Data sharing and reuse under GDPR

Workshop report

Project Title (grant agreement No)	BY-COVID Grant Agreement 101046203
Project Acronym (EC Call)	BY-COVID
WP No & Title	WP6: Engage, train and build capacity with national and international stakeholders
Authors	Simon Saldner, DANS-KNAW/CESSDA Sonja Bezjak, ADP/CESSDA Vasso Kalaitzi, DANS-KNAW/CESSDA Dimitra Kondyli, EKKE/CESSDA Carola Schulz, EMPIRICA Irena Vipavc Brvar, ADP/CESSDA Patricia M. Palagi, SIB Swiss Institute of Bioinformatics
Contributors	Aitana Neves, SIB Swiss Institute of Bioinformatics Mari Kleemola, Finnish Social Science Data Archive
Acknowledgements	The BY-COVID Fest GDPR session speakers: Lilian Mitrou (University of the Aegean), Erdina Ene (BBMRI ERIC), Carola Schulz (Empirica), Sergi Vazquez Maymir (VUB), Irena Vipavc Brvar (ADP/CESSDA), Manolis Terrovitis (Athena Research Center), Aitana Neves (SIB), Mari Kleemola (TAU), and Laura Portell Silva (BSC). The BY-COVID Fest participants.

Funded by
the European Union

History

Date	Version	Who	Description
10/02/2024	0.1	Patricia M. Palagi, SIB Swiss Institute of Bioinformatics; Dimitra Kondyli, EKKE/CESSDA; Irena Vipavc Brvar, ADP/CESSDA; Sonja Bezjak, ADP/CESSDA; Vasso Kalaitzi, DANS-KNAW/CESSDA; Simon Saldner, DANS-KNAW/CESSDA; Carola Schutz, EMPIRICA	First draft
25/04/2024	0.2	Patricia M. Palagi, Dimitra Kondyli, Irena Vipavc Brvar, Sonja Bezjak, Vasso Kalaitzi, Simon Saldner, Carola Schutz, Aitana Neves, Mari Kleemola	Contributions, suggestions
20/05/2024	1.0	Patricia M. Palagi	Revised version

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

Table of contents

List of abbreviations

Abbreviation	**Term**
DMP	Data Management Plan
DPO	Data Protection Officer
EDPB	European Data Protection Board
EHDS	European Health Data Space
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
RDM	Research Data Management

TDR	Trusted Digital Repository
WP	Work package

Executive Summary

The BY-COVID project[1] works towards enabling and improving the accessibility of COVID-19 and other infectious disease data to researchers, policy-makers, and the public. The BY-COVID Fest[2] took place on 23-25 January 2024 in Athens, Greece, as the final event in a series of training events[3] on Research Data Management (RDM) and the General Data Protection Regulation (GDPR).[4] The BY-COVID Fest was organised mainly as a face-to-face event, focussing on knowledge exchange and training regarding data sharing and reuse under the GDPR.

BY-COVID Fest also featured a parallel workshop for the BY-COVID Infectious Diseases Toolkit (IDTk), a knowledge base on infectious disease research-related best practices and solutions.[5] This 'IDTk Contentathon' was aimed to populate and curate the IDTk pages collaboratively.

The BY-COVID Fest started on 23 January, with the co-organisers welcoming the international group of participants with an ice-breaker exercise. This was followed by keynote lectures on *data protection, data privacy, GDPR, and infectious diseases,* and the first workshop on *challenges and issues when dealing with sensitive data*.

The second day was split into two sessions: one dedicated to *Anonymisation and pseudonymisation,* and the other focused on the *FAIRification of sensitive data*. The sessions welcomed lectures, discussions and training sessions from legal and data experts and professionals. Use cases from different disciplinary perspectives were presented, and training on relevant services and tools was provided. The sessions inspired fruitful conversations between the speakers and the event participants, leading to knowledge exchange and highlights deriving from different expertise and experiences. The GDPR track of the BY-COVID Fest concluded with the participants contributing to the IDTK pages on Ethical, Legal and Social issues for all BY-COVID domains.

The present document provides an overview of the organised sessions of the GDPR BY-COVID Fest workshop and their key outcomes.

[1] https://by-covid.org/about
[2] https://by-covid.org/events/by-covid-fest/
[3] https://by-covid.org/events/
[4] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[5] https://www.infectious-diseases-toolkit.org/

1. Introduction

The BY-COVID project aims to provide access to comprehensive, open, and FAIR data and metadata on SARS-CoV-2 and COVID-19, other viruses and diseases. This includes their socio-economic consequences across research fields: from omics, clinical, and epidemiological research to social sciences and humanities. **BY-COVID will accelerate infectious disease research, surveillance, and outbreak investigation**. Since its launch in 2022, the project has brought together 53 partners from 19 countries and 11 Research Infrastructures[6].

BY-COVID is, by design, an interdisciplinary project. Its success and future legacy depend on a thorough understanding of the data provided by partners, adopting best practices and standards to allow data mobilisation and FAIRness, and the best use of the resources and tools provided by the project. Training, capacity building and outreach are essential aspects of this success. They directly contribute to achieving BY-COVID objectives #1 (Enable storage, sharing, access, analysis and processing of research data and other digital research objects from outbreak research) and #5 (Contribute to the Horizon Europe European Open Science Cloud (EOSC) Partnership & European Health Data Space (EHDS), and indirectly the other three objectives[7].

Work Package 6 of the BY-COVID project aimed to engage, train, and build capacity with national and international stakeholders to support the development of an efficient infrastructure for European preparedness for COVID-19 and other infectious diseases. The BY-COVID WP6 partners SIB, EKKE/CESSDA, ADP/CESSDA, and DANS-KNAW/CESSDA co-organised the BY-COVID Fest workshop in Athens, Greece from 23-25 January 2024. BY-COVID Fest had two parallel but strongly connected streams: 1) the **IDTk contentathon** which brought together contributors to the Infectious Disease Toolkit (IDTk[8]) pages, and 2) the training on **Data sharing and reuse under GDPR**. This document reports the main outcomes of the **Data sharing and reuse under GDPR** (**GDPR**) training workshop/stream.

2. Methodology

The **GDPR** track of the BY-COVID Fest workshop was designed to address the critical aspects of data sharing and reuse, particularly in the context of the GDPR. The workshop's primary goal was to foster a more in-depth understanding of GDPR's implications for data sharing and equip participants from all the BY-COVID domains with the knowledge and skills to handle sensitive data responsibly.

---

[6] BBMRI, EATRIS, ECRIN, ELIXIR, EU-OpenScreen, EuroBioImaging, ERINHA, INSTRUCT, MIRRI, CESSDA and PHIRI.

[7] https://by-covid.org/about#objectives

[8] The Infectious Diseases Toolkit (IDTk), developed as part of the BY-COVID project, provides best practices and solutions to data challenges that affect the response to infectious diseases outbreaks. https://www.infectious-diseases-toolkit.org/

The workshop featured a series of lectures from renowned experts who provided insights into the latest developments and best practices in data protection. These sessions were complemented by real-case studies and demonstration sessions, which offered practical perspectives on how GDPR is applied in various scenarios.

Moreover, the workshop included interactive sessions where participants engaged in discussions, sharing their experiences, challenges, and solutions related to **data management**. These collaborative workshops served as a platform for attendees to reflect on their practices, gather new ideas, and learn from the collective knowledge and experiences of the group.

