RAVNETRYKK

UiT goes open: Et festlig skrift til Stein Høydalsvik

DataverseNO: A National, Generic Repository and its Contribution to the Increased FAIRness of Data from the Long Tail of Research

Philipp Conzett

Abstract

Research data repositories play a crucial role in the FAIR (Findable, Accessible, Interoperable, Reusable) ecosystem of digital objects. DataverseNO is a national, generic repository for open research data, primarily from researchers affiliated with Norwegian research organizations. The repository runs on the open-source software Dataverse. This article presents the organization and operation of DataverseNO, and investigates how the repository contributes to the increased FAIRness of small and medium sized research data. Sections 1 to 3 present background information about the FAIR Data Principles (section 1), how FAIR may be turned into reality (section 2), and what these principles and recommendations imply for data from the so-called long tail of research, i.e. small and medium-sized datasets that are often heterogenous in nature and hard to standardize (section 3). Section 4 gives an overview of the key organizational features of DataverseNO, followed by an evaluation of how well DataverseNO and the repository application Dataverse as such support the FAIR Data Principles (section 5). Section 6 discusses how sustainable and trustworthy the repository is. The article is rounded up in section 7 by a brief summary including a look into the future of the repository.

1. The FAIR Data Principles

Data constitute the core assets within many scientific disciplines. New knowledge and insight are often drawn from the analysis of data. However, in traditional scholarly communication, research data have not been granted the attention one would expect given their importance for the advancement of science. Among the different kinds of output from research activities, describing research results in articles and books published in recognized venues is still the most rewarding way of communication for the vast majority or researchers. Data, on the other hand, are – with the exception of some fields – rarely published and shared. This situation is quite surprising considering that in most cases scientific results and claims cannot be verified without access to the underlying data. There are also other unfortunate consequences of data not being made accessible, among

https://doi.org/10.7557/15.5514

© Philipp Conzett. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International license.

others the fact that data cannot be reused by other researchers, which potentially results in duplication of efforts; and that researchers may miss important possibilities for new collaborations arising from shared data.

Research funders and other stakeholders have for quite some time been aware of the unfortunate consequences of research data not being reusable, and the urge to improve reusability of data has been continuously growing as research methods and tools have become increasingly digital. As one of the influential international stakeholders, the Organisation for Economic Co-operation and Development (OECD) addressed the problem of reduced reusability as early as in 2004, when they adopted a declaration on access to research data from public funding. This declaration resulted in the OECD Principles and Guidelines for Access to Research Data from Public Funding, which was published in 2007 (OECD, 2007). Similar guidelines and recommendations have since been adopted by other funding agencies and research organizations (Christian et al., 2020; Crosas et al., 2018; Neylon, 2017). Drawing on this early work, one of the most influential papers in the field of data management saw the light of day in 2016, when a diverse group of stakeholders representing academia, industry, funding agencies, and scholarly publishers postulated a set of principles on how to improve infrastructure supporting the reuse of research data. They referred to these principles as the FAIR Data Principles (Wilkinson et al., 2016).

FAIR data are data that are Findable, Accessible, Interoperable, and Reusable. A key feature that according to the authors distinguishes the FAIR Data Principles from similar initiatives is their emphasis on machine-actionability, meaning that not only humans should be able to find, access and reuse research data, but also machines. While the results from data analysis ultimately are to be understood and interpreted by humans, machine-actionable research data can increase the efficiency of data management and analysis considerably. Also, due to the nature and/or amount of data in some scientific fields, machine-actionability is not only a question of effectiveness, but simply a necessity. In their seminal article, Wilkinson et al. (2016, p. 4) summarize the FAIR Data Principles as follows:

> *To be Findable:*
> F1. (meta)data are assigned a globally unique and persistent identifier.
> F2. data are described with rich metadata (defined by R1 below).
> F3. metadata clearly and explicitly include the identifier of the data it describes.
> F4. (meta)data are registered or indexed in a searchable resource.
>
> *To be Accessible:*
> A1. (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2. metadata are accessible, even when the data are no longer available.

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with detailed provenance.
R1.3. (meta)data meet domain-relevant community standards.

The implication of the different parts of the FAIR Data Principles will become evident from the detailed review presented in section 5.

