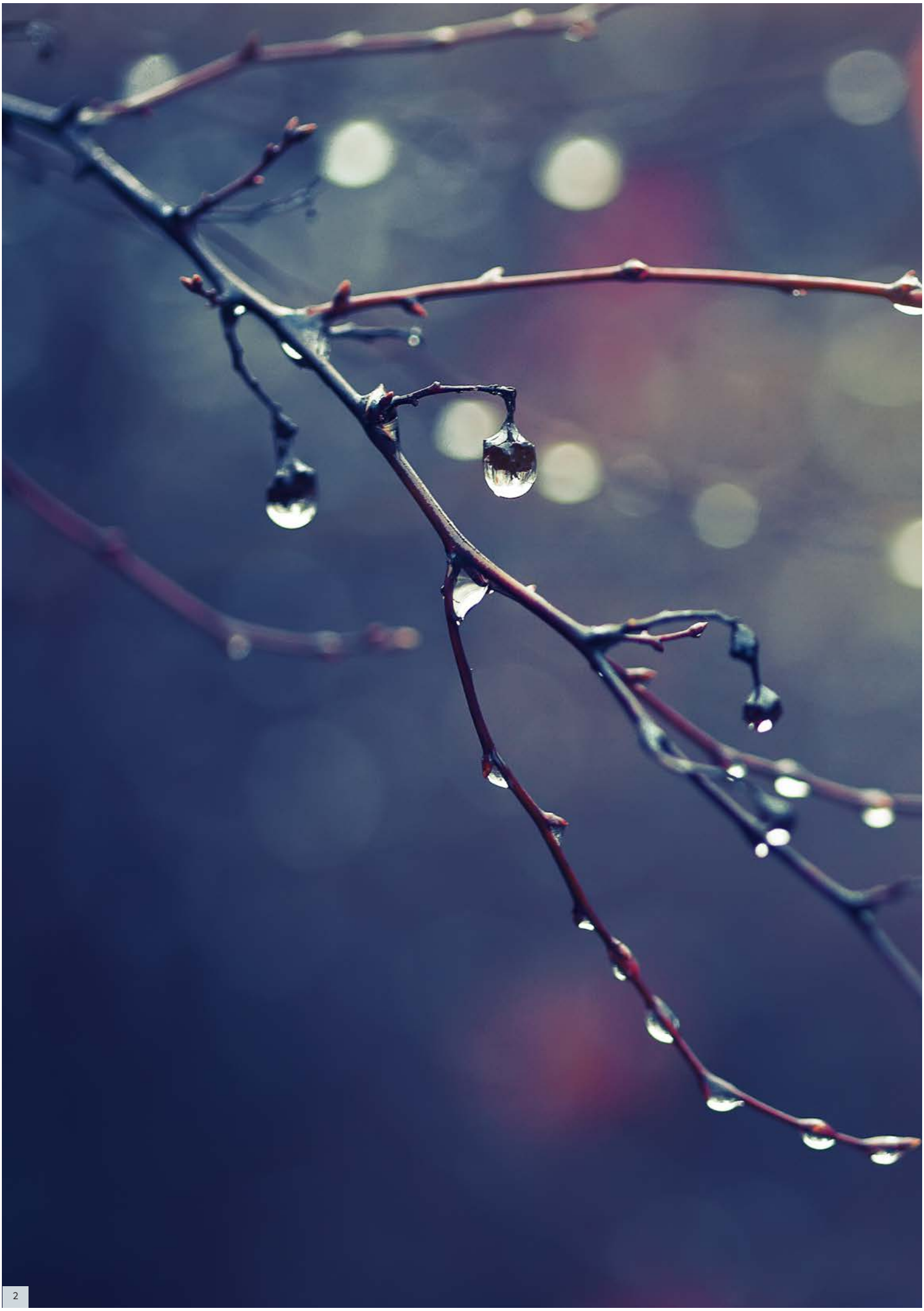
NORDIC E-INFRASTRUCTURE COLLABORATION

THE STATE OF OPEN SCIENCE IN THE NORDIC COUNTRIES

"Enabling data-driven science in the Nordic countries"

A REPORT BY ANDERS O. JAUNSEN ON BEHALF OF NEIC

SEPTEMBER 2018



ACKNOWLEDGEMENTS

The NeIC data management working group (DM-WG) has provided its input via direct comments to the document and via discussions in meetings during the process of preparing this document.

The members of the DM-WG are:
- Dejan Vitlacil
- Anders Sparre Conrad
- Maria Francesca Iozzi
- Ilkka Lappalainen

Special thanks to Michaela Barth (NeIC Executive Manager), Rob Pennington (NeIC Special Advisor) and Gudmund Høst (NeIC Director) for many valuable comments and suggestions. Thanks also for the input on national strategies from Minna Ahokas (CSC), Karl Gertow (Vetenskapsrådet), Beate Eellend (National Library of Sweden), Marte Qvenild (Research Council of Norway), Jyrki Hakapää (Academy of Finland), Anne Sofie Fink Kjeldgaard (RigsArkivet), Jens Begtrup (NordForsk) and Eiríkur Stephensen (University of Iceland).

The Nordic countries are particularly well suited for collaboration among each other due to social and cultural similarities.

CONTENTS

SCOPE OF THIS DOCUMENT

The Nordic countries are particularly well suited for collaboration among each other due to social and cultural similarities. Also, as the countries are individually small, unifying efforts in science and technology to realise common undertakings will generally result in a better end-product and greater impact in the international arena. Finally, a Nordic-wide collaboration reduces the risks of duplication of effort and therefore promotes a more cost-efficient R&D segment within the Nordics.

In 2017 the NeIC Board identified "data management" as a subject that was not being sufficiently addressed in the NeIC project portfolio. The Board recommended that NeIC should direct some effort towards data management and, given the importance of research data in general and the onset of the European Open Science Cloud (EOSC) in the European arena, this seems timely and appropriate.

This document seeks to identify activities that can help to improve the conditions and means for enabling data-driven science in the Nordics. In particular, the relevance to EOSC and the greater visions of Open Science have emerged as an appropriate setting for this report. However, given the context, we limit our discussions and understanding of Open Science to Open Access (of publications) and Open Data (see the following section for a clarification on the terms and their definitions).

01

THE ROAD TO OPEN SCIENCE

01

Open Science is scholarly research that is collaborative, transparent and reproducible and whose outputs are publicly available (OSPP-REC, 2018, doi:10.2777/958647). At its core, Open Science aims at: "increasing research quality, boosting collaboration, speeding up the research process, making the assessment of research more transparent, promoting public access to scientific results, as well as introducing more people to academic research" (Friesike & Schildhauer, 2015, doi:10.1007/978-3-319-09785-5_17).

Many, if not most, governments have acknowledged the need to share federal or governmental data, be it economic or societal statistics. It is not difficult to see the relevance and importance of sharing such data with the public who has ultimately paid for this information. The same argument applies to publicly funded research, which also potentially holds a (reuse) value apart from its original intended purpose. However, the majority of such research data is either not shared by the creators, or is in a state that prevents reuse by others. Data that does not contain *rich* metadata, including intrinsic, contextual, provenance and administrative metadata, by which it can be found, interpreted and assessed is "reuseless" (B Mons, "Data Stewardship for Open Science", ISBN 9780815348184; Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship").

The cross-disciplinary reuse of research data holds particularly great potential added value. Simply providing access to such research data will *not* be enough to foster its reuse. A citizen scientist or colleague from another research field will rarely download specific datasets or have the competence or skills needed to analyse data on which he or she is not an expert. It is necessary to arm the data with meaningful and standardised metadata that exhaustively documents the content of the dataset in such a way that a non-expert may extract relevant information in order to judge its suitability and facilitate its reuse. Crucially, such datasets must be machine-actionable in order to support automated querying and assessment. In other words, the dataset, or the service providing it, must enable a machine to assess the data´s suitability.

OPEN ACCESS AND OPEN DATA

"Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools" (European Commission, doi:10.2777/061652, p.33). The OECD defines Open Science as: "to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction" (OECD, doi:10.1787/5jrs2f963zs1-en, p.7). Some even extend the principles of openness to the whole research cycle, but we will stick to the moderate and more common definition that concerns itself with post-publication transparency.

According to Wikipedia there are six principles on which Open Science is based: Open Educational Resources, Open Access (to publications), Open Peer Review, Open Methodology, Open Source and Open Data.

Two of these principles in particular concern research data: "Open Access" – which refers to the principle of free access to refereed/published scientific results via open access journals or minimal license restrictions, and "Open Data" – which refers to the free access to scientific research data.

See figure 1 on next page.