The event was designed to be face-to-face with a limited number of participants to provide a hands-on, interactive experience, the opportunity for questions, and fruitful and intensive knowledge exchange. The cohort of participants included experts with legal, research, and/or data training and support backgrounds from different disciplines, countries, and professional paths/backgrounds. The format aimed to allow for training, knowledge exchange, and interactive collaboration. The participants were treated as experts in their field with the opportunity to contribute equally to the invited speakers.

3. Description of GDPR workshop sessions

3.1 Introduction to GDPR

The Data Sharing and Reuse under GDPR workshop began with an introduction to GDPR, focusing on what GDPR is and how it applies to data related to COVID-19 and other infectious diseases. The session featured lectures from two legal experts on data protection legislation. Keynote speaker **Prof. Lilian Mitrou**[9], University of the Aegean, gave a lecture on how adequate GDPR is for the purposes of **sharing sensitive data**, as well as the big implications of the COVID-19 pandemic for GDPR and privacy legislation. Next, **Erdina Ene**[10]**,** BBMRI-ERIC, gave a lecture on how people think about **privacy in times of crisis**, based on her research on COVID-19 apps. The session closed with a workshop discussion led by the session chairs **Dimitra Kondyli**,[11], EKKES/CESSDA, and **Irena Vipavc Brvar**,[12] ADP/CESSDA.

[9] Lilian Mitrou is a Professor at the University of the Aegean, where she teaches information and data protection laws: https://gr.linkedin.com/in/lilian-mitrou-a7898b2a

[10] Erdina Ene is a Legal Expert and Data Protection Officer at the BBMRI-ERIC, where she provides expertise on privacy and compliance issues: https://www.linkedin.com/in/erdina-cekani

[11] Dimitra Kondyl is Research Director, National Centre for Social Research (EKKE), Athens, Greece. Academia profile: https://ekke.academia.edu/DimitraKondyli

[12] Irena Vipavc Brvar is the Head Of Department at UL, FDV, Social Science Data Archives: https://si.linkedin.com/in/irena-vipavc-brvar

3.1.1 Data protection in pandemic times: is the General Data Protection Regulation (GDPR) adequate to share sensitive data?[13] - Lilian Mitrou

One of the lecture's important questions was whether respecting human freedoms conflicts with monitoring the spread of an infectious disease such as COVID-19. Answering such a question demands a comprehensive understanding of data protection legislation and research requirements. Lilian Mitrou began by explaining how the COVID-19 pandemic posed new challenges in the realm of data protection and privacy, including for GDPR. Before the pandemic, the main issues related to GDPR and data protection mainly revolved around specific technologies and security or law enforcement policies. COVID-19 introduced challenges around how to manage complex data on a worldwide scale, and whether the fight against a health emergency should justify exceptions to data protection and privacy. **The pandemic prompted many to challenge the restrictions imposed by GDPR** and other data protection measures, in the interest of promoting COVID-19 research or preventative measures. Early in the pandemic, the European Data Protection Board (EDPB) stated that, although the fight against infectious diseases should be supported in the best way possible, personal data must be protected even in times of crisis.[14] The challenge, therefore, was on how to put legal foundations in place that allow research and knowledge to advance, while simultaneously safeguarding the right to (information) privacy, as well as other fundamental rights and freedoms.

Mitrou argued that **COVID-19 can be considered a 'test case' for GDPR's ability to support scientific research.** The pandemic required that scientists understand e.g. contagion trends, the effectiveness of distancing measures, and who are those most vulnerable to the virus. Such **health crisis data requirements** included e.g. health data for secondary purposes (different from those they were originally intended for), as well as telecom, location, and patient data. Although such data has great potential to be used for e.g. enforcing social distancing rules, their use involves significant risks to fundamental rights and freedoms.

To understand how compatible GDPR is with such health crisis requirements, Mitrou provided an overview of **'what is data protection?'** Fundamentally, GDPR requires that everyone should be in control of their data. GDPR seeks a balance between the fundamental rights of the individual, the freedom to conduct business, and legitimate knowledge creation. For this reason, **GDPR has a preferential regime aimed at facilitating research**, granting it exemptions to prohibitions against the processing of special categories of personal data, the purpose limitations principle, storage limitation principle, as well as other provisions.

---

[13] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11217609

[14] EDPB (19 March 2020): 'Statement on the processing of personal data in the context of the COVID-19 outbreak.' URL: www.edpb.europa.eu/system/files/2021-03/edpb_statement_art_23gdpr_20200602_en.pdf

Mitrou continued by describing a number of **key terms and concepts**[15] that are important to understand when discussing GDPR:

- GDPR defines certain **key roles** that determine rights and responsibilities towards personal data for different (legal) persons:
    - Controller: whoever determines the means and purposes of personal data processing;
    - Processor: the natural/legal person that processes data on the controller's behalf;
    - Recipient: anyone to whom the data is disclosed, and;
    - Third parties: a natural/legal person who is not the data subject, controller, or processor.
- **Personal data**: can be defined as any information relating to an identified or identifiable natural person. A person who can be directly or indirectly identified is known as the **data subject**.
- **Identifiability**: In European data protection law, the concept of identifiability is key, which simply means that a person can be distinguished from other individuals.
- **Sensitive data**: is any data which by its nature can pose risks to data subjects when processed, e.g.: racial or ethnic identity; political, religious, or philosophical beliefs; union membership; genetic or biometric information; or data about the health of the individual.
- **Health data**: can be any data related to the physical or mental health of a natural person, including conditions, tests, and other information. This does not have to relate to illness: any health-related data is sensitive.
- **Data processing**: this is a comprehensive concept that can essentially be defined as any operation performed on personal data.
- **Sharing**: this can be defined as the '*communication, disclosure or otherwise making available of personal data*' from the researcher to a third party.
- **Processing for the purpose of scientific research**: although GDPR doesn't provide an explicit definition for this, it should be interpreted broadly, to include e.g. technical, fundamental, and applied research, in accordance with sector-specific standards regarding methodology and ethics.

Mitrou proceeded to introduce the **legal basis for processing health-related data for scientific purposes**. The legal basis for processing *personal data* is specified in Article 6 of GDPR[16]: *consent* (a); *legal obligation* (c); *task carried out in the public interest* (e); and *legitimate interest* (f). To process sensitive *health data,* one must satisfy at least one of the following conditions in Article 9 of GDPR for *special categories* of personal data:[17] *explicit consent* (a); *the data is manifestly made public by the data subject (e); necessity for reasons*

---

[15] For more information and definitions of key GDPR terms and concepts, see GDPR article 4: https://gdpr.eu/tag/chapter-4/

[16] For more information, see: gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/

[17] For more information, see: https://gdpr.eu/article-6-how-to-process-personal-data-legally/

of substantial public interest (g); *necessity for reasons of* public interest in the field of public health (i), or; *necessity for reasons of* scientific research purposes (j)). Regarding consent, it is important to distinguish between **informed consent**, which is necessary for participation in scientific research, and **explicit consent**, which is required to legitimise personal data for scientific research purposes. Mitrou emphasised that the EDPB is very critical of the use of consent for research processing purposes. This is because research participants often have 'decisional vulnerability' due to 'informational and power asymmetries' that may not allow them to give free consent according to Article 7 of GDPR. Therefore, other approaches to gaining consent are specific to unavoidable circumstances and can be gained on a case-by-case basis. For example, **broad consent** which mitigates the requirement of specificity of the consent (where the purposes of data processing cannot be specified at the time of data collection), or **data altruism**, which allows for the use of secondary data for prosocial aims such as fighting the COVID-19 pandemic.

Although processing of sensitive data is prohibited by default, **derogations for scientific research can provide another legal basis.** Such derogations can be made when processing is 'necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes,' according to Article 89(1), and when certain conditions are met. These conditions include using proportionate means for the aims pursued, respecting data protection rights of data subjects, and purpose limitation. Conditions are evaluated on a case-by-case basis, and the data must be subject to appropriate safeguards, such as being anonymised whenever possible.

Mitrou continued by describing some principal ways to safeguard sensitive data. **Anonymisation** means that 'all identifying elements are eliminated from a set of personal data so that the data subject is no longer identifiable.'[18] In order to qualify as such, data must pass a **test of identifiability** stipulated by Recital 26 of GDPR. Mitrou emphasised that anonymisation is often difficult to achieve, due to the risk of re-identification, and the risk of losing the utility of the data. **Pseudonymisation** means replacing sensitive attributes of data with pseudonyms (e.g. codes), such that the personal data can no longer be tied to a data subject without additional information, which are stored separately. Other important safeguards that apply to data controllers are **data minimization** (using the least intrusive data collection solutions for the specific purpose); **transparency and information duties** (sharing as much information as possible about the data processing activities with the data subjects), and compliance with “procedural/functional” obligations such as the **Data Protection Impact Assessment.**

Ending the presentation, Mitrou shared some reflections on GDPR's role as enabler and barrier for research, particularly in times of crisis such as during the COVID-19 pandemic. She emphasised that the **protection of personal data is not an absolute right,** and that this must be balanced against other fundamental rights and the public interest according to the **principle of proportionality**. In this way, **GDPR is in principle able to safeguard**

[18] For more details on (pseudo)anonymisation, see section 3.2 of this report.

personal health data while permitting COVID-19 research. There are however limitations to GDPR, one being that it is often difficult to understand its implication, even by researchers. This is further complicated by the fact that GDPR implementation differs in different EU countries. Emergencies such as COVID-19 also necessitate specific requirements for democratic states when it comes to personal data processing. This includes the application of clear, precise and accessible rules; necessity and proportionality of measures; transparency; as well as independent oversight mechanisms. Such requirements effectively constitute a 'democracy test' for states in emergencies such as COVID-19.

3.1.2 Nationwide technology solutions to COVID-19 and their validity in front of the GDPR: The case of contact tracing applications[19] - Erdina Ene

Erdina Ene used the example of **contact tracing apps**[20] to demonstrate how technological solutions can cause tensions with privacy and data protection concerns. This was based on the research results from a study she performed in 2021[21] on such apps, and how people think about their privacy in times of crisis. These mobile phone apps were widely used to trace contact with people infected with COVID-19 on an individual level. They would warn users if they had been physically near infected people. Such apps were widely used during the pandemic and were promoted, or sometimes even mandated, for their effectiveness in preventing the spread of COVID-19. However, the data collected by such apps raised serious ethical, privacy, and data protection concerns. This presentation therefore explored the privacy implications of such technological solutions in times of crisis, and their compatibility with GDPR, by using two prominent examples of such apps: the British 'NHS COVID-19' app, and the Italian 'IMMUNI' app.[22]

Ene started by giving a historical overview of **the right to privacy.** Although discussions of privacy date back to Aristotle, and although privacy has gained growing public awareness in modern times, the earliest academic publications on the topic did not appear until 1890.[23] Ene pointed out, therefore, that the instinct to protect privacy is not written in our DNA but has been learned throughout modernisation. Various factors contribute to this, prominent among them being that our privacy has never been so exposed as it is today, and contact tracing apps are potentially a prime example of this.

---

[19] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11102997

[20] For more information about contact tracing apps and their use, see: https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1905

[21] Ene 2021, Shielding the Digital Treasure: A Dyad of Economy Giants in their Quest to (for)get your Health Data. URL: https://www.duo.uio.no/handle/10852/89400

[22] Although 'UK GDPR' is domestic legislation which implements GDPR, Erdina Ene emphasised that the data protection principles of the UK GDPR do not differ from those of the EU GDPR.

[23] 'The right to privacy,' written by Samuel D. Warren II and Louis Brandeis, and published in the 1890 Harvard Law Review.

Ene outlined **four key considerations for understanding the implications of tracing apps on privacy**:

- **Communication protocols**: these determine how users' phones communicate with each other and have important implications for where and how data storage takes place. Communication protocols can either be centralised or decentralised, and both use various 'arbitrary identifiers' to identify and trace users and their phones. In centralised systems, identifiers are transferred to, and analysed in, centralised servers that are typically found at the public health authority of the country in question. In decentralised systems, such identifiers are stored and analysed locally on the phone in question.
- **Lawfulness, fairness & transparency**: these are GDPR's core data protection principles. Lawfulness means that a legitimate public interest justifies the processing of data; fairness, although difficult to define, is interpreted by the The European Court of Human Rights (ECHR) as '*processing personal data in a way that complies with the data subject's expectations';*[24] and finally, transparency means that the information should be easily accessible, understandable, and available in multiple languages.
- **Data minimization**: is the principle of not collecting more data than strictly necessary. In the case of tracing apps, this can mean not collecting data about more symptoms than is necessary to determine that someone may have contracted COVID-19.
- **Storage limitations**: that data is stored no longer than necessary. For tracing apps, this means storing data only so long as it is necessary for health protection purposes.

Ene compared the different ways that the NHS COVID-19 and IMMUNI apps navigated these four factors. For example, regarding storage limitations, the NHS app only kept data for 24 - 48 hours, while IMMUNI specified an upper data retention limit at a fixed date in the future. Regarding transparency, the NHS app was offered in 12 languages, while the IMMUNI app was only available in 5, and that having such data may inadvertently reveal sensitive data about the ethnic or racial profile of their users. Regarding data minimization, Ene highlighted how using such apps involves collecting data that may not be strictly health-related but is necessary for maintaining their technical functioning. This may include gathering data about the user's mobile device and location and whether or not the user is infected.

Such examples illustrate how differently privacy considerations can be interpreted, and how concerns about public health protection and privacy issues can come into tension with

[24] Ene 2021, Shielding the Digital Treasure: A Dyad of Economy Giants in their Quest to (for)get your Health Data, p.12.: https://www.duo.uio.no/handle/10852/89400

each other. Ene emphasised the need to find **balancing of public health and privacy concerns** by quoting the European Data Protection Supervisor, Wojciech Wiewiórowski:[25]

> *'Humanity does not need to commit to a trade-off between privacy and data protection from one side, and public health, on the other. Democracies in the age of COVID-19 must and can have them both.'*

3.1.3 Workshop: challenges and issues when dealing with sensitive data

Following the lectures from Mitrou and Ene, the session chairs **Dimitra Kondyli**,[26], EKKE/CESSDA, and **Irena Vipavc Brvar**,[27] ADP/CESSDA, led an open and lively discussion with speakers and participants. This discussion was driven by the questions raised by the participants on their challenges and issues when dealing with sensitive data. Some of these questions were:

- What is the difference between anonymisation, de-identification and pseudonymisation?
- How often do researchers encounter a request for access rights?
- Should apps be seen as medical devices? Would this fall under other legislation?
- Under which conditions non-anonymised data can be kept in the long-term?
- We are keeping very sensitive data which will probably one day be shifted to classical archives? Is this ok?
- Data protection in the light of advances in AI and more and more data? What can we expect in this field?
- How to deal with incidental findings that would be of benefit to the patient but we are handling anonymised data and cannot access the patient?
- Challenges of longitudinal, cohort studies in respect to GDPR?
- We have datasets that include sensitive data collected 50 years ago, they are anonymized, can we preserve them?
- We are dealing with a lot of historical data, researchers conducting interviews, anonymization is no go, it's not a case?

Some lessons learnt from this whole session:

- When processing personal data, one must comply with the European Data Protection Regulation, while being aware of and complying with the national law(s) that may impose additional restrictions.
- When determining the regime of access to sensitive data, one must comply with the law and preserve the trust and confidentiality of the participants.

---

[25] Bertelsmann Stiftung and Algorithm Watch. Automated Decision-Making Systems in the COVID-19 Pandemic: A European Perspective. (1 September 2020)

[26] Dimitra Kondyl is Research Director, National Centre for Social Research (EKKE), Athens, Greece. Academia profile: https://ekke.academia.edu/DimitraKondyli

[27] Irena Vipavc Brvar is the Head Of Department at UL, FDV, Social Science Data Archives. LinkedIn: https://si.linkedin.com/in/irena-vipavc-brvar

- Where there is (domain) specific legislation, e.g. health data, it is worth involving specialised services and experts, such as Data repositories, Data Protection Officers, Data Specialists etc.
- Protocols and defined work processes will facilitate work and contribute to a higher quality of research work, e.g. research data management planning.

3.2 Anonymisation and Pseudonymisation

The session on Anonymisation and Pseudonymisation started with a lecture by **Carola Schulz**,[28] Empirica, contextualising the place of anonymisation and pseudonymisation in GDPR, ISO and EHDS Regulation Proposal[29]. It continued with a talk from **Sergi Vazquez Maymir**,[30] VUB, on de-personalising health data from a complex brain disorders use case, and a hands-on exercise led by **Irena Vipavc Brvar**,[31] ADP/CESDA, where the workshop participants had to identify data in social sciences records that should be anonymised. The session closed with a demonstration of the Amnesia Anonymization Tool by **Manolis Terrovitis**,[32] Athena Research Center.

3.2.1 Setting the scene: Anonymisation and Pseudonymisation in GDPR, ISO and EHDS Regulation[33] - Carola Schulz

Carola Schulz opened the session by explaining the importance of defining anonymisation and pseudonymisation concepts in regulations on data protection and health data and some of the standards used in digital health, because health data is a primary focus of BY-COVID.

Studying **definitions** related to the topic reveals that the GDPR does not have a definition of **anonymisation** per se, only a definition of anonymous information (GDPR, Recital 26). However, it does have a definition of **Pseudonymisation** (Article 4 (5), Chapter 1). Those interested in an explicit definition of anonymisation can find one in ISO 25237:2017.

GDPR does not apply to anonymous data. The Regulation also describes pseudonymisation as a method of data protection, but sets out that it should be accompanied by other

[28] Carola Schulz is Senior Research Consultant and Project Manager at Empirica Communication and Technology Research: https://www.linkedin.com/in/carolaschulz/
[29] This report and the respective presentation consider the EHDS Regulation proposal as of January 2024. It does not consider the changes in the final version, approved on 24 April 2024.
[30] Sergi Vazquez Maymir is a researcher at the Vrije Universiteit Brussel and a project member of the MES-CoBraD project: https://www.linkedin.com/in/sergi-vazquez-maymir-ab46b434/
[31] Irena Vipavc Brvar is the Head Of Department at UL, FDV, Social Science Data Archives. LinkedIn: https://si.linkedin.com/in/irena-vipavc-brvar
[32] Manolis Terrovitis is a Principal Researcher at the Information Systems Management Institute of the Research Center Athena in Athens, Greece, where his work centres on Privacy Preservation and Big Data management: https://gr.linkedin.com/in/manolis-terrovitis-06666b1
[33] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11092756

measures (Recital 28). It also outlines that pseudonymised personal data may be used for research purposes (Recital 156).

The European Health Data Space (EHDS) Regulation Proposal[34] adopts the GDPR definitions of anonymisation and pseudonymisation. These topics are only relevant in the context of secondary use of health data, which is managed by Health Data Access Bodies. These bodies should apply anonymisation and pseudonymisation, among other techniques to preserve privacy (Recital 43). By default, **Health Data Access Bodies should only make data available in anonymised format** - and only exceptionally in pseudonymised format (Recital 49; Recital 50). However, the regulation proposal does not explicitly clarify who should be in charge of the anonymisation/ pseudonymisation of the data, though it appears to be the Health Data Access Body itself. The EHDS regulation proposal also acknowledges that anonymisation does not totally eliminate the risk of re-identification of data subjects, especially for Electronic Health Records, disease registries, biobanks, and person-generated data – which all imply broad identification characteristics. The risk of re-identification is bigger for data from small geographical areas, especially when one considers that new future methods can be available through the evolution of technology and the combination with other data sources (Recital 64). The EHDS regulation proposal also recognises that this risk could endanger the acceptance of secondary use of health data (Recital 64).

Schulz continued by mentioning more recent **ISO standards** that offer definitions of anonymisation and related concepts such as ISO/IEC 27559:2022; ISO/TS 17975:2022(en). She also mentioned recent research on the prevalence of anonymisation and pseudonymisation techniques in Health Data Access Bodies[35] and health-related data infrastructures[36] stating that most of these entities declare to use pseudonymisation.

Schulz summarised the following points for further thought:

1. Anonymisation is often associated with GDPR compliance – however, the GDPR does not define anonymisation per se.
2. Complete anonymisation, with zero risk of (future) re-identification is de facto impossible - even EHDS Regulation Proposal acknowledges this. The danger lies in the possible linkage with various data sources. Health data is more challenging, because it is very personal and diverse in type.
3. Currently, most of the shared data is found in pseudonymised format.

---

[34] https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en

[35] Landscape analysis using a health-related data catalogue matrix (HealthyCloud) (2023) – S. Cosgrove, I. Kesisoglu, P. Derycke (Sciensano). https://zenodo.org/records/10226557

[36] HealthData@EU Pilot identifies common elements for health data access and data use within the legal frameworks of the participating nodes (HealthData@EU Pilot) (2023). Available at: Website publication_2023_WP7 landscape analysis_rev (ehds2pilot.eu)

3.2.2 A DPO's account on health data de-personalisation: the MES-CoBraD case[37] - Sergi Vazquez Maymir

Sergi Vazquez Maymir gave an overview of the Multidisciplinary Expert System for the Assessment & Management of Complex Brain Disorders platform (**MES-CoBraD**), which was developed in the context of an EU-funded project that aims to improve the diagnostic accuracy and therapeutic outcomes of complex brain disorders, such as epilepsy, dementia, and sleep disorders. The MES-CoBraD system architecture is built on two main components: an Edge Module and Cloud Services. The "Edge Module'' is the one that collects and anonymises data from data holding institutions and sends it to a cloud where the analysis will happen.