As summarized in Figure 1, FAIR research data may be considered as both the output of and the input to good data stewardship throughout the entire lifecycle of research data. Let us start with the processes illustrated with the orange arrows. Early in a research project, data are collected or generated, and possibly processed (e.g. annotated or enriched in other ways). During this active phase of the research project, it is essential to have good routines in place for organizing and describing the data as well as for data storage. Processed data are then analysed, and the findings from data analysis are usually presented in articles or books. At the same time, the background data for these publications should be archived and shared with the scientific community and the greater public. An important step before data sharing is to document the data to enable other researchers to understand and reuse them. The grey box at the top of the figure indicates that all these activities should be carried out in line with policies, standards and good practice recommendations for research data management, and that good planning lays the foundation for successful data management. However, contrary to widespread belief, planning is not a one-off task to get done with at the outset of a research project. Rather, a data management plan (DMP) only reveals its full potential when used as an active document that is updated and revised throughout the project. In Figure 1, this is indicated by the multiple grey arrows below the grey box. The ideal output of the research activities described so far are FAIR research data, illustrated in the green arrow-box.

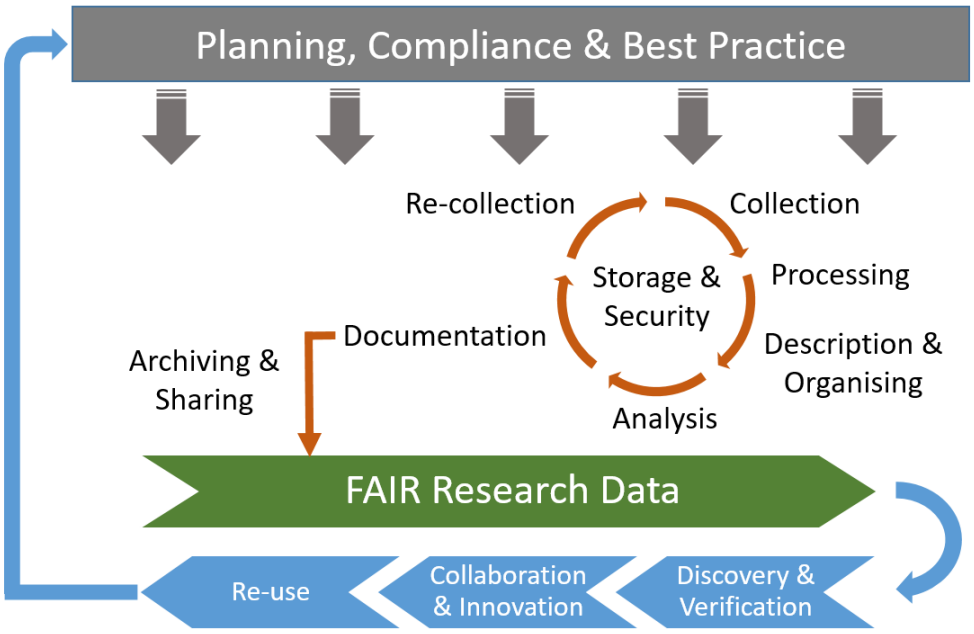

Figure 1: The FAIR Research Data Lifecycle.

These data may, in turn, serve as input to other projects, illustrated by the blue arrows and arrow-boxes. FAIR research data can be discovered, and they may be used to verify the scientific findings and claims that are set forth based on them. Discoverable data is a driving force for new collaborations between researchers as well as innovation within the scientific community, but also in society more broadly. In more general terms, FAIR research data may be reused, which brings us to the beginning of the next round of the lifecycle of research data. Considering potential reuse of data should be a natural part of the planning of any research project in need for empirical evidence.

2. How to turn FAIR into reality?

The FAIR Data Principles are a set of general guidelines, they do not include specific advice on how these principles may or should be implemented. Many research support services provide guidance for researchers and others involved in research data management on how to organize and document their data to make them reusable. There are also numerous high-level initiatives working for the uptake of the FAIR Data Principles on a more global scale. The most prominent example in Europe is probably the establishment of the European Open Science Cloud (EOSC) (European Commission, n.d.).

As has been pointed out by several experts, FAIR is not a binary concept. Instead, data – or more broadly: digital objects – may be more or less FAIR, and we can thus speak of degrees of FAIRness. Recommendations on FAIR data management are thus meant to support a transition towards increased FAIRness of research data at large.

Ultimately, the transition towards increased FAIRness requires change in research culture. Brian Nosek uses two theories to illustrate the mechanisms and progress involved in research culture change. The first one is the Theory of Diffusion of Innovation, proposed by Everett Rogers in 1962 (Rogers, 2003). Nosek suggests that the normal distribution graph used in Rogers' theory and redrawn in Figure 2 also applies to the spread of change in research culture (Nosek, n.d.).

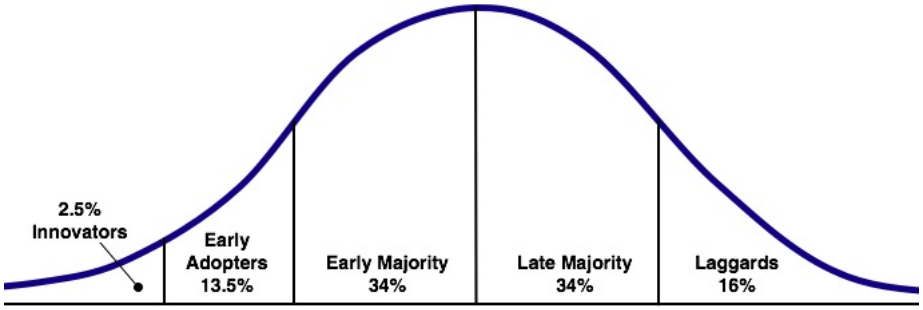

Figure 2: The Diffusion of Innovation. Redrawn from Rogers (2003, p. 281).

According to this view, a change or innovation in research culture is first initiated and spearheaded by a small group of innovators, followed by a somewhat larger group of early adopters, before the change is adopted by an early majority and afterwards by a late majority of members of the research community. Finally, the change spreads to the last group of adopters, called laggards in Rogers' theory.

The second theory adopted by Nosek is the Theory of Human Motivation, originally proposed in 1943 by Abraham Maslow (Maslow, 1943). Maslow postulates a hierarchy of needs to illustrate how human behaviour is governed by motivation. Adapting Maslow's hierarchy of needs to the realm of research, Nosek uses the pyramid in Figure 3 to illustrate the different motivational factors and driving forces behind cultural change. In this view, basic infrastructure including tools and skills are necessary to make change in research culture possible. Turning this infrastructure more user-friendly makes it easy for members of the research community to adopt new practice. Once new practice has spread to and is recommended by (a large part of) a research community, its adoption may be considered normative. As a further step in advancing the uptake of new practice, incentives may be introduced that make the adoption of the practice rewarding. Finally, the implementation of new practice may be made required by policies.

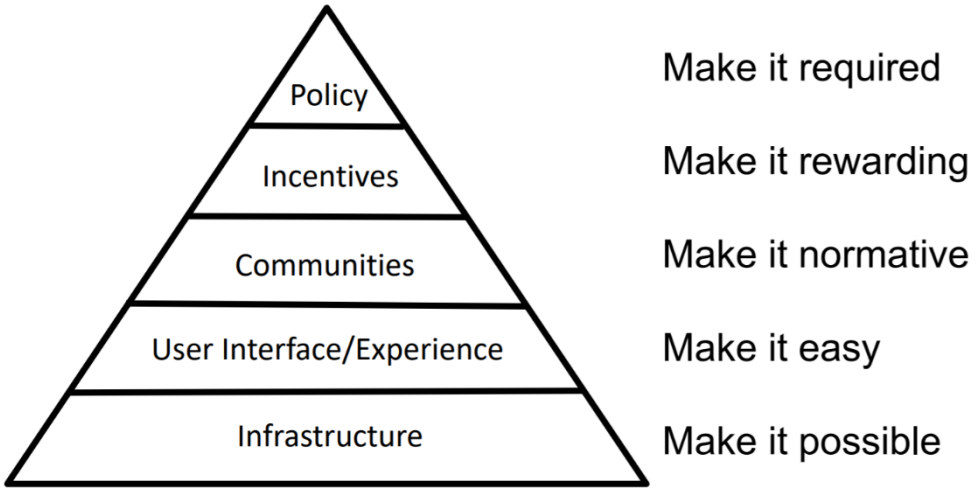

Figure 3: Motivational factors and driving forces behind change of research culture. From Nosek (n.d.), licensed under CC0 1.0 Universal.

In Figure 4, Nosek combines his adaption of Rogers' Theory of Diffusion of Innovation and Maslow's hierarchy of need. In this view, basic infrastructure making the change of research culture possible is sufficient motivation for the small group of innovators to adopt new practice. Early adopters go along with the innovators once adoption has been made easy. Adoption by the early majority is mainly driven by the new practice becoming part of community norm, while rewarding incentives are the main motivation needed for the late majority to change their behaviour. Finally, policy requirements seem to be the last resort to motivate the group of laggards to comply with what by then probably is recognized as a de facto standard in the research community.

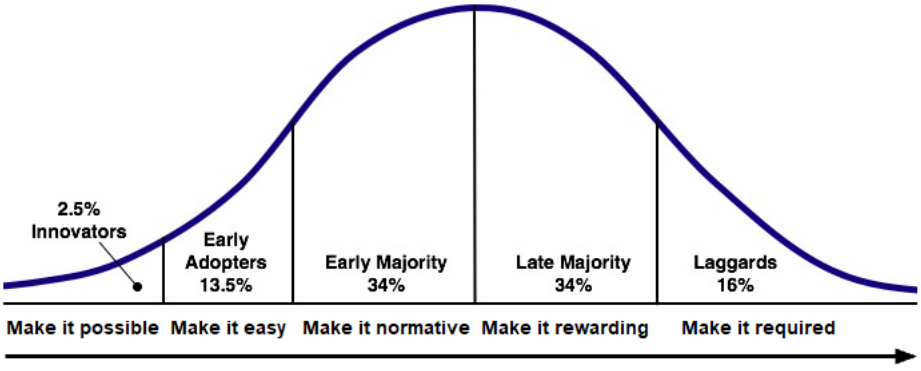

Figure 4: Change in research culture: diffusion and motivation. Slightly adapted from Nosek (n.d.), licensed under CC0 1.0 Universal.

To my knowledge, Nosek's view of how change in research culture diffuses in research communities has not been verified by empirical data. However, applied to the process of increasing the FAIRness of research data, there are some observations I have made in my work with providing support services for research data management for the last six years which to some degree substantiate Nosek's view.

First, there are parts of research communities that started adopting the FAIR Data Principles without there being other resources in place

than the basic infrastructure making the adoption possible. Parts of the bioinformatics community may here serve as an example.

Second, for quite some time researchers have had access to data repositories and other support services making it rather easy to make their data at least partly FAIR. Nosek suggests that – seen from a post-hoc perspective once the culture change has been accomplished – the group of early adopters amounts to 13.5% of the research community. To get a rough indication of how large the group of early adopters of the FAIR Data Principles currently might be in Norway, one could compare the number of unique authors of published data-sets with the number of unique authors of publications of research results in anthology chapters, articles and monographs (books). In a small case study, I obtained these numbers limited to researchers affiliated with my own university, UiT The Arctic University of Norway (UiT), and limited to outputs published in 2019. The background data for this small investigation including a description of the methods and tools used to obtain these numbers are available in Conzett (2020). There were 20 unique UiT-affiliated researchers who published one or more datasets in 2019. In the same year, there were 1736 unique UiT-affiliated researchers who published research results in recognized publishing venues. The group of UiT-affiliated researchers who published data in 2019 may thus be said to represent 1.15% of all UiT-affiliated researchers who published research results in the same year. These numbers are based on a very small selection of researchers and limited to the span of only one year. I still argue that they can give us a rough indication. The results from this small investigation suggest that the percentage of researchers making use of easily accessible resources to make data more FAIR is still much lower than the 13.5% indicated by Nosek for the group of early adopters.

Third, there seems to be considerable agreement among research data professionals that strong support and recommendations from within the research community as well as rewarding incentives are crucial to advance the further uptake of the FAIR Data Principles. On the other hand, the effectiveness of policy requirements may vary somewhat. Some research funders and journals have introduced policies that require researchers to make their data openly available. Such policies are most effective when non-compliance results in direct negative consequences like repayment of funding or the pre-clusion of article or book manuscripts from being accepted for review. Journals' data policies are thus an effective incentive for researchers to make their data available. Currently, most universities do not reinforce the compliance with their research data policies. However, institutional policies are still important as they help provide useful legitimation for institutional support services (see for example Figenschou in this volume).

In addition to the driving forces described above also other factors affect the way in which the FAIR Data Principles may be turned into

reality. The properties of research data themselves is one such factor, which is the topic of the next section.

3. