKMB, Arifuzzaman, BM, Siddik, SAN, et al., 2022, 'Semantic Web in Healthcare: A Systematic Literature Review of Application, Research Gap, and Future Research Avenues'. *Int J Clin Pract*. 2022 Oct 18;2022:6807484. DOI: https://doi.org/10.1155/2022/6807484

Hermjakob, H, Kleemola, M, Moilanen, K, et al., 2022, 'BY-COVID - D3.1 - Metadata standards. Documentation on metadata standards for inclusion of resources in data portal'. *Zenodo*, DOI: https://www.doi.org:10.5281/zenodo.6885016

Holub, P, Kohlmayer, F, Prasser, F, et al., 2018, 'Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health'. *Biopreserv Biobank* 16, 97–105. DOI: https://doi.org/10.1089/bio.2017.0110

Hrynaszkiewicz, I, Simons, N, Hussain, A, et al., 2020, 'Developing a Research Data Policy Framework for All Journals and Publishers'. *Data Science Journal* 19, 5. DOI: https://doi.org/10.5334/dsj-2020-005

Hutson M, 2022, 'Taking the Pain out of Data Sharing'. *Nature*. Vol 610: pages 220-221. DOI: https://doi.org/10.1038/d41586-022-03133-5

Jacobs, JA, Humphrey, C, 2004, 'Preserving research data'. *Commun. ACM* 47(9), 27–29. DOI: https://doi.org/10.1145/1015864.1015881

Kesisoglou, I, Cosgrove, S, Derycke, P, et al., 2022. 'Glossary of commonly used terms in the field of health data research' - *developed by the EU project HealthyCloud*. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6787119

Lin, D, Crabtree, J, Dillo, I, et al., 2020, 'The TRUST Principles for digital repositories'. *Sci Data* 7, 144 (2020). DOI: https://doi.org/10.1038/s41597-020-0486-7

Michener, WK, 2015. 'Ten Simple Rules for Creating a Good Data Management Plan'. *PLoS Comput Biol* 11, e1004525. DOI: https://doi.org/10.1371/journal.pcbi.1004525

Ohmann C, Banzi R, Canham S, et al., 2017, 'Sharing and reuse of individual participant data from clinical trials: principles and recommendations', *BMJ Open*;7:e018647. DOI: https://doi.org/10.1136/bmjopen-2017-018647

Pergl, R, Hooft, R, Suchánek, M, et al., 2019, 'Data Stewardship Wizard: A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning'. *Data Science Journal*, 18(1), p.59. DOI: http://doi.org/10.5334/dsj-2019-059

Rocca-Serra, P., Gu, W, Ioannidis, V,, et al., 2022, 'The FAIR Cookbook - the essential resource for and by FAIR doers'. *Zenodo*, DOI: https://doi.org/10.5281/zenodo.7156792

Sansone, S-A, McQuilton, P, Rocca-Serra, P, et al., 2019. 'FAIRsharing as a community approach to standards, repositories and policies'. *Nat Biotechnol*, 37: 358–367. DOI: https://doi.org/10.1038/s41587-019-0080-8

Smale, N, Unsworth, K, Denyer, G, et al., 2020. A Review of the History, Advocacy and Efficacy of Data Management Plans. *International Journal of Digital Curation*, 15(1), 1–29. DOI: https://doi.org/10.2218/ijdc.v15i1.525

Stall, S, Specht, A, Corrêa, PLP, et al., 2020. 'PARSEC Data and Digital Output Management Plan and Workbook'. *Zenodo,* DOI: https://doi.org/10.5281/zenodo.3891426

Soiland-Reyes, S, Sefton, P, Crosas, M, et al., 2022a, 'Packaging research artefacts with RO-Crate'. *Data Science* 5 (2). DOI: https://doi.org/10.3233/ds-210053

Soiland-Reyes, S, Sefton, P, Jael Castro, L, et al., 2022, 'Creating lightweight FAIR Digital Objects with RO-Crate', 1st International Conference on FAIR Digital Objects (FDO 2022) *Research Ideas and Outcomes* **8**:e93937. DOI: https://doi.org/10.3897/rio.8.e93937

Miksa, T, Walk, P, Neish, P, 2020. RDA DMP Common Standard for Machine-actionable Data Management Plans. RDA recommendation DOI: https://doi.org/10.15497/rda00039

Miksa, T, Jaoua M, Arfaoui G, 2020, 'Research Object Crates and Machine-actionable Data Management Plans', *First Workshop on Data and Research Objects Management for Linked Open Science(DaMaLOS)* at *The 19th International Semantic Web Conference*. (ISWC 2020) DOI: https://doi.org/10.4126/frl01-006423291

Vines, TH, Albert, AYK, Andrew, RL, et al., 2014, 'The Availability of Research Data Declines Rapidly with Article Age', *Current Biology* 24, 94–97. DOI: https://doi.org/10.1016/j.cub.2013.11.014

Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., et al., 2020, ‘Wikidata as a knowledge graph for the life sciences’. *eLife* Mar 17;9, e52614. DOI: https://doi.org/10.7554/eLife.52614

Wilkinson, MD, Dumontier, M, Aalbersberg, IjJ, et al., 2016. ‘The FAIR Guiding Principles for scientific data management and stewardship’. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Williams, M, Bagwell, J, Nahm Zozus, M, 2017, 'Data management plans: the missing perspective'. *Journal of Biomedical Informatics*. 71: 130–142. DOI: https://doi.org/10.1016/j.jbi.2017.05.004

Wittner, R, Mascia, C, Gallo, M. et al., 2022, ‘Lightweight Distributed Provenance Model for Complex Real–World Environments’. *Sci Data* 9, 503 (2022). DOI: https://doi.org/10.1038/s41597-022-01537-6

Wittner, R., Holub, P., Müller, H. et al., 2021, ‘ISO 23494: Biotechnology – Provenance Information Model for Biological Specimen And Data’ In: Glavic, B., Braganholo, V., Koop, D. (eds) Provenance and Annotation of Data and Processes. IPAW IPAW 2020 2021. Lecture Notes in *Computer Science*, vol 12839*. Springer International Publishing*, Cham, pp. 222–225. DOI: https://doi.org/10.1007/978-3-030-80960-7_16

Websites & web pages:

- **ANR.** [WWW Document], n.d., 2019. Available at URL: https://anr.fr/en/latest-news/read/news/the-anr-introduces-a-data-management-plan-for-projects-funded-in-2019-onwards/ (Last accessed 12.14.22)
- **ARIA.** [WWW Document], n.d.. Available at URL: https://instruct-eric.org/help/about-aria (Last accessed 12.15.22)
- **BioImage Archive.** [WWW Document], n.d.. Available at URL: https://www.ebi.ac.uk/bioimage-archive/ (Last accessed 12.14.22)
- **BY-COVID.** [WWW Document], n.d.. Available at URL: https://by-covid.org/ (Last accessed 12.14.22)
- **Canserv.** [WWW Document], n.d.. Available at URL: https://www.canserv.eu/ (Last accessed 12.15.22)
- **CoData.** [WWW Document], n.d.. Available at URL: https://codata.org/ (Last accessed 12.14.22)
- **Cool URIs.** [WWW Document], n.d.. Available at URL: URIs https://www.w3.org/Provider/Style/URI.html.en (Last accessed 12.15.22)
- **CORBEL.** [WWW Document], n.d.. Available at URL: https://www.corbel-project.eu/ (Last accessed 12.14.22)
- **COVID-19 Data Portal.** [WWW Document], n.d.. Available at URL:https://www.covid19dataportal.org/ (Last accessed 12.14.22)
- **COVID-19 workflows.** [WWW Document], n.d.. Available at URL: https://covid19.workflowhub.eu/ (Last accessed 12.14.22)
- **Data Policy.** [WWW Document], n.d.. Available at URL: https://stats.oecd.org/glossary/detail.asp?ID=4454 (Last accessed 12.14.22)
- **DCC_FAIRsharing.** [WWW Document], n.d.. Available at URL: https://dcc.ac.uk/blog/fairsharing-and-dcc-collaborate-align-policy-metadata (Last accessed 12.14.22)
- **DMPOnline.** [WWW Document], n.d.. Available at URL: https://marketplace.eosc-portal.eu/services/dmponline (Last accessed 12.14.22)
- **DSW.** [WWW Document], n.d.. Available at URL: https://ds-wizard.org/ (Last accessed 12.14.22)
- **ECBD.** [WWW Document], n.d.. Available at URL: https://ecbd.eu/ (Last accessed 12.14.22)
- **ELIXIR.** [WWW Document], n.d.. Available at URL https://elixir-europe.org/ (Last accessed 12.14.22)
- **EMPIAR.** [WWW Document], n.d.. Available at URL: https://www.ebi.ac.uk/empiar/ (Last accessed 12.14.22)
- **Ensembl.** [WWW Document], n.d.. Available at URL: https://www.ensembl.org (Last accessed 12.14.22)
- **EOSC.** [WWW Document], n.d.. Available at URL: https://eosc.eu/ (Last accessed 12.14.22)
- **EOSC4Cancer.** [WWW Document], n.d.. Available at URL: https://eosc4cancer.eu/ (Last accessed 12.14.22)
- **EOSC-Life.** [WWW Document], n.d.. Available at URL: https://www.eosc-life.eu/ (Last accessed 12.14.22)
- **ERINHA.** [WWW Document], n.d.. Available at URL: https://www.erinha.eu/ (Last accessed 12.14.22)
- **EU-OPENSCREEN.** [WWW Document], n.d.. Available at URL: https://www.eu-openscreen.eu/ (Last accessed 12.14.22)
- **Euro-BioImaging.** [WWW Document], n.d.. Available at URL: https://www.eurobioimaging.eu/ (Last accessed 12.14.22)

- **EVAg.** [WWW Document], n.d.. Available at URL: https://www.european-virus-archive.com/access-evag-ressources (Last accessed 12.15.22)
- **FAIR Cookbook.** [WWW Document], n.d.. Available at URL: https://faircookbook.elixir-europe.org (Last accessed 12.14.22)
- **FAIRsharing.** [WWW Document], n.d.. Available at URL: https://fairsharing.org/ (Last accessed 12.14.22)
- **FAIRsharing community curation.** [WWW Document], n.d.. Available at URL: https://fairsharing.org/community_curation (Last accessed 12.14.22)
- **FAIRsharing EOSC-Life.** [WWW Document], n.d.. Available at URL: https://fairsharing.org/EOSCLife (Last accessed 12.14.22)
- **FAIRsharing policy record metadata.** [WWW Document], n.d.. Available at URL: https://fairsharing.gitbook.io/fairsharing/additional-information/policy-content-and-scope (Last accessed 12.14.22)
- **GDPR.** [WWW Document], n.d.. Available at URL: https://gdpr-info.eu/ (Last accessed 12.14.22)
- **Go-FAIR.** [WWW Document], n.d.. Available at URL: https://www.go-fair.org/ (Last accessed 12.14.22)
- **HE DMP.** [WWW Document], n.d.. Available at URL: https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf (Last accessed 12.14.22)
- **ICPSR_DMP.** [WWW Document], n.d.. Available at URL: https://www.icpsr.umich.edu/web/pages/datamanagement/dmp/framework.html (Last accessed 12.14.22)
- **IDR.** [WWW Document], n.d.. Available at URL: https://idr.openmicroscopy.org/ (Last accessed 12.14.22)
- **IDTk.** [WWW Document], n.d.. Available at URL: https://www.infectious-diseases-toolkit.org/ (Last accessed 12.14.22)
- **INSTRUCT.** [WWW Document], n.d.. Available at URL: https://instruct-eric.org/ (Last accessed 12.14.22)
- **ISIDORe.** [WWW Document], n.d.. Available at URL: https://isidore-project.eu/ (Last accessed 12.14.22)
- **ISO 20691.** [WWW Document], n.d.. Available at URL: https://fairsharing.org/ISO20691 (Last accessed 12.14.22)
- **IVOA.** [WWW Document], n.d.. Available at URL: https://fairsharing.org/IVOA (Last accessed 12.14.22)
- **interoperability resource within ELIXIR.** [WWW Document], n.d.. Available at URL: https://elixir-europe.org/platforms/interoperability/rirs (Last accessed 12.14.22)
- **Mons, B.** [WWW Document], n.d., 2020. Available at URL: https://www.nature.com/articles/d41586-020-00505-7 (Last accessed 12.14.22)
- **NIH.** [WWW Document], n.d.. Available at URL: https://www.nia.nih.gov/research/blog/2022/09/data-management-and-sharing-nih-policy-details-and-guidance (Last accessed 12.14.22)
- **NSF DMP.** [WWW Document], n.d.. Available at URL: https://www.nsf.gov/bfa/dias/policy/dmp.jsp (Last accessed 12.14.22)
- **OBO Academy's.** [WWW Document], n.d.. Available at URL: https://oboacademy.github.io/obook/tutorial/project-ontology-development/ (Last accessed 12.14.22)
- **OECD.** [WWW Document], n.d.. Available at URL: https://www.oecd.org/digital/ieconomy/enhanced-data-access.htm (Last accessed 12.14.22)
- **OSTP.** [WWW Document], n.d.. Available at URL: https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/ (Last accessed 12.14.22)
- **PDB.** [WWW Document], n.d.. Available at URL: https://www.rcsb.org/ (Last accessed 12.14.22)
- **RDA.** [WWW Document], n.d.. Available at URL: https://www.rd-alliance.org/ (Last accessed 12.14.22)
- **RDA FAIRsharing WG.** [WWW Document], n.d.. Available at URL: https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html (Last accessed 12.14.22)
- **RDAMSC.** [WWW Document], n.d.. Available at URL: https://rdamsc.bath.ac.uk/ (Last accessed 12.14.22)
- **RDMKit.** [WWW Document], n.d.. Available at URL: https://rdmkit.elixir-europe.org/ (Last accessed 12.14.22)
- **RI-VIS.** [WWW Document], n.d.. Available at URL: https://ri-vis.eu/ (Last accessed 12.14.22)
- **RO-Crate.** [WWW Document], n.d.. Available at URL: https://w3id.org/ro/crate (Last accessed 12.14.22)
- **Uniprot.** [WWW Document], n.d.. Available at URL: https://www.uniprot.org/ (Last accessed 12.14.22)
- **WDS.** [WWW Document], n.d.. Available at URL: https://worlddatasystem.org/ (Last accessed 12.14.22)
- **WorkflowHub.** [WWW Document], n.d.. Available at URL: https://workflowhub.eu/ (Last accessed 12.14.22)
- **Zenodo.** [WWW Document], n.d.. Available at URL: https://zenodo.org/ (Last accessed 12.14.22)

Supplementary material

Data management tools

Using general-purpose open repositories operating under the FAIR principles for non-sensitive project outcomes (such as research papers, data sets, research software, reports, slides, posters) can improve the overall project data management. An example is the [Zenodo] multi-disciplinary open repository maintained by CERN. A digital object identifier (DOI) is automatically assigned to all files (and new versions of files) uploaded in Zenodo.

The Research Data Management toolkit for Life Sciences [RDMkit], developed by [ELIXIR] with contributions from other European Biomedical Research Infrastructures, is a community-lead Research Data Management Toolkit for best practices and guidelines to support FAIR policies in data management. The RDMkit provides narrative context and explanations for steps in the data lifecycle, including Data Management Planning; common data tasks, such as data brokering; specialised data management for domains and data types, such as bioimaging and rare diseases; and examples of RDM tool assemblies that support data journeys. The RDMkit provides a signposting gateway to resources such as the FAIRCookbook and the DSW common knowledge model, as well as smart contextual indexing into registries for training materials (the TeSS training portal); tools (Bio.tools); standards (FAIRsharing) and workflows (WorkflowHub). The [EOSC4Cancer] project, in a close collaboration with [CanServ] project will produce a dedicated 'Cancer Data view' in the RDMkit that captures good practices, training resources and connects to data experts across Europe.

The Infectious Diseases Toolkit [IDTk] , developed in the context of the BY-COVID project, provides a convergence point for knowledge exchange on efforts made in research response to infectious diseases. The IDTk aims at collecting and showcasing past and present responses to challenges on data handling, analysis and visualisations, while extracting best practices and guidelines to enhance preparedness. In addition, national resources and experiences can be showcased and shared even when direct access cannot be provided. The IDTk adopts the same ways of working and technical background as the RDMkit. As such, the IDTk is able to connect users with multiple tools and resources, including the RDMkit itself. The IDTk is a key asset open for the ISIDORe partners and users to showcase their data journeys, identify common solutions, and direct to the appropriate resources.

The [FAIR Cookbook] (Rocca-Serra et al., 2022) is an online, open and live resource with recipes that help users to make and keep data FAIR. Developed by data professionals and FAIR experts from ELIXIR, as well as the academic and industry sectors, the FAIR Cookbook guides users through the key steps of a FAIRification journey via recipes, which provide levels and indicators of FAIRness, the maturity model, the technologies, the tools and the standards available, as well as the skills required, and the challenges, to achieve and improve FAIRness. Each recipe tells you the audience type, reading time, level of difficulty, and the level of FAIR maturity it allows you to reach. Recipes are citable via their unique identifier, and their authors are credited, and have cross-references to FAIRsharing (for standards and repositories), RDMkit (for additional reading material), the DSW (to ensure FAIRness by design) and other resources in the ELIXIR ecosystem, and beyond. Supported by ELIXIR Nodes, and strong from the participation of the NIH Office of Data Science Strategy, and major large pharmaceutical companies, the FAIR Cookbook is open to contributions by the ISIDORe community, which can write recipes to showcase community tools and resources, as well as examples of FAIRified data, to help with training, and sharing of common practices.

The EOSC service [WorkflowHub] is a registry of computational workflows, developed by projects including EOSC-Life and BY-COVID that initially targeted Life Sciences and [COVID-19 workflows] (Goble et al., 2021). The workflows and computational notebooks deposited in the hub now cover a wide range of communities including biodiversity and earth sciences. Registries are essential to support FAIR Computational Workflows (Goble et al., 2022), e.g. persistent identifiers, versioning, attribution and detailed metadata. As many ISIDORe data practices involve computational analysis, capturing the implied workflow as a method is important for reproducibility and reuse.

[RO-Crate] is a community-developed method for lightweight packaging of diverse scientific research data with structured metadata, based on FAIR principles and Linked Data standards JSON-LD and schema.org (Soiland-Reyes et al., 2022a). RO-Crate has been integrated with machine-actionable Data Management Plans by using DMPs as templates for building FAIR data packages (Miksa, Jaoua & Arfaoui, 2020), and using exemplar crates to kick-start populating DMPs (Soiland-Reyes et al, 2022b), both approaches now being adopted by ELIXIR's DSW (Eguinoa et al, 2022). RO-Crate has rich support for describing software and workflows, utilised by the ELIXIR Software Management Plans (Giraldo et al, 2022) and EOSC-Life workflow ecosystem including WorkflowHub (Goble et al 2021). ISIDORe can utilise RO-Crate as a way to connect DMPs and data deposits with extensible metadata profiles for different domains, capturing lightweight data derivation provenance with the Common Provenance Model (CPM) (Wittner et al., 2022).

The CPM is a provenance model developed under the auspices of the EOSC-Life project. The main purpose is to support the traceability of data precursors in distributed environments, address reproducibility issues, and enable a more effective data quality and fit-for-purpose assessment. The model is designed for complex, heterogeneous, and multi-institutional environments with different data access restrictions for sensitive data, and subsequently, their provenance. CPM solves the access and reputability challenge by creating multiple provenance bundles, each documenting one environment data and its precursors, such as biological material or environmental specimens, and links these bundles into a common provenance chain. In addition, the model forms a conceptual foundation for an ISO 23494 series (Wittner et al., 2021) (still under development), and is being integrated with the RO-Crate specification within the BY-COVID project (Soiland-Reyes et al., 2022a). Finally, the standard and the underlying model aim for a very high level of provenance interoperability within life sciences domains. Despite its simplicity, usage of the model requires some decisions and design considerations, which may be described in a DMP.

[FAIRsharing] is a community-driven, cross-disciplinary resource, with users and collaborators across all disciplines, that describes community-driven standards, databases, repositories and data policies and the relationships among them. It is also a recommended [interoperability resource within ELIXIR] and EOSC-Life, and an output of the [RDA

FAIRsharing WG]. Plans are to work with ISIDORe to create a branded FAIRsharing Collection that displays live descriptions and network graphs of the resources recommended within their DMP as many have already done, such as [IVOA] , [FAIRsharing EOSC-Life] and the [ISO 20691] specification. Next, we will also ensure that key ISIDORe representatives join the [FAIRsharing community curation] programme to ensure that resources developed and used by ISIDORe are described as richly as possible, providing the ISIDORe community with a well-curated landscape graph of resources relevant to them. Traversal of the ISIDORe graph can aid resource discovery, gap analysis and further targeted outreach and engagement. Additionally, any data management policies created by ISIDORe will be added to FAIRsharing to increase the findability, accessibility and reusability of the policy descriptions through the FAIR data policy 'workflow' described earlier.

