PaNOSC
Photon and Neutron Open Science Cloud
H2020-INFRAEOSC-04-2018
Grant Agreement Number: 823852

Deliverable: 8.1 Report on lessons learned and future prospects for adopting best practices data stewardship at the PaNOSC facilities

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Project Deliverable Information Sheet

Project Reference No.	823852
Project acronym:	PaNOSC
Project full name:	Photon and Neutron Open Science Cloud
H2020 Call:	INFRAEOSC-04-2018
Project Coordinator	Andy Götz (andy.gotz@esrf.fr)
Coordinating Organization:	ESRF
Project Website:	www.panosc.eu
Deliverable No:	8.1
Deliverable Type:	Report
Dissemination Level	Public
Contractual Delivery Date:	31/08/2021
Actual Delivery Date:	01/12/2021
EC Project Officer:	Flavius Alexandru Pana

Document Control Sheet

Document	Title: Report on lessons learned and future prospects for adopting best practices data stewardship at the PaNOSC facilities
	Version: 1
	Available at: https://github.com/panosc-eu/panosc
	Files: 1
Authorship	Written by: Teodor Ivănoaica, T. H. Rod, Martin Dostál, Jiří Bartoš, Stella d'Ambrumenil
	Contributors: A. Götz, Fabio Dall'Antonia
	Reviewed by: A. Götz
	Approved: Jordi Bodera

List of participants

Participant No.	Participant organisation name	Country
1	European Synchrotron Radiation Facility (ESRF)	France
2	Institute Laue-Langevin (ILL)	France
3	European XFEL (XFEL.EU)	Germany
4	The European Spallation Source (ESS)	Sweden
5	Extreme Light Infrastructure Delivery Consortium (ELI-DC)	Belgium
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

Table of Content

Project Deliverable Information Sheet 2

Document Control Sheet 2

Table of Content 3

1. Introduction 4

2. FAIR Data Policy implementation driving FAIR Data Stewardship 4

2.1. Data Policy implementation challenges and next steps 5

2.2. Data commons 5

2.3. Data Management Plan 6

2.4. Data curation and PID strategies 7

2.5. Findable data and data analysis 8

3. Conclusions and future perspectives 9

Examples of workshops & presentations 10

Examples of teaching material 11

Pan-learning.org course 11

Final Conclusions 12

References 13

Further Reading 13

1. Introduction

The FAIR Data Policy Framework, developed in the context of the PaNOSC project to support the adoption of FAIR principles, now serves as a community reference for the Photon and Neutron facilities involved in PaNOSC and beyond. The framework supports the facilities in developing their data policies following the FAIR principles. Moreover, the partners have developed a set of necessary "Guidelines for implementing a Research Data Policy"[1] that helps facilities with identifying and structuring the internal organizational support for adapting the FAIR Data Policy and update processes.

Adopting a FAIR Data Policy is not only a commitment to follow community standards but, in most cases, also a change in the culture of an organization, a change in the users' and staff's mentality which has to be carefully socialized and implemented and for which each PaNOSC partner has to prepare the necessary awareness and training campaigns.

In this context, the current report aggregates and presents the progress and challenges in adopting FAIR standards for data and services, including associated training, identified by the photon and neutron community members during the PaNOSC project.

The collected list of challenges and solutions have been presented, in different workshops during the last year, where common topics like DOIs, PIDs, ORCHID, NeXus, were introduced as solutions or candidates for solving the photon and neutron (**PaN**) community Data Governance and Data Stewardship challenges.

2. FAIR Data Policy implementation driving FAIR Data Stewardship

In the first part of the PaNOSC project, PaN partners have initiated the process of adapting the FAIR Data Policies. This way, the core principles of a FAIR Data Governance plan have been introduced, preparing the member RIs to become FAIR Data providers for their scientific communities.

Data stewardship encompasses the management operations, services, and oversight of all data governed by each partner RI's specific Data Policy. It is the integrated process that will be structuring the collaboration between different stakeholders, IT groups, and users, driving the implementation of data processes supporting the Data Policy.

Data Policies based on the PaNOSC Data Policy Framework is the core document expressing the support of the top management of each partner RI for the FAIR standards and are now in different stages of the approval process. The implementation of a Data Policy is a long-term operation, involving changes at multiple levels, changes that require careful preparation, including user awareness, and even technical training.