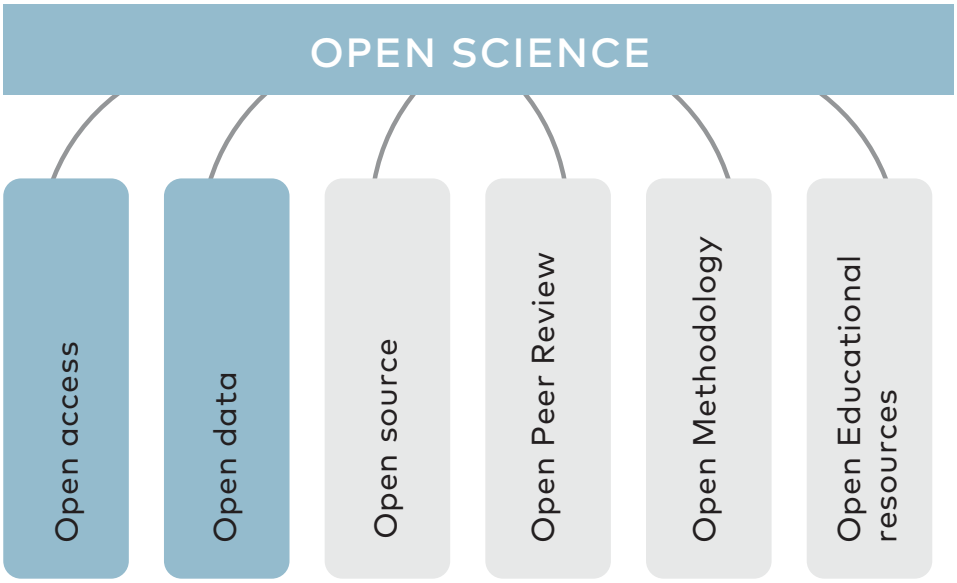


FIGURE 1 The six principles of Open Science and the two data-related themes that are the subject of this report (source Wikipedia; Open Source)

The "Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020" (H2020 pilot guide) provide the following definition:

"Open access (OA) refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable. 'Scientific' refers to all academic disciplines. In the context of research and innovation, 'scientific information' can mean:

1. *peer-reviewed scientific research articles (published in scholarly journals) or*
2. *research data (data underlying publications, curated data and/or raw data)."*

The H2020 definition of "open access" concatenates open access and open data. Open access is one of the means of achieving open science. Along with the six principals open data, open source, open methodology, open peer review and open educational resources – open access to research data refers to the right to access and reuse digital research data under the terms and conditions set out in the Grant Agreement under which it was generated. Research data refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation.

In a research context, examples of *data* include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form. Users can normally access, mine, exploit, reproduce and disseminate openly accessible research data free of charge as long as the data creators or owners are appropriately and correctly cited. The precise conditions for reuse of data are best communicated by issuing a *data usage license* when publishing the dataset.

The term *open access* is often used in reference to both open data and access to science publications. In this document, we employ the Wikipedia definition of the two terms in order to clearly differentiate between the aspects:

- Open Access refers to research outputs which are distributed online and free of cost or other barriers, and possibly with the addition of a Creative Commons license to promote reuse. Open access can be applied to all forms of published research output, including peer-reviewed and non peer-reviewed academic journal articles, conference papers, theses, book chapters, and monographs.
- Open Data represents the concept of openly sharing research data in raw or processed form. A piece of data is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike.

02

NATIONAL STRATEGIES ON OPEN SCIENCE

02

Open science is a strategic priority for the EU. The Open Science goal is materialising in the development of a European Open Science Cloud and access to scientific data generated by Horizon 2020 projects. Here we present a brief summary of the national strategies in the Nordic countries on working towards Open Science, but limit ourselves to the two most relevant aspects in our research data context, *Open Access* and *Open Data*.

DENMARK

A new National Strategy for Open Access was published in June 2018. The strategy states that the implementation of Open Access is to take place through the green model – i.e. parallel filling of quality-assured research articles in institutional or subject-specific archives (repositories) with Open Access. In order to monitor the transition to Open Access publications, an OA indicator service has been established.

On Open Data the following is stated: "In Denmark there is a long tradition for data management and a number of Danish initiatives and organisations work on opening the access to data and making data accessible to research and development". The Danish Agency for Science and Higher Education has commissioned Oxford Research and Højbjerre Brauer Schultz to carry out a preliminary analysis of the potential for implementing FAIR data in Denmark.

The preliminary analysis points to the following key factors to implement FAIR data in Denmark:

- National coordination and cooperation is needed across research actors, libraries and research funding actors;
- Competence and culture must be supported to give the researchers the necessary skills and incentives for sharing data
- Access to research data is a precondition for encouraging researchers to take part in sharing and using digital research data
- A wide range of technical and legal barriers must be addressed – at the national as well as international level
- Further investments are needed in specialised infrastructure for storage, handling, processing and dissemination of research data

The implementation of FAIR principles appears to be closely related to Open Science.

ICELAND

Since 2012 the official policy in Iceland requires researchers who are funded by public research funds to disseminate scientific results in the form of open access publications (Open Access policy). During the fall of 2018 work on a national policy on open data will begin in accordance with the strategy of the Science and Technology Policy Council 2017-2019.

FINLAND

The Open Science and Research initiative (ATT) was established in 2014 by the Finnish Ministry of Education and Culture to incorporate open science and research into the entire research process. The impact of the Initiative is analysed in an external evaluation. The target groups of the evaluation were the research organisations and their staff members, research funders, the national stakeholders, representatives of the innovation ecosystem

and international organisations (UNESCO, OECD, European Commission, NordForsk and Nordic Council of Ministers).
In 2016, the Ministry of Science stated: "Open science is one of the spearheads of Finnish science policy and it must be promoted by all means necessary. The Ministry of Education and Culture has outlined that Finland will become one of the leading countries in open science and research by 2017. The objective is to have open access to all scientific publications by 2020."

Recently, the universities and other higher education institutions, together with the Academy of Finland, have established a national action plan, the "Open Science and Data Action plan". The main objective is that in Finland, Open Science is to be part of the daily life of science at all levels. One of the earliest tasks, currently in a planning phase, is the establishment of a national coordination office for open science at the Federation of Learned Societies.

NORWAY

According to Meld. St. 18 (2012-2013) "Long-term perspectives: knowledge provides opportunities" (white paper on research), "In principle, it is the Government's view that all research that is wholly or partially funded through public allocations must be made openly available."

In 2017 the Norwegian Government published "National goals and guidelines for open access to research articles", which states that all publicly funded Norwegian research articles should be made openly available by 2024. The document contains guidelines and measures for open access to research articles in Norway.

The national strategy on access to and sharing of research data was published in December 2017 and states three basic principles for publicly funded research data in Norway:

- Research data should be as open as possible and as closed as necessary.
- Research data should be processed and adapted in such a way that the content of the data can be exploited in the best possible way.
- Decisions on archiving and facilitating research data must be taken in the research communities.

The Government established a new directorate (UNIT) in 2017 that, in addition to offering services, will also coordinate and harmonise IT services, increase synergies, reduce duplication of efforts and oversee the implementation of the aforementioned principles.

The Research Council of Norway (RCN) revised its Principles for Open Access to Scientific Publications in 2014:

- The Research Council requires all scientific articles resulting from research wholly or partially funded by the Research Council to be openly accessible.
- All articles with such funding must be made available in OA repositories.
- In a transition period 2014–2019 the Research Council manages a separate funding scheme to cover all OA publication fees incurred by Norwegian research institutions. The scheme will be evaluated in 2018.
- The Research Council seeks to encourage the use of gold open access to scientific articles, and therefore recommends that the institutions only cover the cost of publication fees in journals registered in the Directory of Open Access Journals.
- RCN signed the San Francisco Declaration on Research Assessment 9 May 2018.

The Research Council's Policy on Open Access to Research Data was revised in 2017.

Main principles in the policy are:

- Research data must be as "open as possible, as closed as necessary".
- Research data must be FAIR.
- Access must be provided at the lowest possible cost, preferably at no more than the marginal cost of dissemination; user-fees are approved costs in all RCN funding schemes.

SWEDEN

The Swedish Research Bill 2016/17:50 (Sec 8.4) states that the Government's goal is that all scientific publications resulting from publicly funded research should be made available immediately after they are published. Likewise, research data underlying scientific publications should be made available at the same time as the associated publication. A transition to open access to research results should be gradual in order to ensure that it is done in a responsible way. For scientific publications, the transition can begin immediately, while further investigations of the forms of open access to research data and scientific works may be required. Notably, Sweden recently decided on the non-renewal of the Elsevier agreement.

The National Library of Sweden holds the national coordination task related to Open Access to publications. As of December 2017, the National Library of Sweden (NLS) has been assigned responsibility by the Government to

develop criteria for assessing whether a scientific publication resulting from public funding meets the national objective of open access. Further, the NLS shall propose a method that provides a comprehensive overview of the extent to which both scientific publications and research data meet the FAIR principles. A report will be submitted in February 2019. Also, the NLS is instructed to continuously monitor and report on the total cost of publication for scientific publications in Sweden. When doing so the NLS shall pay particular attention to costs regarding subscriptions, publication charges (APCs), and administrative expenses.

The Swedish Research Council (VR) has the national coordination task related to Open Data. A government commission was appointed in December 2017 with the mission to develop assessment criteria to follow the development towards an open science system. Specifically, VR is to develop criteria to assess the extent to which research data, wholly or partly generated by public funding, meets the FAIR principles. The mandate of the commission states that the goal is to fully implement a transition to open access to research results, including scientific publications, artistic works and research data, in a ten-year perspective. The commission is expected to conclude its efforts and submit a report by December 2018.

STRATEGY SUMMARY

From the above it is clear that the Nordic countries have very similar ambitions on Open Science. The majority of the countries have already policies in place requiring researchers to publish their results in Open Access journals, and in the few cases where policies are not yet established, there are at the very least concrete goals. For Open Data the vision seems to be going in the same direction, although it has not matured to the same extent as Open Access publications. In many of the countries an Open Data policy is being drawn up, while a few countries have already imposed policies and others have published guidelines.

Most recently, an initative by Science Europe and the European Research Council, named Plan-S, has vowed to accelerate the transition to Open Access by requiring that participating countries commit to a list of requirements by 2020, involving copyrights to authors and not to accept hybrid open access journals. Norway, Sweden and recently also Finland has joined this cOAlition-S.

03

RESEARCH PROCESS AND THE DATA LIFE CYCLE

03

The scientific process starts with the collection and/or selection of data and ends with the publication of the data and/or science results, preferably in such a way that they can easily be reused.

According to DataONE, a distributed framework and sustainable cyberinfrastructure for environmental science, *the data life cycle provides a high level overview of the stages involved in successful management and preservation of data for use and reuse.* Alternate versions of the data life cycle exist with differences attributable to variation in practices across domains or communities. The life cycle serves as an underlying framework for the development of tools and services.

- Data collection/selection: observations/measurements are made using sensors of some kind. Alternatively, suitable existing data is selected for reuse.
- Processing / Analysis: if necessary, the data is processed and calibrated, then analysed in order to extract (new) scientific knowledge.
- Deposit or archive data: potentially useful data are located and obtained, along with the relevant information about the data (metadata) and potentially reused.
- Publish: the raw and/or processed data are somehow made available for verification and scientific results are published via an appropriate channel (e.g. journal)

FIGURE 2 Data Life Cycle – an ideal reuse cycle of scientific data

04

METADATA AND METADATA STANDARDS

04

Metadata is documentation that describes data. In a lab setting, much of the content used to describe data is initially collected in a notebook; metadata is a more formal, sharable expression of this information. It can include content such as contact information, geographic locations, details about units of measure, abbreviations or codes used in the dataset, instrument and protocol information, survey tool details, provenance and version information and much more.

Properly describing and documenting data allows users to understand and track important details of the work. In addition to describing data, having metadata about the data also facilitates search and retrieval of the data when deposited in a data repository.

INTRINSIC METADATA

Intrinsic metadata is information that is relevant to the specific piece of data. It is therefore tightly tied to the data - often generated automatically by the instrument/pipeline that created the data. There are many tools that consume these kinds of metadata and it is therefore generally best to leave intrinsic metadata "intact" and make it available in its native format.

CONTEXTUAL METADATA

Contextual metadata is information about data that pertains to a collection of data. Contextual metadata may be tied to a specific piece of data, or may be tied to a dataset, or even a set of datasets (e.g. an entire funded project). It may also be necessary to create multiple contextual metadata records ("record" or sections of larger records).

PROVENANCE METADATA

Provenance metadata is historic information about the data and may describe how the data were processed for analysis. Provenance metadata is usually attached to all levels - specific piece of data, a dataset, or a set of datasets (e.g. an entire funded project). It will be necessary to create multiple provenance metadata records.

METADATA STANDARDS

While data curators, and increasingly researchers, know that good metadata is key for research data access and reuse, figuring out precisely what metadata to capture and how to capture it can be a daunting task. Fortunately, many academic disciplines have supported initiatives to formalise the metadata specifications the community deems to be required for data re-use.

Specific disciplines, repositories or data centres may guide or even dictate the content and format of metadata, possibly using a formal standard. Because creation of standardised metadata can be difficult and time consuming, another consideration when selecting a standard is the availability of tools that can help in generating the metadata. It is good practice to be well informed about any established or agreed-upon community standards or requirements. This is ideally secured by creating a data management plan in collaboration with qualified individuals. It is generally a good idea to start by contacting a local librarian, data manager or data steward, if available, who has a better overview of the best practices in a given community or domain field.

The Digital Curation Centre provides a catalogue of some common metadata standards. Some specific examples, both general and domain specific, are listed on next page:

Dublin Core
- domain agnostic, basic and widely used metadata standard

DDI
Data Documentation Initiative) - common standard for social, behavioural and economic sciences, including survey data

EML
(Ecological Metadata Language) - specific for ecology disciplines

ISO 19115 and FGDC-CSDGM
(Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata) - for describing geospatial information

MINSEQE
(MINimal information about high throughput SEQeuencing Experiments) - Genomics standard

FITS (Flexible Image Transport System)
- Astronomy digital file standard that includes structured, embedded metadata

MIBBI - Minimum Information for Biological and Biomedical Investigations

05

THE FAIR PRINCIPLES

05

"The FAIR Principles largely revolve around the specifics of achieving (open) access to research data. They put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals." (Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship" doi:10.1038/sdata.2016.18)

The existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments. Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archiving, data stewardship includes the notion of "long-term care" of valuable digital assets, with the goal that they should be discoverable and re-usable for downstream investigations; either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, enable high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies.

The emphasis placed on the application of FAIRness to both human-driven and machine-driven activities, is a specific focus of the FAIR Guiding Principles that distinguishes them from many peer initiatives. Humans and machines often face distinct barriers when attempting to find and process data on the Web.

With the FAIRification of data one aims to support existing communities in their efforts to enable valuable scientific data and knowledge to be published and made reusable. The FAIR principles - in short - are as follows:

- Findable - (meta)data is uniquely and persistently identifiable and data should have basic machine readable descriptive metadata.
- Accessible - data is reachable and accessible by humans and machines using standard formats and protocols.
- Interoperable - (meta)data is machine readable and annotated with resolvable vocabularies/ontologies.
- Reusable - (meta)data is sufficiently well-described, contain clear and accessible usage license, detailed provenance and follow community standards.

The GO-FAIR web-pages contain a comprehensive description of the specifics of the FAIR principles, but we summarise them below for completeness.

FINDABLE

A globally unique persistent identifier (PID) for metadata and data is among the fundamental requirements of FAIR and the vision of Open Science. An identifier consists of an internet link (e.g., a Uniform Resource Locator (URL) that resolves to a web page) that defines the concept it is meant to represent, and is thus essential to the human-machine interoperation. When considering a potential PID service it is important that it guarantees a globally unique identifier that cannot be reused and that the identifier remains active for the foreseeable future. It typically requires both time and money to provide a service such as this.

Metadata should be rich enough that the dataset can be discovered and its relevance to a potential user can be assessed without retrieving the actual data. The metadata should also clearly and explicitly include the (persistent) identifier to the data that it describes. To ensure that the (meta)data can be found they must be registered or

indexed in a searchable resource. If the availability of a digital resource such as a dataset, service or repository is not known, then nobody (and no machine) can discover it.

ACCESSIBILITY

Once the user has identified the data she/he requires, the meta(data) should be retrievable by the identifier using standardised communication protocols (Principle A1) and data retrievable by a protocol that is open, free and universally implementable (Principle A1.1). FAIR data retrieval should be mediated without specialised tools or communication methods. The metadata should clearly define who can access the actual data, and specify how. This does not imply that a method that is not fully mechanised would not be acceptable under certain conditions, e.g. for sensitive or restricted data. The data access protocol should be open, free and universally implementable (Principle A1.2). The protocol must also allow for authentication and authorisation when necessary. Note that the "A" in FAIR does not necessarily mean "open" or "free". Rather, it implies that one should provide the exact conditions under which the data are accessible. Finally, metadata should be accessible even when the data itself is no longer available (Principle A2). So, while the metadata must always be open, there are acceptable reasons for restricting the access to the data itself. Examples include sensitive data that should not be made freely available (accessibility restrictions), copyrighted material that is only accessible when certain conditions are met (authorisation required) or when the data itself is not available at all due to privacy issues or having been deleted. That said, queries about access protocols and/or access conditions should be clearly stated in the metadata, including machine parsable responses, unless this is impractical or unreasonable.

INTEROPERABILITY

The data often need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. To achieve this (meta)data should use a formal, accessible, shared and broadly applicable language for knowledge representation. The controlled vocabulary used to describe (meta)data needs to be documented and resolvable using globally unique and persistent identifiers (following FAIR principles). The (meta)data should include qualified references to other (meta)data, enriching the contextual knowledge about the data and specifically indicate if one dataset builds on another, if complementary data are found in another dataset, or additional data are needed to complete it. Furthermore, all datasets need to be properly cited.

REUSABILITY

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. It will be much easier to find and reuse data if there are many labels attached to the data. The R principle is related to F, but R focuses on the ability of a user (machine or human) to decide if the data are actually *useful* in a particular context.

The (meta)data should be released with a clear and accessible data usage license. While the "I" principle contains elements of technical interoperability, this reusability principle covers legal aspects of interoperability. Non-licensed data, although "open" in the mind of most academics, is "reuseless". For a potential user it is essential to have a clear appreciation and legal consent to use the data in question. An unlicensed dataset will/should be avoided by most users or major companies, due to the potential legal consequences of using data with unclear conditions.

For data to be reusable it will require provenance metadata, including the origin of the data (i.e., clear story of origin/history), a description of the workflow, how it was processed/manipulated, who to cite and/or how you wish to be acknowledged.

Finally, the (meta)data should conform with domain-relevant community standards. It is easier to reuse datasets if they are similar: same type of data, data organised in a standardised way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary.

06

DATA STEWARDSHIP

06

As we have seen in the previous section, compliance with the FAIR principles will for some tasks require skills that a typical researcher can not be expected to possess. A scientist will generally not prioritise such tasks or carry them out with the necessary dedication. From their point of view it simply does not serve the driving force of their science project and would likely be perceived as adding to their administrative tasks and leading to further dilution of their research time. For this reason it is important to brief researchers of the overall benefits and long-term goals of going FAIR, and assure them that they will receive qualified assistance.

The full implementation of FAIR will eventually lead researchers to regain some of the time typically lost to data wrangling. The potential of this recovery should not be underestimated, as some studies (e.g. CrowdFlower) show that as much as 78% of a researcher's time can be consumed by so called data wrangling (transforming or mapping data from one form to another). The growth of FAIR data points, including semantic models, use of established vocabularies, and linked (knowledge triplets) data, will lead to greater science impact, better research transparency and consequently strengthen the legitimacy of the science.

To tackle the non-trivial tasks involved in FAIRification of research data, researchers will require access to qualified "data stewards" who have expertise (or a network of expertise) related to policies, standards, licensing, metadata, vocabularies, semantic data modelling, domain vocabularies and ontologies, overview of metadata types and data formats, best practices for quality control, data provenance, data publishing, archiving and overseeing the optimisation for reuse of the data.

Data stewardship is instrumental in implementing the FAIR principles – and Open Science for that matter (B Mons, "Data Stewardship for Open Science", ISBN 9780815348184). As mentioned above, the list of tasks and expertise is quite significant, although it is not expected that a data steward should have all these skills or be able to personally assist in addressing the tasks required to prepare the dataset for publication. Rather, the data steward should supervise the FAIRification process and ensure that the right experts are involved in completing the process.

List of likely tasks for a data steward:

- Assist in identifying suitable metadata standards
- Locate suitable expertise for data modelling and relevant vocabularies
- Data (and metadata) quality control
- Ensure data integrity and provenance (metadata)
- Ensure archiving requirements are met and licenses issued
- Data preservation (countering data decay) and DMP (see next section) follow-up

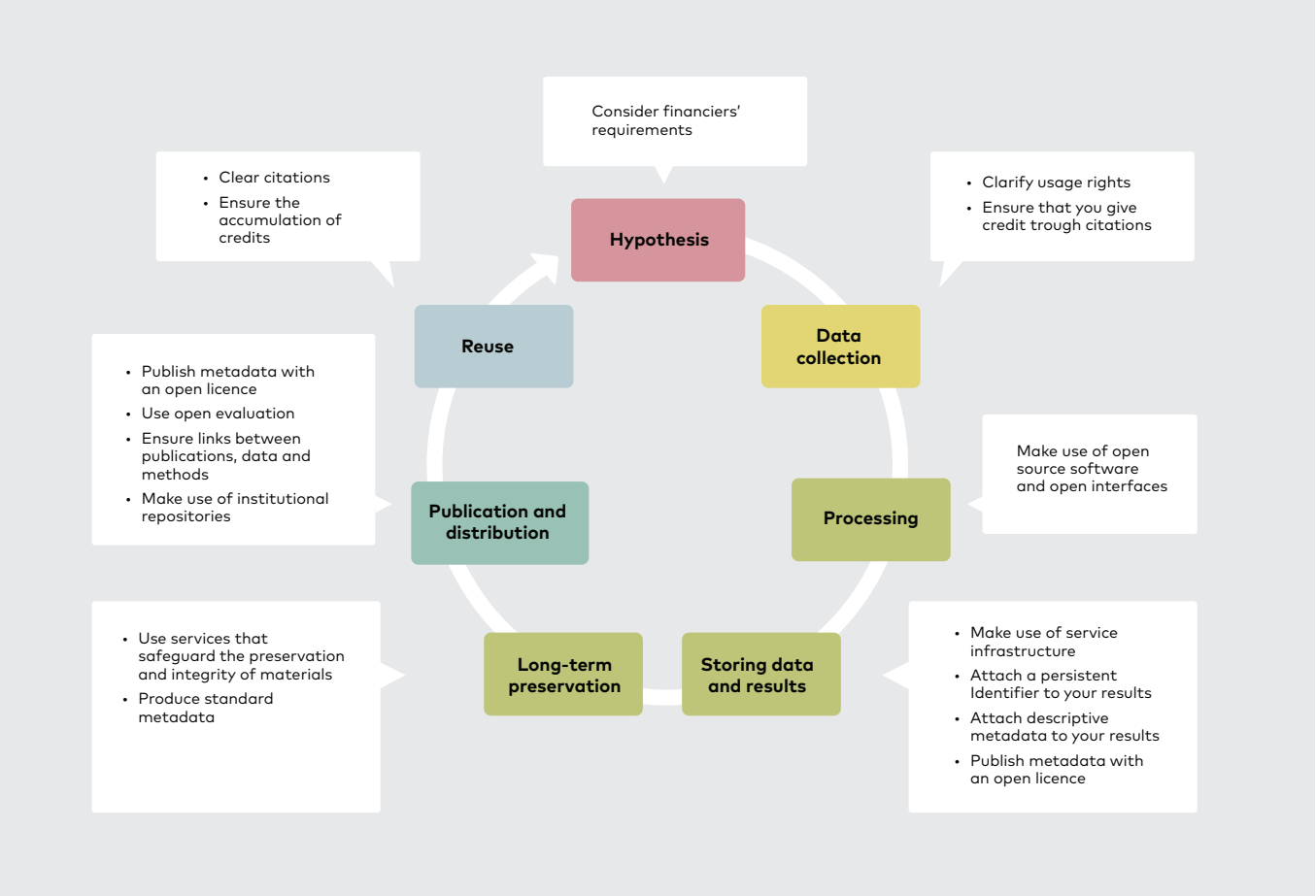


FIGURE 3 The Open Science approach according to FOSTER with notes on how to ensure openness and FAIRness during the various stages of the Data Life Cycle

For further information about data stewardship, see:
Mons, B., "Data Stewardship for Open Science", ISBN 9780815348184
USGS https://www2.usgs.gov/datamanagement/plan/stewardship.php
Peng et al. 2016, "Scientific Stewardship in the Open Data and Big Data Era", doi:10.1045/may2016-peng

07

DATA MANAGEMENT PLANS

07

Data Management Plans (DMPs) are documents that describe the procedures and practices concerning the data a researcher plans to collect or use in a project. DMPs are a means of collecting investigators' intentions and commitments relating to the metadata standards employed, data sharing, resource needs and data publication. DMPs are also excellent for arguing the need for external expertise and planning for such resources early on in the project – data stewardship resources included. It is also a helpful tool to request in connection with e-infrastructure resources, allowing providers to plan and make cost-efficient upgrades to the infrastructures in a timely fashion.

Relevant questions that a researcher or a research team should ask themselves prior to or during the ramp-up of a project are;

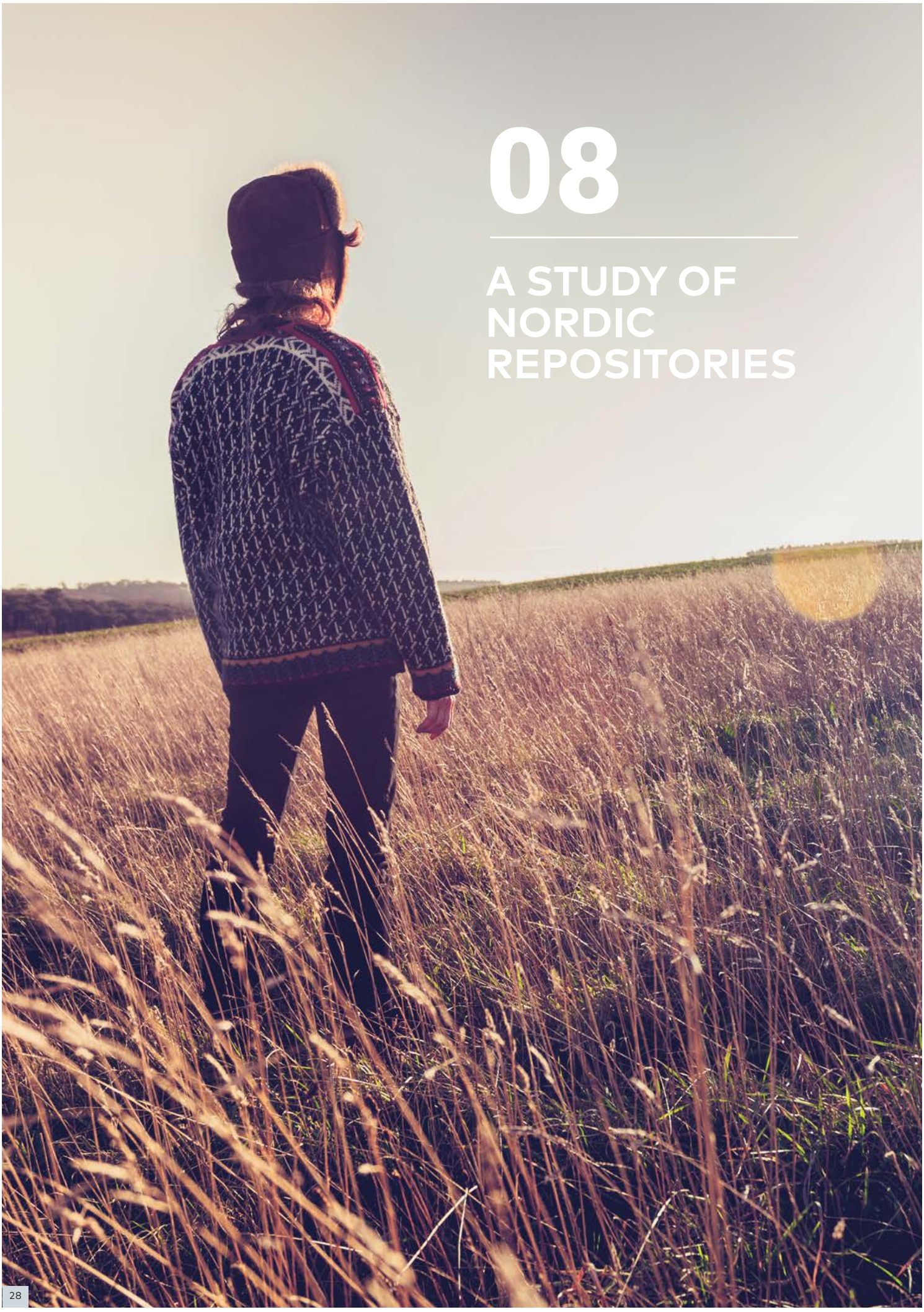
- Are there existing data policies (domain, institutional, faculty or nationwide)?
- What metadata standards or templates should be used?
- What data formats, (meta-)data sources and data rates are expected?
- Is data provenance information secured during the life cycle?
- Is the data sufficiently secure during the research project?
- Will the data be made publicly available, if so how, when and under what license?

DMPs in the form of a human-only readable document are restrictive and experience suggests that commitments made by the researchers are best verified early on and adjusted if necessary. It is expected that DMPs need to be updated several times during the project life span. To achieve dynamically verifiable commitments in DMPs it is necessary to streamline the process of documenting and quality checking the intentions set out in the document. This can be achieved using machine-actionable data management plans.

MACHINE-ACTIONABLE DMPS

Machine-actionable data management plans (maDMPs) are data management plans that consist of digital entities with linked data that can be automatically verified via online services using supported protocols. These maDMPs have a much greater potential to assist researchers in creating and executing a data management plan, in that the commitments and prerequisites are precisely specified in the form of linked metadata and can be automatically verified without the involvement of humans. In the presence of such DMPs, stakeholders may regularly generate reports that summarise the progress on commitments a project has made. The researchers, on their side, may alter the plan during the project period to reflect changes to the plan, whether these are at the strategic level or changes in the delivery timeline (e.g. due to delays in the data capturing process or problems with the analysis).

- maDMPs allow smoother planning and provisioning of services for the researchers;
- Institutions (e.g., librarians or data stewards) to provide researcher support
- Funders to monitor the commitments made by grantees (e.g. sharing of data)
- Infrastructure providers to plan their resources and provision of suitable services
- Researchers to manage, share, and discover data more easily



08

A STUDY OF NORDIC REPOSITORIES

08

Using the re3data.org repository we extracted attributes about data providers in the Nordic countries in June 2018. The selection was based on repositories associated with at least one of the Nordic countries from any field of subject.

We have eliminated duplicates and find a total of 61 repositories among the five Nordic countries. The sample is available in Table 2.

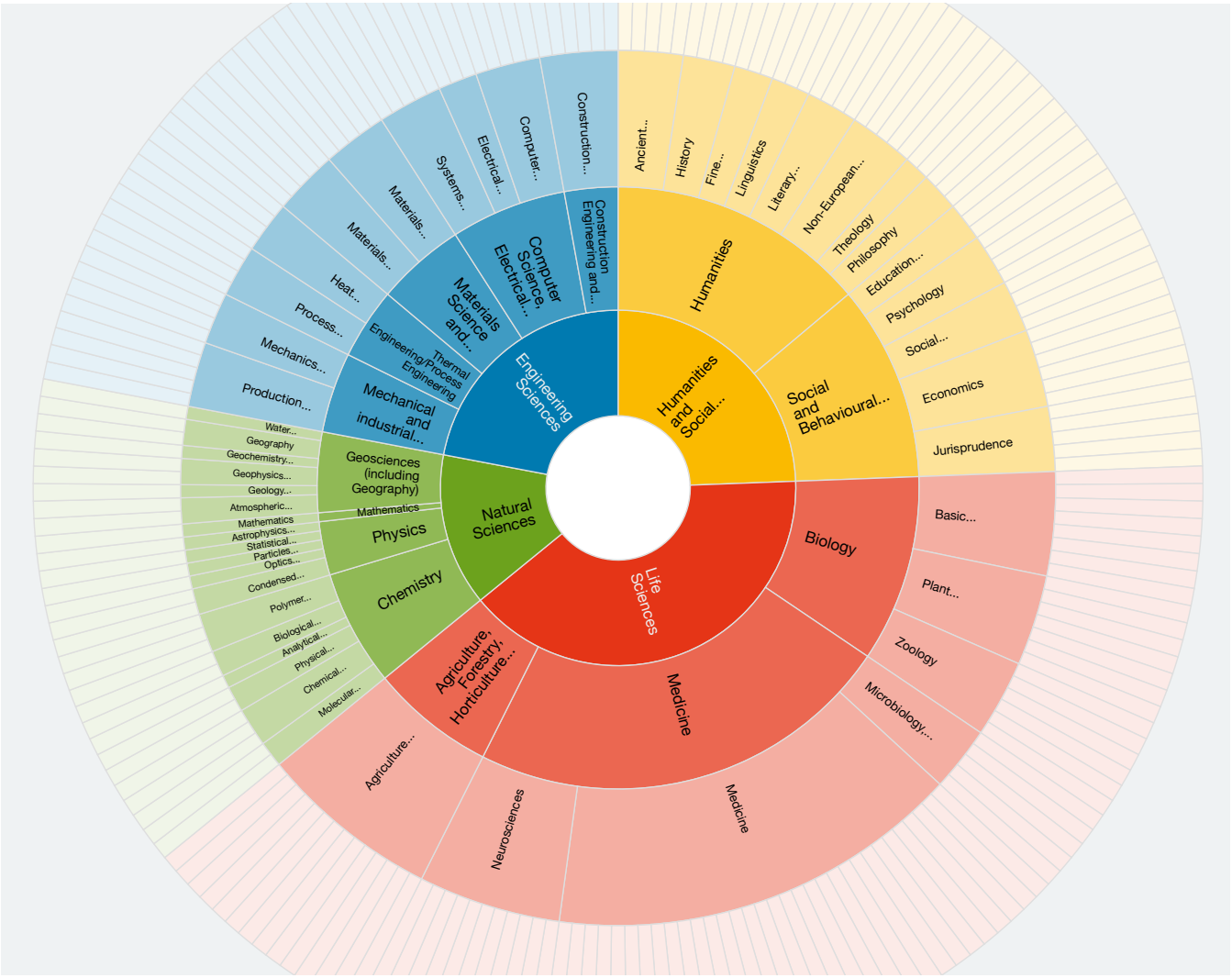


FIGURE 4 Division of major subjects of science (centre circle) and their multiple-layers of subfields.

A QUANTITATIVE SUMMARY OF NORDIC REPOSITORIES

One of the surprising findings is that very few (5%) of the 61 repositories have participation from two (or more) of the Nordic countries. Only three of the repositories had an additional Nordic country on their list of partners. This is surprising given that we expect there to be similar needs, and therefore partnering, among the Nordic countries.

The repositories providing research data in the Nordic countries were concatenated from multiple queries for each of the four major subjects (Humanities and Social Sciences, Life Sciences, Natural Sciences and Engineering Sciences) and participating country (Denmark, Finland, Norway, Sweden and Iceland). Additionally "providerTypes=dataProvider" was included in the selection criteria. Note that the data source (re3data) may contain erroneous or outdated information about some of the repositories (or even omissions). The objective is not a complete or even 100% accurate table of all the data repositories, but to obtain a representative selection that can be used to study trends and key attributes that relate to open access policies, licensing, certifications, employed metadata standards and more.

Among the most surprising discoveries is that 60% of the repositories do not issue a PID. This severely reduces the discoverability of the data. Finding a dataset without a PID will require a direct link or a mechanism to search and find the dataset (typically a repository search mechanism, which obviously does not scale). The problem with direct links is that they tend to break over time (link rot). Even with a "modest" link rot frequency of 5% per year, the majority of links will have broken in ten years.

To counter this it is strongly recommended to employ persistent identifiers (PIDs), and preferably in a form which provides guarantees of longevity and maximises discoverability (e.g. using metadata distribution). Although a minority of the repositories employ PID services, about 27% do use DOI, which is among the most trusted and relied on PID services available.

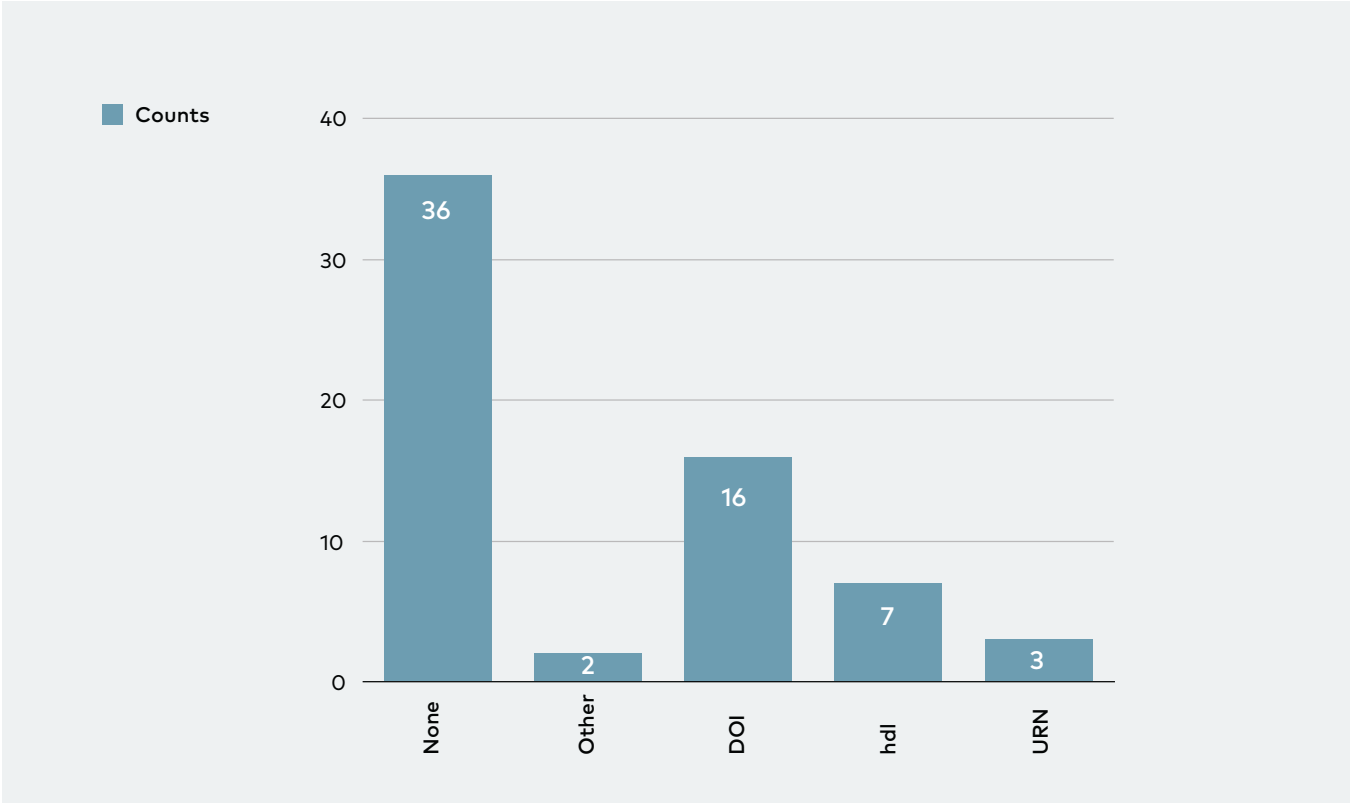


FIGURE 5 Histogram of supported PIDs for selected Nordic repositories (source re3data.org)

While almost all repositories provide open access to their metadata, it is important to recognise that the majority (70%) of those repositories do not provide open access to all their data. Typically, some of the data are shared, while some remain restricted. There may be a number of reasons for this, e.g. licensing, terms and restrictions. Furthermore, some data are likely sensitive (containing person sensitive information), in which case it is expected that the data will not be shared openly. We do not have (easily accessible) information to evaluate the degree to which sensitive data, licensing or other restrictions cause the data to have restricted access. It is also important to note that there is no conflict in requiring restricted data access and the data being compatible with FAIR (see Accessibility).