Definitions

This section aims to provide definitions of common terms used in the document (the source is Kesisoglou et al., 2022, unless otherwise stated)

Data Management Plan (DMP): A DMP is a structured document that describes data management during and after completion of a research project. It includes information on data capacity, data production, data quality, data safety and protection measures, as well as a role description of people dealing with these tasks. Through recommendations and curation tools, DMPs make data available for reuse in a sustainable way (Bishop et al., 2020). Good research data management prevents data loss, ensures description for appropriate reuse through metadata, keeps data secure and facilitates data sharing. National and international entities have put forward specific requirements to be included in the DMP. In August of 2022, the White House Office of Science and Technology Policy [OSTP] announced that, by 2025, scientific data from all new federally funded research must be made accessible to the US public. Similarly, the European Commission pushes for open access to scientific information, including support for the development of the European Open Science Cloud [EOSC] with major investment from the European Horizon 2020 and following Horizon Europe research and innovation programmes.
[Data policy]: A set of broad, high level principles which form the guiding framework in which Data Management can operate.
Dataset: Collection of data that is represented in a particular form. Datasets will vary depending upon the type of intended use, and how the collecting organisation has decided to organise their data upon collection. Dataset is essentially a heterogeneous term that could be made up of any type of collection for any type of data.
Data steward: A person who has an administrative role; they do not really use the data. They create guidelines to make data FAIR and advice on how to do it. Stewards might have direct responsibility on the data at hand (processors) or not.
Metadata: A set of data that defines and describes a resource (e.g. data, dataset, sample...) so that it can be understood, discovered and reused. There are different levels of metadata. Since metadata can be used to describe different aspects of data, we can group metadata properties in terms of quality, availability, provenance, processing, among others. Then there are metadata catalogues that can be developed to describe the available data collections in a repository or hub. Metadata is important to make data understandable, and can contribute to increase the findability, accessibility, interoperability and reusability of the data. Metadata can be collected or compiled in repositories to improve the FAIRness level of the data collections. '
Personnel data: The stated aim of the GDPR was to enable data subjects to have greater control over their 'personal data', whilst unifying the European data protection rules. The definition of 'personal data', as outlined in Article 4(1) of GDPR, includes 'any information relating to an identified

or identifiable natural person (data subject)'. This includes names, surnames, home address, email address, or an identifier number or data held by a hospital or laboratory that could be used to identify a living individual. In addition, the existence of special categories of personal data, referred to as sensitive personal data, adds another layer of complexity. Sensitive personal data are outlined in Article 9(1) GDPR and include data pertaining to ethnicity, sexual orientation, religious beliefs, trade union membership, and genetic data. Genetic data is defined as 'personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained'.

Quality assurance and control: Regardless of whether a project re-uses / creates new data, or not, their quality assurance and control throughout and beyond the project must be reflected in the DMP. This can be particularly challenging in those consortia with a large number of participants.

Quality ultimately depends on the purpose, so the data flow must be clearly stated in the DMP, indicating the specific objectives of the different working groups and the data use/management in each of them. Quality measures related to each of these steps have to be taken into account in the DMP, without losing the focus on the overarching structure and main goals of the entire project. In addition, not only the quality control procedures must be reflected in the DMP, but also where they are going to be collected. This is important in order to ensure that the data as well as the QA/QC procedures follow the FAIR principles and are available for other projects. (discussion from ISIDORe)

RIs: The European Commission (EC) is defining Research Infrastructures (RIs) as facilities that provide resources and services for research communities to conduct research and foster innovation. They can be used beyond research e.g. for education or public services and they may be single-sited, distributed, or virtual. They often include: i) major scientific equipment or sets of instruments, ii) collections, archives or scientific data, iii) computing systems and communication networks, iv) any other research and innovation infrastructure of a unique nature that is open to external users

Sensitive data: Information that is regulated by law due to possible risk for plants, animals, individuals and/or communities and for public and private organisations. Sensitive personal data include information related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership and data concerning the health or sex life of an individual. These data could be identifiable and potentially cause harm through their disclosure.