The project's real-world data is to be anonymised at the point of collection, at the institutions holding the data. It was challenging to make this happen from a technical and legal point of view because close to 600 data sets had to be collected from 4 medical institutions involved in the project.

Before the data can be sent to the platform, it needs to be harmonised. It is thus a **de-personalisation tool that prepares data for upload to the MES-CoBraD platform**. The data is subsequently uploaded into a data lake in the Cloud Services, which has an integrated ability for analytical queries. Vazquez Maymir stated that the output of this module is not necessarily anonymised data - which is why it was called "de-personalisation module", and it can encompass both anonymised and pseudonymised data.

The data protection methodology in MES-CoBraD mapped key data protection aspects to a series of questions. A data management plan was used as a tool for this, jointly with data processing agreements and others. As a general rule, data can be processed in the project cloud if it is not re-identifiable anymore - in which case the GDPR would not apply to it. However, it turned out to be a challenge to determine if data is anonymised or not - and what constituted personal data in the first place.

Vazquez Maymir then elaborated on the **legal semantics of personal data, pseudonymisation and anonymisation** in the GDPR. He highlighted that, unlike with anonymised data, pseudonymised data can still be linked to data subjects and is thus still considered personal data. EU courts also interpret this concept quite broadly, ensuring widest data protection.

For the work in MES-CoBraD, this means that the decision whether data is personal or not depends whether it (possibly) identifies a **natural person**, when combined with other information. Also, data might be pseudonymous for the controller (since they possess additional information for re-identification), but anonymous for the recipient. To evaluate

[37] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11220535

the risk of re-identification it needs to be checked whether this would be possible using reasonable measures.

This led to a re-interpretation of a de-personalisation module: GDPR might still apply to the controller, but not to the recipient. Thus, this question needs to be clarified individually for each dataset.

In the following question and answer session, one participant highlighted the risk of re-identification of "anonymised" MRI images. Vazquez Maymir suggested that in this case, it makes sense to agree on a **threshold for anonymisation** and judge if each case falls above or below it. Another participant mentioned that in case of doubt, it is always safer to assume that data is pseudonymous rather than anonymous - to which Vazquez Maymir agreed.

3.2.3 Hands on social sciences data anonymisation[38] - Irena Vipavc Brvar

Irena Vipavc Brvar started the session by pointing out that data anonymisation in social sciences is situated between **Ethics (what we should do)** and **Legal (what we must do)**. These aspects affect all stages of the research data life cycle.

Irena repeated the definition of (sensitive) personal data. As a particular challenge of social sciences, she mentioned that participants in social science research, especially in interviews, often reveal data on third persons - on behalf of others - who have not given **explicit consent** to data collection.

Workshop participants then worked on a first exercise[39]. They identified information to be anonymised in a transcript of an interview on Learning at Swiss Elementary Schools. In the sample solution, Vipavc shared possible **information substitutes** for elements like location, school, name and biographic information. She highlighted that, depending on the research purpose, anonymisation of certain aspects could hamper the analysis.

In a second example, participants proceeded in a similar manner with an exercise on Managing Suffering at the End of Life[40]. Since this example contains much more detailed personal as well as health information and refers to one single individual and their immediate family, the sample solution for anonymising this content is much more complex. Discussion acknowledged challenges with anonymisation protocols in international studies, since one element, e.g. type of health organisation, might be more revealing in some than in other countries.

The session was wrapped up by Vipavc reminding the group of the GDPR principles. These are 1) Lawfulness, fairness and transparency; 2) Purpose limitation; 3) Data minimization;

[38] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11123132
[39] Available at: https://shorturl.at/dntGN
[40] Available at: https://shorturl.at/quHRY

4) Accuracy; 5) Storage limitations; 6) Integrity and confidentiality; and 7) Accountability. **These 7 principles of GDPR**[41] are a quick guide for researchers on how personal data should be handled. One should remember to collect information which is necessary for one's research and delete or anonymise it when it is not needed anymore.

3.2.4 Amnesia Anonymization Tool - Data anonymization made easy (openaire.eu)[42] - Manolis Terrovitis

Manolis Terrovitis presented the **anonymisation tool Amnesia**[43], which researchers can use to remove identifying information, while preserving most of the statistically interesting part. Like previous speakers, he also mentioned a general confusion of the concepts of anonymisation and pseudonymisation.

He proceeded to outline the different purposes of anonymisation and encryption. **Encryption** is used to protect data from unauthorised third parties, but not from the recipient. **Anonymisation** is used when the recipient should not get access to the personal data. Researchers mainly anonymise data for publication or because the data subjects have not consented to data sharing. Many scenarios might need both encryption and anonymisation.

Terrovitis then explained that there is no guarantee for total (future) anonymisation, citing cases of **linkage attacks** via quasi identifiers. Since any personal data can potentially be linked to other data, in case of doubt, data should be considered pseudonymous rather than anonymous.

He then introduced the concept of **k-anonymity**,[44] which is the basis of the Amnesia anonymisation algorithm. Using this method, it is unavoidable to lose some information. Terrovitis conceded that choosing the value of the "k" might be challenging and that even after anonymising with this method, it might still be possible to infer if a certain data subject is a member of the data set. It might also be challenging if institutions only have very small datasets.

He then proceeded to demonstrate how **Amnesia lets the user anonymise data in five steps**: 1) Data upload; 2) Creation of generalisation hierarchies; 3) Anonymisation processing; 4) Choice of output; 5) Finalisation. He stressed that the tool also works locally, without any online connection, so that the data remains with the controller. Terrovitis emphasised that when choosing the anonymisation parameters, the user should always keep in mind the kind of analysis intended. Finally, Terrovitis clarified the difference between anonymised and synthetic data, highlighting the advantages and drawbacks of both for research.

---

[41] https://gdpr-info.eu/art-5-gdpr/
[42] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11217653
[43] https://amnesia.openaire.eu/
[44] https://en.wikipedia.org/wiki/K-anonymity

In the following question and answer session, one participant asked what are typical use cases for Amnesia. Terrovitis replied that they span several areas and institutions. Clinical researchers typically only use it when sharing data with third parties. There were also several questions on how to improve the performance of Amnesia and manage large amounts of variables.

3.3 How to make your sensitive data FAIR, challenges and use-cases session

The third and final part of the workshop approached the topic of how sensitive data can be made FAIR[45], that is Findable, Accessible, Interoperable, and Reusable. Given the particular privacy and data protection considerations that apply to sensitive data, this poses certain challenges when trying to make such data FAIR. The session featured three lectures which presented experiences and use-cases for how to navigate and overcome challenges around FAIRifying sensitive data in the life & health sciences by **Aitana Neves**,[46] SIB, and in the social sciences by **Mari Kleemola**,[47] FSD/CESSDA; as well as an overview of tools that can be used to **FAIRify** sensitive data by **Laura Portell Silva**,[48] CSC. Finally, the three speakers took part in an interactive panel discussion chaired by **Vasso Kalaitzi**.[49]

3.3.1 Use-case in life and health sciences[50] - Aitana Neves

Aitana Neves began her presentation by introducing the Swiss Pathogen Surveillance Platform (SPSP),[51] a One Health data platform to support surveillance of pathogens, co-developed by the SIB. SPSP is an online platform that collects pathogen sequencing data and their associated clinical and epidemiological metadata from various Swiss health institutes. The platform then curates and processes this (meta)data, before rapidly sharing it with external stakeholders at the European and international level, including with BY-COVID's COVID-19 Data Portal,[52] as well as the Swiss federal government. Data sharing through the SPSP has been so effective that it has made Switzerland the 3rd largest contributor of open SARS-CoV-2 data in the world, despite its relatively small population.