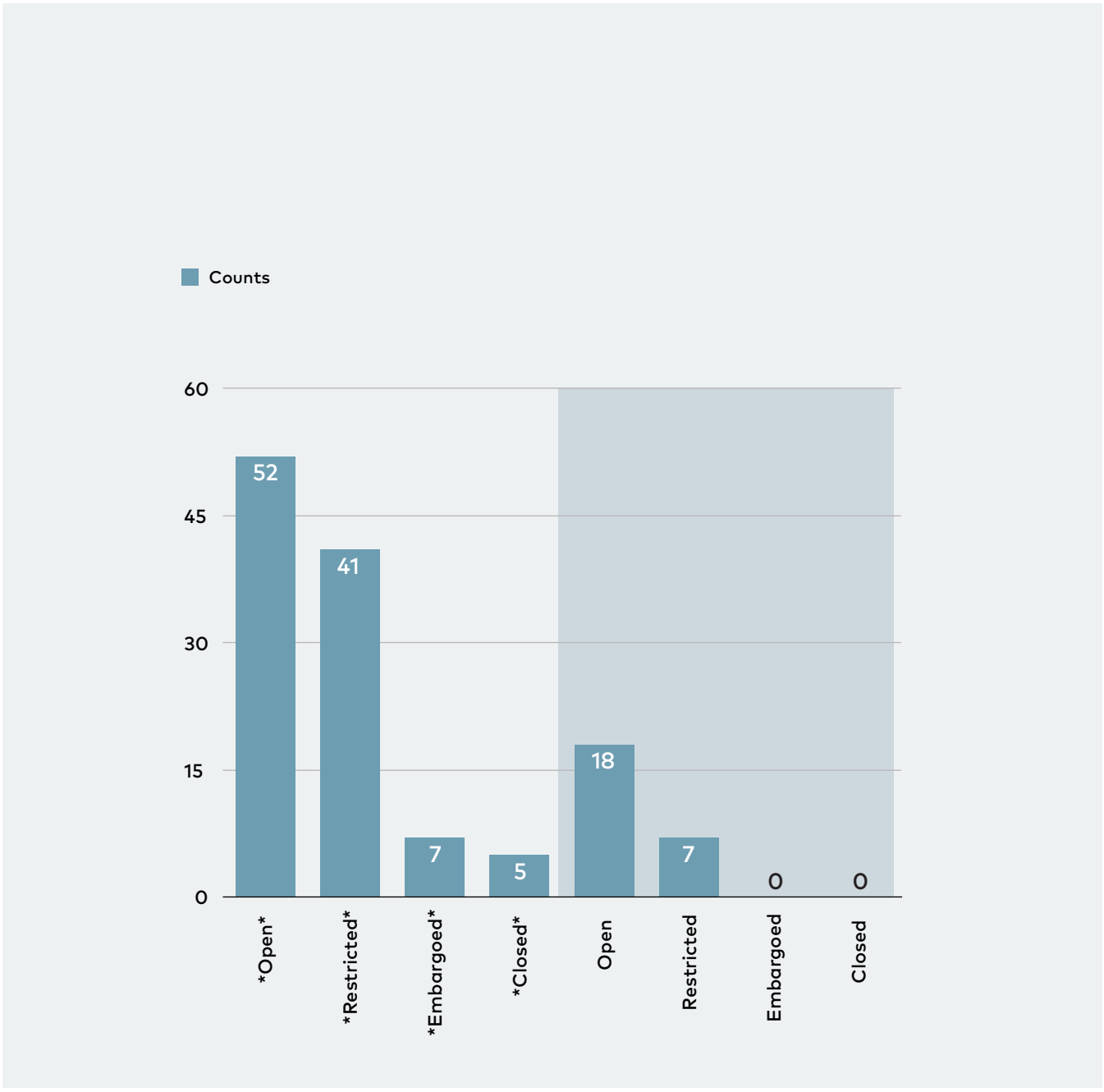


FIGURE 6 Histogram of type(s) of data access for selected Nordic repositories (source re3data.org). Columns tagged with asterix represent substrings and may consist of combinations of multiple policies. Hence, there are 18 purely "Open" and seven purely "Restricted" repositories.

Less surprising is that about 80% of the repositories are not certified or do not follow established archive/repository standards. In total eight repositories have employed / obtained the Core Trust Seal (CTS) or the associated World Data System (WDS). Another five repositories have chosen the Data Seal of Approval (DSA). A few of the repositories (five) have been awarded the Clarin-B seal.

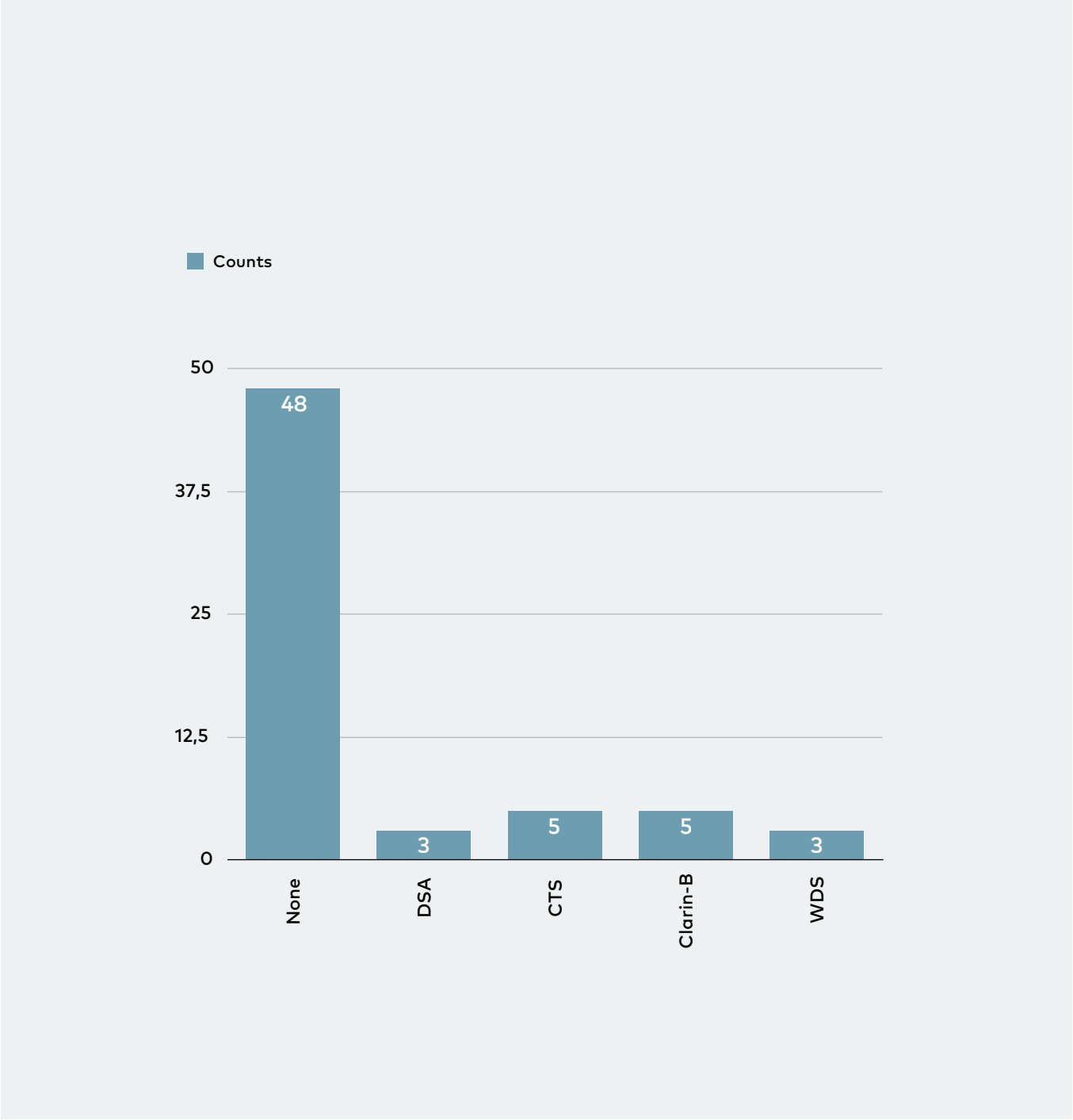


FIGURE 7 Histogram of certifications or repository standards for selected Nordic repositories (source re3data.org). Note that some repositories may have multiple entries.

Among the repositories, the majority (56%) do not employ any metadata standard. There is a large selection of standards available, some of which are more generic and less demanding to employ. A rich metadata standard (often domain specific) will typically be extensive to implement and exhaustive in coverage.

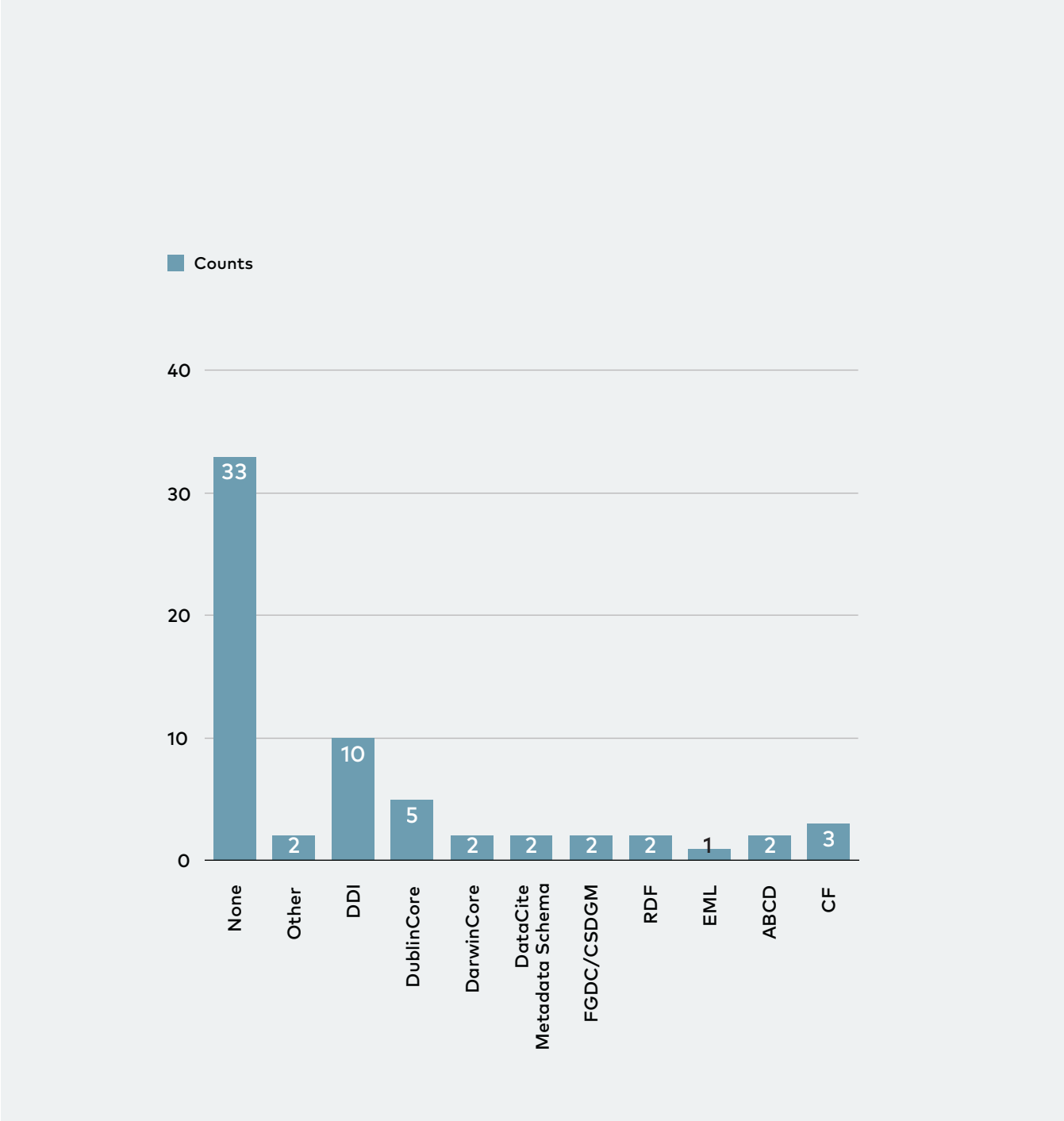


FIGURE 8 Histogram of supported metadata standards selected Nordic repositories (source re3data.org). Note that some repositories may have multiple entries.

Three major science fields are equally represented among the repositories in the sample, while the Engineering field is less frequently represented among the repositories.

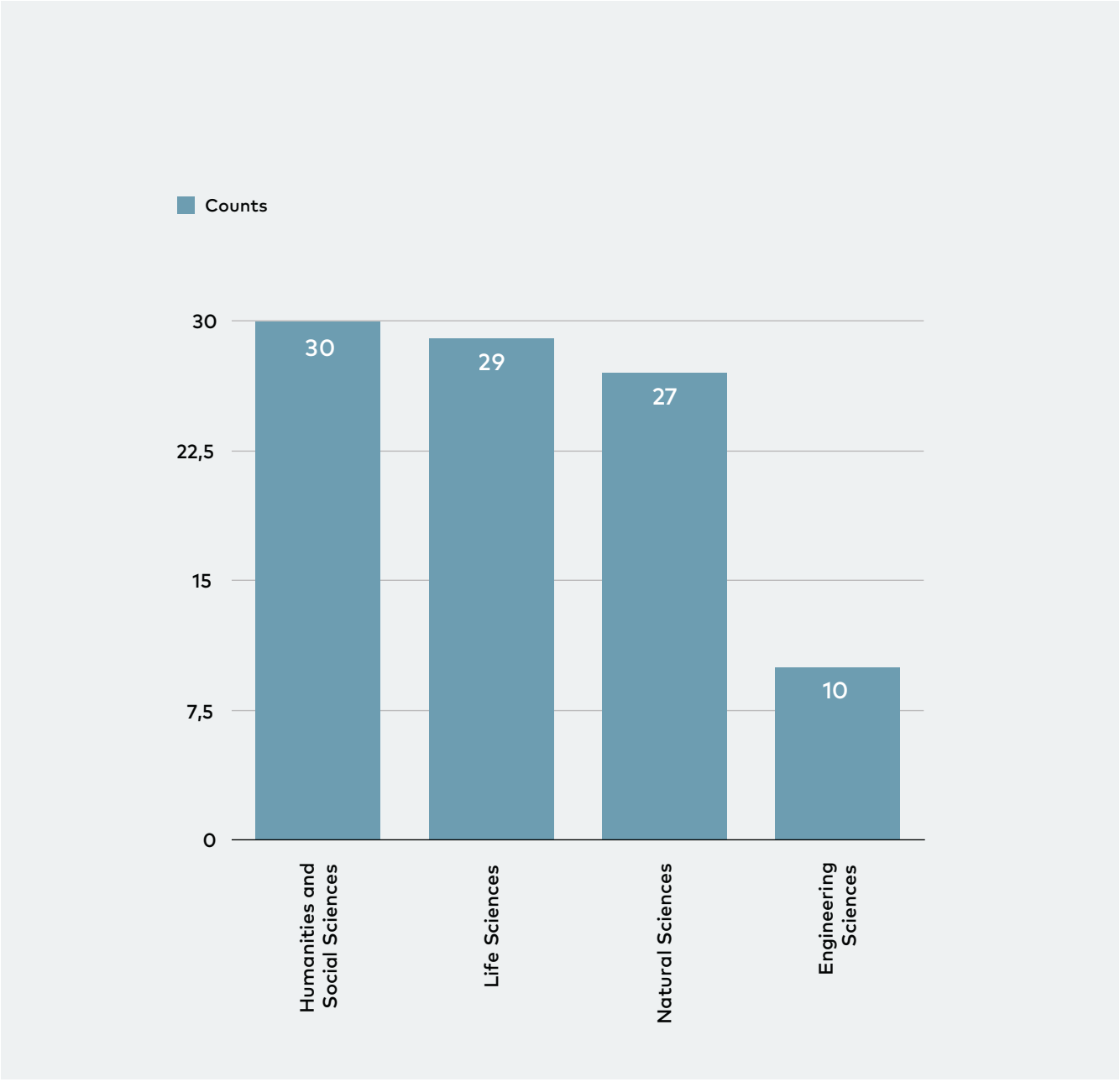


FIGURE 9 Histogram of major supported science fields for selected Nordic repositories (source re3data.org). Note that some repositories may have multiple entries.

A majority of the repositories (67%) state that they provide "other" licenses for their data, which very often means some form of restricted license. It is hard to assess the nature of these in a systematic way, so we assume these to be restricted licenses (e.g. not open). Some form of Creative Commons license (CC) is stated for 25 (41%) repositories, but it would require considerable effort to determine the exact license that is intended in each of these cases. For a sub-sample of these, nine (15%) are specified as being CC0 (all rights waived, effectively public domain). This last category can be combined with an additional seven public domain licenses, bringing the total amount of public domain-associated licenses to about 26%. Eight (13%) employ copyright licenses, although it is not clear how restrictive these licenses are (this has not been explored).