To gain momentum, each PaNOSC of the pan partners' teams have been actively working on identifying common sets of guidelines and best practices facilitating the Data Policy adoption process, and, together with these processes, they have also identified new tools and services that could support the implementation.

[1] https://zenodo.org/record/4899344/files/PaNOSC-D2.3-FINAL.pdf

2.1. Data Policy implementation challenges and next steps

In the process of implementing FAIR Data Policies, the Data Stewardship and Data Custodian are, from the policy point of view, the key challenges that have to be solved in order to have consistent implementation of the Data Policy. This challenge can only be solved by developing the necessary support procedures and processes to structure the organizational support needed for the teams to develop and implement a set of data policy-compliant tools and services supporting both the facility operations and users across different scientific disciplines. For a successful and reliable implementation, it is equally important to have both facility staff and users performing experiments that are data-aware and understand the data mechanisms and data services and the principles governing the data processes at each facility; a process that starts with the definition of a common Data Management Plan (DMP) Template (PaNOSC Deliverable D2.2 and ExPaNDS D2.2). This process is the first step towards the adoption of fair standards and initiates FAIR data lifecycle management and data stewardship at each PaNOSC partner facility.

From the Scientific Community perspective, having Data policies implemented by more RIs, even if those policies are based on a common set of data commons, in this case, developed in the context of the PaNOSC Data Policy Framework, will not necessarily provide consistency across the community, as each implementation process can be impacted by the specifics of each organization. In this context, guidelines and best practices are playing a crucial role in preserving consistency across the entire scientific community, facilitating not only the definition of the data commons but making data commons a reality.

Each PaN partner facility is now addressing the Data Policy implementation challenges for which data stewardship and data management best practices are playing a crucial role, helping them identify and integrate a common set of tools and services, together with necessary roles and responsibilities supporting FAIR data standards.

2.2. Data commons require common understanding and data-aware actors

PaNOSC[2] and its sister project **ExPaNDS**[3] are bringing together a significant number of major world-class European photon and neutron research facilities, joining forces to initiate the development of a FAIR and integrated pan-European data platform supporting the complete scientific cycle from experiment proposal to publication.

As a first step towards this common goal, the **PaNOSC**, and **ExPaNDS** partners are developing Data Commons, meaning services and tools for managing data, data storage, data analysis, and simulation and the governing principles for this, for the many scientists from existing and future disciplines using data from photon and neutron sources.

The PaNOSC Deliverable 2.1 “FAIR Data Policy Framework[4]”, is the cornerstone on which the PaNOSC partners are building or updating their Data Policies.

The framework presents a common set of standards and definitions for the PaN community (but not limited to) member facilities, also presenting the key benefits of having well-defined formats:

[2]PaNOSC - Photon and Neutron Open Science Cloud (https://panosc.eu)

[3]ExPaNDS - European Open Science Cloud (EOSC) Photon and Neutron Data Service (https://expands.eu)

[4] https://doi.org/10.5281/zenodo.3862701

- making previously measured data available for further analysis without the necessity to repeat the experiment;
- promoting data use, cross-disciplinary research, and machine learning and other data science research techniques;
- raw data becomes open to scrutiny by other researchers, which ensures scientific integrity and reproducibility of experiments;
- scientists can mine data in previously unknown ways or reapply new methods to existing data;

It is already a known fact for the entire scientific community that any changes at the level of Data Policies will require more than just upgrades of the IT Systems to support the policy implementation. Such a change produces a change of culture and mentality for the scientific users and facility staff and, at the same time, it is extremely important to build and train a common understanding of the standards and procedures and thus attract a critical mass of scientists supporting the evolution of the services and also advocate FAIR standards and policies.

2.3. Data Management Plan, a community template streamlining the implementation

Integrating a scientific research data management policy (RDP) or just a Data Management Plan (DMP) supporting the implementation of a Data Policy into an already functional scientific research facility that is already dealing with data and users is not an easy task.

The DMP is the 'living document' aiming to support researchers through the entire lifecycle of an experiment. Its life starts together with the experiment planning and design process, followed by data collection processes and data curation activities, and ends with the data publication and subsequent archival.

The Data Management Plan challenge, addressed by the PaNOSC and ExPaNDS partners together and, by sharing the know-how, the challenges, and expertise of the entire PaN community is now working on developing a DMP template for the facility users. The template is offering the necessary support to the facilities working on the adoption of a DMP strategy. At the same time, circulating such a template with users from all the PaNOSC partner facilities also promotes these common definitions and standards, and this way the collaboration aims at reaching a necessary critical mass of scientists, to review, evaluate, validate the quality suggested approach.

For the above-presented approach, both PaNOSC and ExPaNDS projects partners, have already followed community standards and even best practices to agree on DMP templates[5], this will streamline the implementation of standard machine-readable DMPs for the PaN community experiments.

[5] https://ds-wizard.org/about - Data Stewardship Wizard - a tool facilitating the creation of experiment projects DMP

2.4. Data curation and PID strategies - what and how to address the persistent identifiers challenge

According to PaNOSC Data Policy Framework, "*All raw data will be curated in well-defined formats, for which the means of reading the data will be made available by the facility*[6]", because the "*Metadata that are automatically captured by instruments will be curated and stored in a catalogue or similar repository which links the metadata to the raw data they are describing*[7]" and also the fact that " *The experimental team shall be able to create a DOI for one or more specific datasets to be cited in a publication*[8]", all these are parts of the data curation process and are one of the most important steps of the data policy implementation process. For all of the above-presented processes and definitions, the PaN member facilities have now a standard and mature understanding which allows the members to properly evaluate and select the community best practices facilitating the implementation of the PIDs compliant with the Data Policies and, at the same time, respecting the requirements of their specific user community.

These definitions, part of the data commons across the photon and neutron facilities and, though they are developed based on the Data Policy Framework and start from common definitions, at the level of the facility they could be treated differently or, impacted by already existing strategy (compliant or partially compliant with the data policy) could result in a parallel/not standard approach.

This is why, in most cases, adopting the domain-relevant data-stewardship best practices facilitate the adoption of a common set of data services that will improve the FAIRness of the data sets.

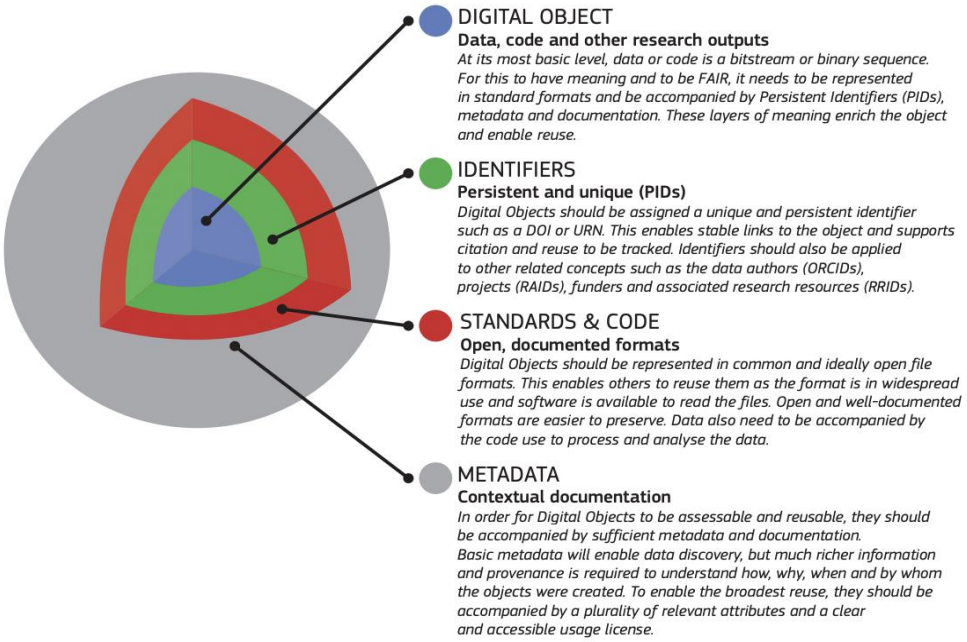

Fig1 Example model for FAIR Digital Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable[9]

For a successful and reliable implementation, it is equally important to have both facility staff and users performing experiments are data-aware and understand the data mechanisms and data services and the

[6] https://doi.org/10.5281/zenodo.3738497 - Raw data and metadata - 3.4.4
[7] https://doi.org/10.5281/zenodo.3738497 - Raw data and metadata - 3.4.5
[8] https://doi.org/10.5281/zenodo.3738497 - Persistent Identifiers - 3.3.3
[9] https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_0.pdf

principles governing the data processes at each facility, a process that starts with the definition of a common Data Management Plan Template (PaNOSC Deliverable D2.2 and ExPaNDS D2.2).

Digital Object Identifier system (http://doi.org) - This is a system that enables digital identifiers to be defined for any kind of object or device/system, the system was started to identify physical devices and, over the years, has been extended to also identify data sets. It makes the datasets traceable during the entire data lifecycle, from the moment the data sets are produced to the moment they are published or cited in scientific publications.

All partners are also adopting a distinct service, but an equally important one for the users is the DOI landing page, a web page associated with each DOI name is derived from this concept and it usually contains high-level metadata information (author, title, date, categories) and, in some cases, it can even facilitate access to the specific data set.

Electronic logbooks are an essential part of an experiment, they are an extremely important tool for collecting valuable metadata from scientists. Among the core common features expressed by PaN partner facilities users we mention:

- Online shared editing, a collaborative approach allowing multiple users to insert data at the same time;
- Powerful search facilities allow the user to perform searches;
- Access rules during the embargo period, the ideal experiment logbooks allow scripting (so that users might even be able to pull data from the CS or data archivers in real-time)
- Offer a clear picture of what happened during the experiment;

The common challenges are the one thing bringing together facilities and users that share a common belief, a common strategy, allowing them to share know-how and come up with new solutions, built to serve a wider scientific community and thus contributing both to finding a solution as well as to train the future users' community.

2.5. Findable data and data analysis - data is a common language that needs to be properly

In terms of findability, the entire PaNOSC community has agreed on using common standards and most of the partners are using, planning or testing, a possible NeXus (a standard using HDF5 file format) approach for their data. The reasons for choosing this as a possible standard for a community are determined by the fact that NeXus is already providing a standard vocabulary for the scientific community, while the HDF5 data format is one of the most common data formats of the PaN community members.

Even if common standards for data and metadata, discussed together with all PaN members are agreed upon, having standards like NeXus and HDF5 implemented is still going to add more challenges, at least for new facilities that have to identify and engage a critical mass of beam scientists to build and validate the necessary application definitions specific to their research areas.

Having data in well-defined formats enables the adoption of common file cataloguing solutions by the community, supporting common data transfer solutions, allowing users to have data replicated at their home institute or even on their computers or laptops to be analyzed.

The common data cataloguing solutions considered, like ICAT, SCICAT, or even INVENIO test setups (as possible competitive solutions) are now tested or updated to support the ingestion of FAIR data sets. These

solutions are adding another challenge, the implementation of a federated search API and this way integrating the RI resources with EOSC.

3. Conclusions and future perspectives - It's all about FAIR (ness).

The FAIR Data Policy Framework, developed in the context of the PaNOSC project to support the adoption of FAIR principles, now serves as a community reference for the Photon and Neutron facilities involved in PaNOSC and beyond. The framework supports the facilities in their data policies following the FAIR principles. Moreover, the partners have developed a set of necessary "Guidelines for implementing a Research Data Policy" that help facilities with identifying and structuring the internal organizational support for adapting the FAIR Data Policy.

This report documents case studies from each PaNOSC partner and lists some lessons learned for implementing FAIR Data Commons at the facilities. It is clear from that report that the involvement of top management is a necessity and also that the question about open science and FAIR data has never been in question at the newer facilities (ELI, ESS), partly due to the influence from the European Commission, which is an important stakeholder for such Pan-European research infrastructures. However, at established facilities with existing data, it tends to be a more cumbersome and slow process. The data volume also naturally plays a role in the willingness to adopt the FAIR Data Commons as exemplified by ILL, a well-established neutron source, and ESRF, a well-established photon source. ILL has stored all their data from Day One (in 1973) probably because its data production always has been rather modest compared to available storage technology. The cost of data storage has therefore not been a hindrance. On the other hand, ESRF produces significantly more data and the storage of those carries a high price tag that can be a hindrance for adopting the FAIR principles. That being said, the promotion of a FAIR data policy framework, although implicitly FAIR, a decade ago (REF - pan-data), and the use of a common data format across PaN sources, in the form of NeXUS, has undoubtedly reduced the barrier for adopting the FAIR principles at the Pan-European PaN sources.

Whilst there is agreement among the partners and their governing bodies, that they should implement FAIR data policies and associated services, the socializing of the FAIR principles and their impact among staff and users, provides a more fragmented picture. Referring to the previous report, we note that training activities related to the FAIR principles at the partner facilities are essentially absent, whereas plenty of cross-facility workshops has been held by evangelists, particularly those sponsored to do so through projects like ExPaNDS or PaNOSC or other EOSC related projects funded by the European Commission. At these workshops, common topics like DOIs, PIDs, ORCHID, or NeXus, were introduced as solutions or candidates for solving the PaN community Data Governance and Data Stewardship challenges. Likewise, there is plenty of information and teaching material available on the web.

We consider it important to socialize the FAIR principles because adopting a FAIR Data Policy is not only a commitment to follow community standards. Implementing FAIR principles also requires a change of culture and mentality for the scientific users and facility staff and, at the same time, it is extremely important to build a common understanding of the standards and procedures and get a critical mass of scientists to understand the benefits to advocate and translate those to the community. It is commonly understood that many users are reluctant to make their data open (accessible) and if they have to do so, reluctant to make them reusable by providing sufficient information about the experiment (e.g., the sample).

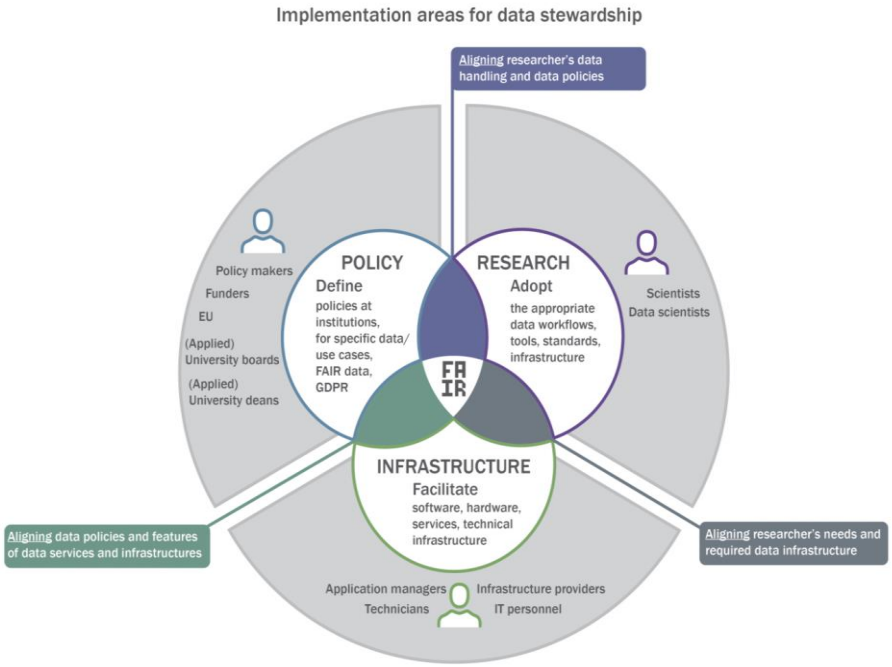

fig2: interplay between the three actors10; 1) Policymaker, 2) Facility, and 3) user

Examples of workshops & presentations

Name	Organizer / event	Sponsor	Comments
Plenum presentation	ESS / ILL User Meeting 2020	ESS & ILL	Presented by Andy Götz in plenum. Critical comments from more senior scientists.
Laserlab-Europe, ELI and CASUS Workshop "Better Data for Better Science" 28-29 October 2021	LaserLab Europe	LaserLab &ELI	The importance of a scientific data management system/DMP Presented by Andy Götz
ELI Scientific Data Management System	ELI ERIC	ELI ERIC & ELI Beamlines	Presented by Teodor Ivănoaica, presenting ELI Fair Data Policy principles and implementation challenges
2nd online workshop of the Battery2030+	Battery 2030+	Battery 2030+	Presenter Andy Götz - PaNOSC experience in Research Data

[10] https://fairtoolkit.pistoiaalliance.org/methods/data-stewardship/

Name	Organizer / event	Sponsor	Comments
Initiative			Management presented at the Battery2030+ Initiative on RDM
The FAIR Workshops – 1st and 2nd October 2020	ExPaNDS	ExPaNDS	Multiple presenters, Jean-Francois Perrin - presenting DOI's importance for linking proposals and publications

Examples of teaching material

Name	Organizer	Sponsor	Comments
FAIRsFAIR-CODATA-RDA Data Steward Training Series	FAIRsFAIR	FAIRsFAIR	https://www.fairsfair.eu/events/training/fairsfair-codata-rda-data-steward-training-series-0
FAIR Data Stewardship Training Fellowship	ELIXIR	ELIXIR-UK &UKRI	FAIR data management in the Life Sciences practices by UK researchers and their research support staff.
FAIRsFAIR & EOSC Synergy Data Steward Instructor Training Workshops	FAIRsFAIR & EOSC Synergy	FAIRsFAIR & EOSC Synergy	Data steward instructor training workshops

Pan-learning.org course

The development of an e-learning course on FAIR data and data stewardship, presenting principles and best practices, has been started on pan-learning.org using existing teaching material and examples provided by PaN partners in different workshops. The course is intended as a container for relevant information related to data stewardship standards, presenting best practices and guidelines for users and facilities adopting FAIR standards.

Considering the specifics of each PaNOSC work package developing FAIR tools and services (WP2-WP5), the development of PaNOSC specific training material, employing proper data stewardship procedures and best practices used in the design of the PaNOSC services has been started. The PaNOSC specific training material will be presented in a series of events, starting with a Data Stewardship Workshop which will be prepared in the context of Task 8.4 and D8.2 activities, focusing on lessons learned for adopting data stewardship and e-learning platform at the PaNOSC facilities.

Final Conclusions

It is extremely important from the organization's perspective that "Data *Stewardship explains everything you need to know to successfully implement the stewardship portion of data governance, including how to organize, train, and work with data stewards, get high-quality business definitions and other metadata, and perform the day-to-day tasks using a minimum of the steward's time and effort*."

At the same time, for the scientific communities, **Data Stewardship** is a process involving more than just the specific research facility providing instruments and services, aiming at producing FAIR Data, it also includes a user perspective, making it a community challenge. Since scientific data is defined together with the users, the Data Stewardship can be seen as "*The process and attitude that makes one deal responsibly with one's own and other people's data throughout and after the initial scientific creation and discovery cycle*".

As an effect of FAIR standards and principles being adopted by PaNOSC members, new Data Governance and Data Stewardship challenges arise, challenges that are now introducing a change of mentality, culture for both users and facility staff. This change adds a new dimension to the users' data, to the users' data analysis habits which need to be properly trained and prepared in order to maximize its potential.

Among these challenges, proper presenting data stewardship, in such a way that the same standard best practices are reaching their target audience (scientists, staff, IT groups) a set of community-specific training has to be developed and promoted. In this context, multiple training events, workshops and conferences where FAIR principles and Data Stewardship core practices have been organised, introducing data standards to a very diverse audience, combining senior scientists, students and IT professionals. Though the feedback is positive, each scientific community requires custom presentations, tailored to their specific areas of interest or implementations derived from scientific use cases that could present a facility-specific FAIR data management approach.

In this context, considering the specifics of each PaNOSC work package developing FAIR tools and services (WP2-WP5), the development of PaNOSC specific training material, employing proper data stewardship procedures and best practices used in the design of the PaNOSC services has been started. The PaNOSC specific training material will be presented in a series of events, starting with a Data Stewardship Workshop which will be prepared in the context of Task 8.4 and D8.2 activities, focusing on lessons learned for adopting data stewardship and e-learning platform at the PaNOSC facilities.

More PaNOSC specific training is needed for supporting the adoption of FAIR Data Stewardship best practices across the PaN community members, as well as for building a new generation of FAIR data-aware users.

References

1. Gotz, A., Perrin, JF., Fangohr, H., Salvat, D., Gliksohn, F., Markvardsen, A., … Matthews, B., (2020), *PaNOSC FAIR Research Data Policy framework* (Version 1.1). Zenodo, https://doi.org/10.5281/zenodo.3826039
2. Gotz, Andy; Taylor, Jonathan; Dimper, Rudolf; Perrin, Jean-François; Gliksohn, Florian; Roccella, Dario; Wrona, Krzysztof; Ivănoaica, Teodor; Malka ,Janusz; Collins, Stephen https://zenodo.org/record/4899344#.YZyHGL1udfU

Further Reading

1. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, European Commission Expert Group on FAIR Data DOI: 10.2777/1524- https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_0.pdf
2. Mark D. Wilkinson et al.# https://www.nature.com/articles/sdata201618 DOI: 10.1038/sdata.2016.18
3. Professionalizing FAIR Data Stewardship in the life sciences: defining job criteria & skills https://www.dtls.nl/2019/10/21/professionalizing-fair-data-stewardship-in-the-life-sciences-defining-job-criteria-skills/