---

[45] For a brief overview of the FAIR Principles, see: https://force11.org/info/the-fair-data-principles/

[46] Aitana Neves is the Associate Director of Clinical Bioinformatics at SIB Swiss Institute of Bioinformatics in Geneva, Switzerland: https://www.linkedin.com/in/aitananeves/

[47] Mari Kleemola is a Development Manager at the Finnish Social Science Data Archive (FSD), in Tampere, Finland. Kleemola participates in the BY-COVID as a member of CESSDA: https://www.linkedin.com/in/mari-kleemola-31098a3/

[48] Laura Portell Silva is a Research Engineer at the Spanish National Bioinformatics Institute (INB) of the Barcelona Supercomputing Center (BSC), and a BY-COVID Project member: https://www.linkedin.com/in/laura-portell-silva/

[49] Vasso Kalaitzi is Senior project manager and acquisitor at DANS (Data Archiving and Networked Services) in The Hague, Netherlands, and contributes to the BY-COVID project as a member of CESSDA: https://nl.linkedin.com/in/vasso-kalaitzi-a8a93722

[50] The presentation slides can be found at: https://zenodo.org/doi/10.5281/zenodo.11066608

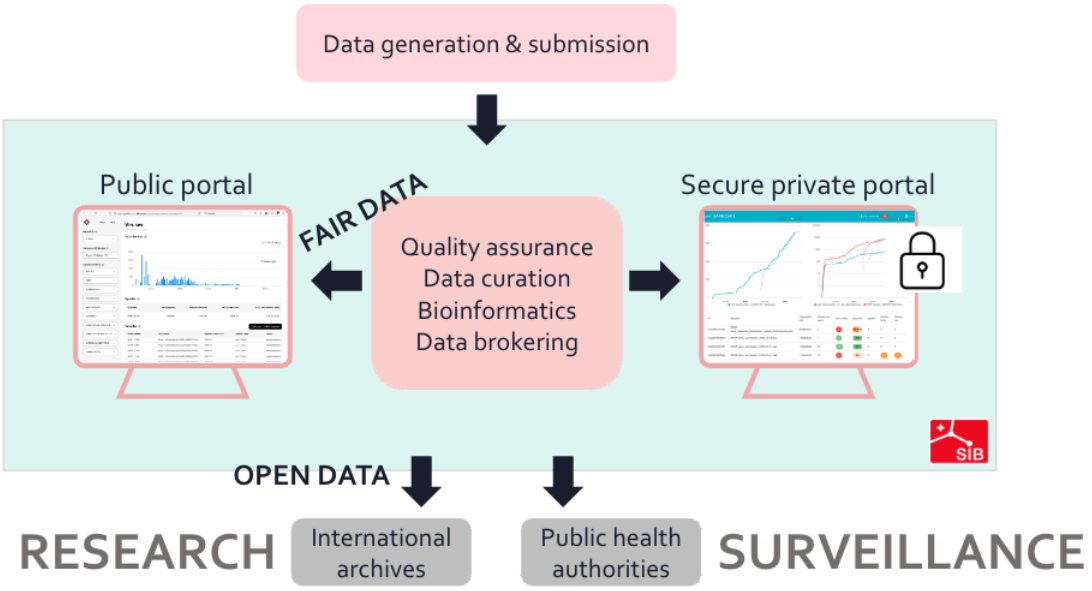
[51] https://spsp.ch/; https://doi.org/10.1099/mgen.0.001001

[52] www.covid19dataportal.org

However, SPSP was mostly limited to sharing what was considered non-sensitive data, such as raw pathogen data, genomes, and minimal contextual data.

Moving beyond non-sensitive data, Neves continued by describing two data portals that SPSP has to deal with sensitive data: a secure private portal where users can access their data or the data they have been given access to, and a public portal where FAIR data can be shared openly (Figure 1). SPSP is hosted on the secure IT infrastructure BioMedIT, based in Lausanne. Running a secure web platform that can also continuously communicate and exchange data with external sources requires a complex IT infrastructure that has matured over several years of development. The guiding principle in this development has been to make the infrastructure itself **FAIR**, with the aim that this will, in turn, produce FAIR data. Providing timely and FAIR data is an integral part of SPSP's mission to support epidemiological research, surveillance, and preparedness & response.

Figure 1. SPSP data platform conceptual model[53]



Describing how SPSP complies with each of the FAIR principles, Neves started by detailing how data is made **Findable**. This is achieved by providing rich, globally unique, searchable, and ENA (The European Nucleotide Archive) compulsory metadata with persistent identifiers.[54] Data is made **Accessible** by using a standardised and open communications protocol, where sensitive data is discoverable, but only accessible through an authentication and authorisation procedure, for a minimum of 10 years. **Interoperability** is ensured by making all (meta)data structured, and by assigning an ontology to each submission field, and by including qualified references to other (meta)data through

[53] Adapted from Aitana Neves' presentation referred to above.

[54] For the latest metadata template used by SPSP, see: public.spsp.sib.swiss/docs/instructions-for-new-registered-groups.html#get-the-metadata-template-file

International Nucleotide Sequence Database Collaboration (INSDC) identifiers, which make it possible to link data with data from other datasets. Finally, data is made **Reusable** by providing rich and well-described (meta)data with multiple relevant attributes, and by providing a clear ethical and legal framework for defining data access and reuse.[55] Detailed data provenance is also provided using the Nextflow workflow manager as much as possible, and efforts are made to comply with domain-relevant community standards through participation in international initiatives such as ELIXIR, PHA4GE, and GMI.

A key consideration throughout data processing is ensuring (meta)data quality. At SPSP, this is done on the server side (as opposed to the client side), by using parsers that automatically check the data (e.g. that it complies with required formats & vocabularies, and that there are no missing files). Data is returned to the user for resubmission if necessary. Neves emphasised that the effort required for such critical **user support and data curation** is often underestimated: for SPSP, this requires 0.6 FTE/year to process roughly 50,000 samples.

Finally, Neves ended by highlighting some next steps for SPSP. This included the need to **make metadata more machine-findable and accessible**, and by making metadata field ontologies available through APIs (Application Programming Interfaces) on the Portal. SPSP is also working on further developing the Maturity Model for Pathogen Data Platforms,[56] to ensure that FAIR principles are embedded and applicable to infectious diseases data platforms, including within BY-COVID.

During the Q&A session that followed her talk, Neves was asked **how the SPSP approach could be replicated in other countries**, or at the EU level. Neves responded by saying that this would require overcoming challenges at multiple levels: firstly, the ethical and legal framework would have to be adapted to country-specific conditions, an important process which can take over 18 months to accomplish. From a technical perspective, although some code can be shared and reused, not all code can be made open for security reasons. The code used to develop the SPSP was also heavily tailored to fit the particular needs of SPSP and may therefore not be suitable to others.