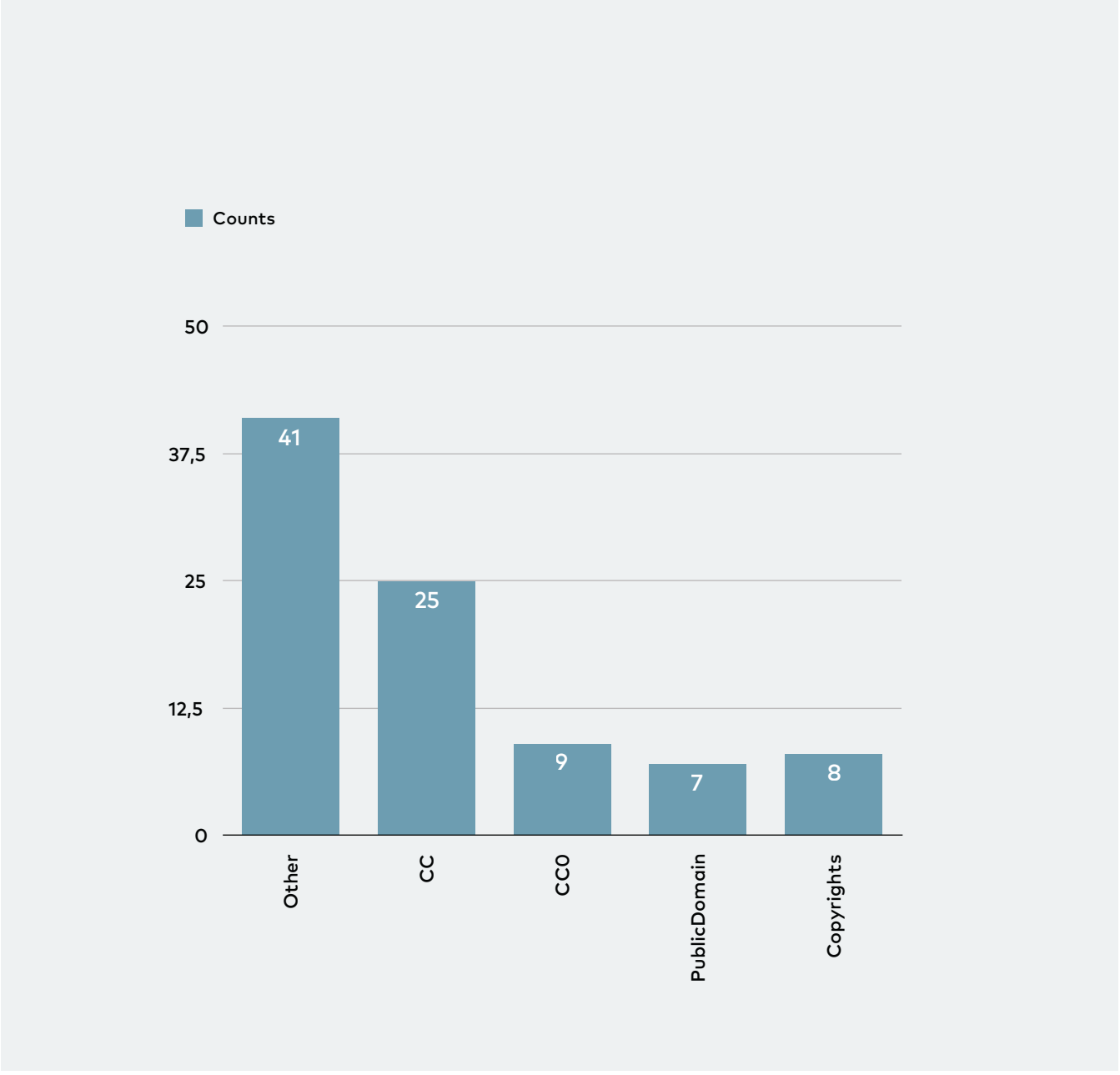


FIGURE 10 Histogram of type(s) of licenses supported for selected Nordic repositories (source re3data.org). Note that some repositories may offer a combination of multiple license agreements.

09

DEVELOPING OPEN SCIENCE IN THE NORDIC COUNTRIES

09

SUMMARY

Development of the concept of open science is still in its infancy and it will require significant effort and funding to fully realise the potential of aligning research practices in modern science with the capabilities offered by semantic metadata modelling, linked data and knowledge graphs. To get there, it is necessary to build the essential infrastructures to support this vision.

There are several opportunities for improvement in order to achieve better open science in the Nordic countries, both short-term and long-term. This report identifies the following opportunities:

I. Making legacy data findable and reusable
II. Enabling FAIR data – machine actionable protocols, templates and standards
III. Data stewardships: a fundamental pillar to aid science
IV. Training the researchers
V. Preparing for the future: FAIRification of research data

I MAKING LEGACY DATA FINDABLE, ACCESSIBLE AND REUSABLE

Goal: The very first level of actions one can initiate to address the FAIR principles for legacy data is to improve the probability that the data can be found, that it can be accessed and that is can be reused.

In their simplest form, the FAIR principles centre around the aspect of findability, accessibility and reusability of legacy data. Interoperability and certain aspects of the F, A and R principles are harder to achieve and will realistically only be applied in research groups that have made a strategic choice to produce FAIR data.

The study presented in the previous section shows that a majority (60%) of the repositories included in the survey do not offer a PID service, meaning that the data is not findable in any systematic or sustainable way. Furthermore, a little over 50% of the repositories provide open/public domain data licenses, while the rest (probably) do not. These shortcomings are relatively simple to address and one should therefore direct attention to taking advantage of these low-hanging fruits. Accessibility of data requires open access to rich metadata so that the reuse relevance of data can be assessed. The access mechanism to the data itself should be realised using standard protocols that can be evaluated in a machine actionable way.

Proposed action: It is possible to measure the compliance to FAIR using FAIR metrics (FAIRmetrics.org) in order to get a quantified measure of the degree of compliance. To increase the FAIR metrics score it is common to focus on the simplest and more trivial tasks that would lead to an improvement of the score. This has been demonstrated to be an effective strategy to improve the FAIRness of data repositories in the form of what is referred to as "hackathons". The selected participants (representatives of repositories) are invited to evaluate and calculate the FAIR metrics for their repository prior to the event and during the event work alongside colleagues to improve the FAIR metric score by addressing tasks that can be achieved in relatively short time.

II ENABLING FAIR DATA BY MACHINE-ACTIONABILITY

Goal: Establish an overview of available tools and services that support FAIR and specifically enable machine actionability (used for metadata assessment, metrics and maDMPs).

Machine actionability is a crucial component in support for FAIR metadata. It must be possible to query a dataset for metadata such as license, access protocols, etc., and it is crucial that the replies are parsable and interpretable.

A successful FAIRification of research data and relevant repositories in the Nordic countries will benefit from a larger degree of coherence for the relevant tools and services.

Categories of tools and services;
- Authentication and authorisation infrastructures
- Persistent identifiers
- Person identifiers
- Current Research Information Systems
- Data licenses
- Repository support, data access and integration

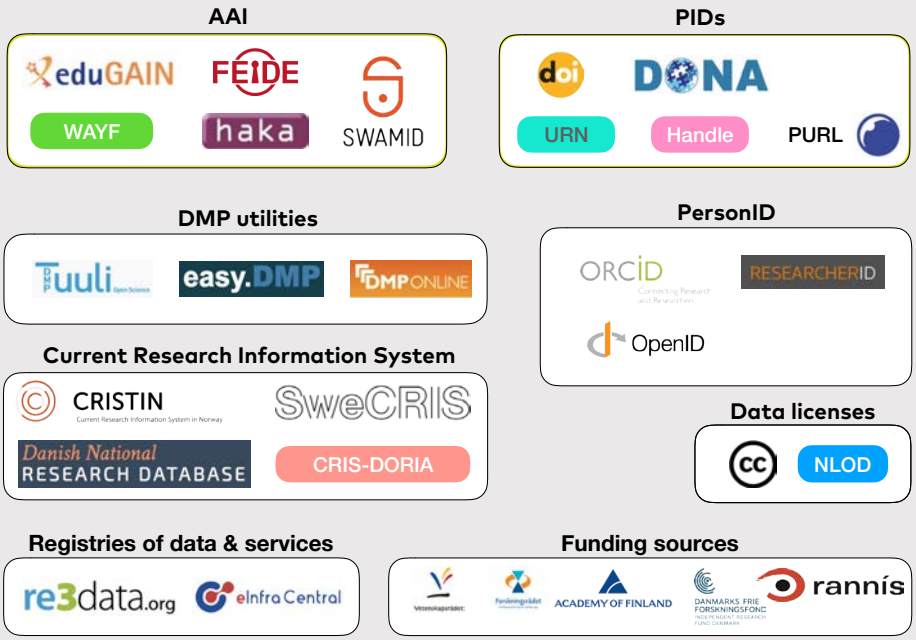
Proposed activities:
- Document requirements for protocols/PIDs (persons, resources, papers, publications, datasets)
- Encourage harmonisation if needed or build supporting platforms for Nordic maDMP executions

III DEVELOPING A NORDIC DATA STEWARDSHIP PROGRAMME

Goal: Recruit, train and fund long-term positions for data stewardship. Utilise the Nordic factor to benefit an extended network of data stewards in order to eventually realise the full vision of open science in the Nordic countries.

Proposed activities:
- Lobby the need and benefits of data stewardship and achieve high-level political support for the funding of such data experts
- Train data stewards through a series of pan-Nordic seminars, possibly in collaboration with GO-FAIR (in particular if this network is established in the Nordic regions or in some of the member countries)
- Establish community specific metadata templates in the Nordic countries
- Strive to implement heterogeneous data stewardship plans within the Nordic countries

FIGURE 11
Technologies & protocols relevant for enabling machine actionability for FAIR data and sources of administrative data



IV TRAINING THE RESEARCHERS

Goal: Develop a Nordic-wide multi-track training programme for students, researchers and data experts/ stewards with primary focus on reuse of FAIR data and the FAIRification of scientific knowledge and data (publishing science and data).

This activity can be arranged and announced among the Nordic countries and supported with contributions from GO-FAIR and other experienced offices that have key competence relevant to FAIR and open science.

V PREPARING FOR THE FUTURE: FAIRIFICATION OF NORDIC RESEARCH DATA

Goal: Identify actions needed for a successful implementation of the FAIR principles in the Nordic countries.

This activity goes beyond the "low-hanging fruit" goals discussed in Action I. Here we build on Action III and the focus is on semantic modelling using community-specific vocabularies and ontologies, the creation of linked data and FAIR data points.

Notes:
FAIR sharing (standards, databases and policies): https://fairsharing.org/
FAIR metrics: http://fairmetrics.org/

A REPORT BY

ANDERS O. JAUNSEN
ON BEHALF OF NEIC
CONTACT: AJAUNSEN@GMAIL.COM

Adress:
c/o NordForsk
Stensberggata 25,
NO-0170 Oslo

Phone: +47 476 14 400

E-mail:
kine.nordstokka@nordforsk.org
gudmund.host@nordforsk.org