3.3.2 Use-case in Social Sciences[57] - Mari Kleemola

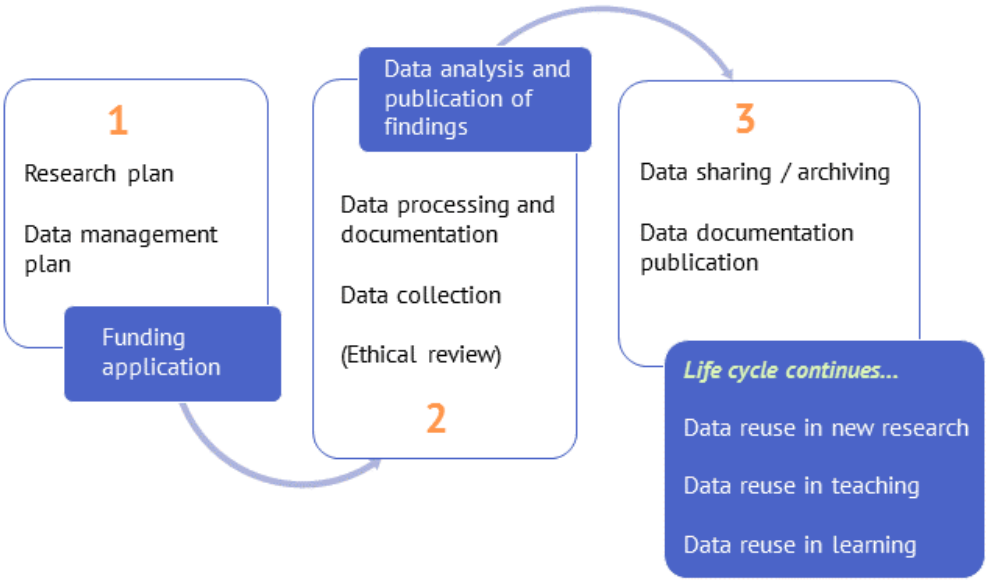
Mari Kleemola started by describing what sets social science data apart from other types of data. Depending on the research questions, social science research data typically consists of e.g.: questionnaire surveys, interviews, focus group discussions, written material, recordings, official documents, archival material, websites, register data or social media data. Examples of quantitative social science data include survey data, while qualitative data can include interviews or audiovisual data. This presentation focused on data

[55] For details on access & reuse, see: https://public.spsp.sib.swiss/docs/data-access-and-re-use.html
[56] elixir-europe.github.io/pdp-maturity-model/
[57] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11066716

collected through surveys and interviews. Most social science datasets are small yet complex, and are produced by a small team or even a single researcher. **Sensitive data is common in the social sciences**, since researchers often collect personal data from individuals. Obtaining informed consent is therefore crucial. Although social science researchers have become much more willing to share data compared to 25 years ago, GDPR is regularly used as an excuse not to share sensitive data.

Kleemola proceeded to provide an overview of the FAIR principles, and how these apply to social science data. Firstly, Kleemola emphasised that **making (meta)data FAIR starts with good data management**. This involves ensuring that (meta)data is organised and preserved in a way that ensures that data remains "accessible, understandable and reliable." Data protection and security must be ensured throughout the entire data life cycle (Figure 2).

Figure 2. Data life cycle[58]



When it comes to the sharing of sensitive personal data, there is frequently a perceived tension between data sharing and data protection. However, **even sensitive data can be shared**. By a combination of gaining consent, anonymising data (see section 3.2 in this report for details), and providing clarity regarding data copyright and access conditions, ethical and legal sharing of sensitive data is possible. Informed consent, which is often required in social science research, should also include information about the storage, archiving and reuse of this data. It is also important to recognise that GDPR does not apply to anonymised data. However, researchers are often reluctant to anonymise their data, in part because it is often a resource-intensive process.

[58] Adapted from Mari Kleemola's presentation at the BY-COVID Fest. Image source: www.fsd.tuni.fi/en/services/data-management-guidelines/why-are-research-data-managed-and-reused/

Kleemola emphasised that it is important to recognise that **FAIR data does not imply open and free data**. Restricted data can still be entirely FAIR, provided that the data is findable, and the terms of access and re-use are clear (how this is defined varies by discipline and research community). Therefore, sensitive data can be made FAIR by making its metadata FAIR, and the more metadata is available, the better for potential re-users. Metadata should be made available even if the data is restricted. In fact, **metadata is essential for restricted data**, making it easier for users to identify whether the data is suitable for their purposes without needing access to the data itself. Besides bibliographical information about data creators, general metadata is anonymous, and variable-level metadata typically poses no security risks in social science research.

Both metadata and data should be made as FAIR as possible. However, this can be a resource-intensive process, particularly in smaller research teams without a dedicated data steward, as is often the case in the social sciences. Multiple FAIR assessment tools exist that can test the FAIRness of metadata. However, such tools usually only assess the metadata, not the actual data. In addition, it is often difficult to know when (meta)data is 'FAIR enough.' Nonetheless, **FAIR assessment tools like F-UJI**[59] can be useful in improving the FAIRness and data management of data. To use F-UJI, you simply enter the DOI or URL of the data, and the tool will return an assessment of FAIRness, including social science metrics.

Kleemola continued to emphasise that you should always think about '**FAIR + time**', meaning that while making your (meta)data FAIR now is essential, securing that data for the long term is equally important. As Kleemola summarised, 'a database will not stay FAIR unless someone takes care of it.' **Sharing your data via a Trustworthy Digital Repository (TDR)** is an excellent way to ensure this: TDRs can 'make the researcher's life easier' by assisting with data curation; making and keeping (meta)data FAIR; enabling metadata harvesting/sharing; anonymisation; and periodic *residual risk assessment* of anonymised data. This usually means sharing the metadata rather than the data itself for sensitive data. For social sciences data in Europe, choosing a local CESSDA-affiliated[60] and/or CoreTrustSeal-certified[61] TDR is a reliable choice. These TDRs can ensure that your data fulfils all the 'FAIR metadata basics,' including providing a persistent identifier (PID); using a common (domain-specific) metadata standard (e.g. Dublin Core, or DDI for the social sciences); providing licence conditions; making it machine-readable; and including provenance information.

Following the presentation, Kleemola responded to a question on **why more researchers do not make their data FAIR**. Kleemola responded that the main reason is usually a lack of resources, that researchers underestimate the time and effort required to make their data FAIR, or that they do so too late in the data creation process. Regarding informed consent,

[59] https://www.f-uji.net/
[60] https://www.cessda.eu/About/Consortium-and-Partners/List-of-Service-Providers
[61] https://amt.coretrustseal.org/certificates

another workshop participant asked **how consent forms can be preserved** to prove that consent was obtained. Kleemola responded that, in her organisation's repository, they typically do not ask for consent forms since this is not required for anonymised data. If a researcher wants to deposit non-anonymised data, they would ask the researcher to maintain the consent forms themselves. Finally, a participant asked about the **most essential capacity-building measures for making data FAIR**, particularly for health data holders (a.k.a. 'Health Data Access Bodies'[62]). Kleemola suggested that the most important consideration is that metadata is made machine-actionable, mainly to ensure that holders of health data and e.g. of social science data can connect effectively, and avoid creating silos, as frequently happens.

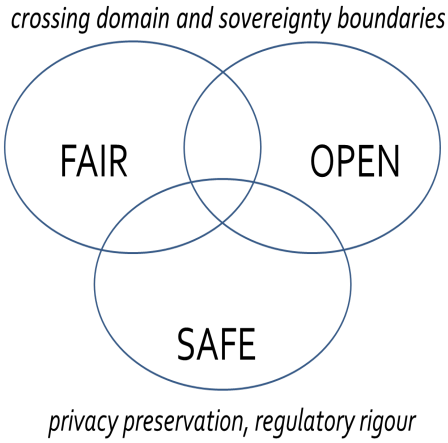
3.3.3 ELIXIR tools to make your sensitive data FAIR[63] - Laura Portell Silva

Laura Portell Silva began her lecture by introducing the principles of Research Data Management (RDM), which the tools described in the presentation should help improve. There are **two main principles of RDM**: 1) it involves the whole data lifecycle, and; 2) it complies with FAIR principles. The data lifecycle revolves around planning, data collection, data processing, analysis, preservation, sharing, and reuse (see also Figure 2).

When it comes to the FAIR principles, Portell Silva focused on two aspects that are particularly significant for sensitive data, the first of which is **Access**. As Kleemola explained in the previous lecture, FAIR is not the same as open (see Figure 3). Therefore, you may need to store sensitive data by using a trusted, **controlled access repository**. These repositories can determine the access conditions to the data, including demanding authentication where necessary, and can be either generalist and subject-specific. A good repository should always adhere to the FAIR principles; be indexable by search engines; be accessible using standard, open, and free machine-accessible protocols; reliable, and be suitable for your research domain.

[62] https://www.european-health-data-space.com/
[63] The presentation slides can be found at: https://doi.org/10.5281/zenodo.11191575

Figure 3: FAIRness, Openness, and Safety[64]



It may however be difficult to identify a repository that suits the data and discipline-specific requirements in question. To identify suitable repositories, several useful tools are available: **FAIRsharing.org**[65] has several data resources tools that can be used to submit data yourself, or to identify data submitted by others for re-use in your own research domain. The **Elixir Core Data Resource**[66] can be used to identify repositories and data tools in the life-sciences. Finally, **re3data.org**[67] is a comprehensive registry of research data repositories.

As explained in other parts of this report as well, Portell Silva emphasised the importance of having **well-described and detailed metadata.** This is especially important for sensitive data, since users can only evaluate the suitability of the data based on its metadata before requesting access. What metadata should be used greatly depends on the type of digital object and its purpose. In order to identify the metadata standard that suits your data, FAIRsharing.org lists over 1,600 different standards for various data types and purposes.

Navigating the FAIR data landscape can be difficult, and data holders may need assistance with questions such as where to deposit their data, which standards to use, and which (meta)data to collect. **ELIXIR**[68] **provides FAIR data management support for the life sciences** via their national nodes throughout Europe. Portell Silva highlighted four different tools/resources that can assist you in various parts of the data lifecycle (see Figure 4):

- **The RDMkit**[69] provides guidance and tools for all parts of the data lifecycle in the life sciences, or for particular roles and tasks, disciplines, and national contexts. The RDMkit also contains various guides and tools specific to sensitive data and GDPR

[64] Source: Laura Portell Silva's presentation slides. See the footnote under the header of this section for details.
[65] https://fairsharing.org/
[66] https://elixir-europe.org/platforms/data/core-data-resources
[67] https://www.re3data.org/
[68] https://elixir-europe.org/
[69] https://rdmkit.elixir-europe.org/

issues, including examples of how different data holders have navigated associated challenges.

- **The FAIR Cookbook**[70] contains various 'recipes' for making data FAIR, including guidance and references for how to make data FAIR for various topics and disciplines. For example, a specific recipe can be found on how to make a FAIR data protection impact assessment.

- **Data Stewardship Wizard (DSW)**[71] is a tool for creating data management plans (DMP) for various disciplines and purposes. By simply filling out a questionnaire about your data and research, the tool produces a document with a suitable data management plan, which is compatible with the requirements of the European Commission for EU-funded projects. The DSW also assesses how FAIR your data is, and contains specific questions related to sensitive data.

- **FAIRsharing.org**[72] contains a host of information, training resources, and services related to FAIR data for engineering, the humanities, as well as natural- and social sciences. Besides the resources already mentioned above, FAIRsharing contains 'collections' on various topics: e.g. the BY-COVID collection on data sources and standards used in the project.

Figure 4: FAIR data support tools and the data life cycle[73]

In the **Q&A** after Portell Silva's presentation, one of the participants asked how FAIR data practices for sensitive data can be adapted to the particular needs of different domains and research communities and what challenges remain in this area. Portell Silva responded that in the BY-COVID project, which includes four different research domains, health data and biomolecular data have proved particularly challenging. This is because such data is often

[70] https://fairplus.github.io/the-fair-cookbook
[71] https://ds-wizard.org/
[72] https://fairsharing.org/
[73] Source: Laura's presentation slides. See the footnote under the header of this section for details.

particularly sensitive, and regulations often vary from country to country, making it particularly difficult to use such data. The project has made progress in tackling these challenges to make use of sensitive data from these domains, and Portell Silva hopes that the lessons from this work can also be applied to other domains in the future.

3.3.4 General Q&A session

Following the presentations, Kalaitzi invited the audience to ask any questions to the three presenters, and share their experiences from their own organisations and work regarding making sensitive data FAIR.

One of the participants raised the issue with **DMPs**. The participant described how many academic researchers lack knowledge about how to create them, and that universities often lack resources that can help them with this. Furthermore, he described how it can be difficult to appreciate the requirements of data management, and what one needs to know, especially for someone who comes from a different background of the research in question. The participant therefore found tools such as the **Data Stewardship Wizard (DSW)** tool described in Portell Silva's presentation to be particularly useful for people in his position.

During the session on FAIRification of sensitive data, a few key take-home messages were identified by the speakers and the participants:

- Making sensitive data Findable, Accessible, Interoperable, and Reusable (FAIR) is essential and possible.
- It is primordial to consider FAIRification at the early stages of research and in the Data Management Plan (DMP) while continuously working towards FAIRification throughout the research project. This includes implementing measures to sustain the data's FAIR characteristics over time.
- RDM should encompass the entire data lifecycle.
- RDM must adhere to FAIR principles.
- FAIR does not equate to open access. Data can be "as open as possible, as closed as necessary."
- A domain-specific and trustworthy repository is recommended for storing research data relevant to specific fields.

4. Conclusions

The training stream on Data sharing and reuse under GDPR was particularly impactful. It not only highlighted the importance of compliance with data protection laws but also emphasised the need for transparency and accountability when dealing with sensitive information. Participants were encouraged to become proactive in their approach to data management, seeking advice and support from professionals within their institutions to navigate the complexities of GDPR.

Overall, the BY-COVID Fest workshop aimed to empower both data users and producers in the BY-COVID community and beyond. By the end of the event, attendees were expected to feel more confident in their ability to submit or use sensitive data, articulate their questions and concerns to data protection experts, and implement effective solutions to ensure the integrity and security of the data they handle. This initiative was a significant step towards promoting a culture of data sharing that respects privacy and fosters trust among stakeholders in the data ecosystem and in the BY-COVID project.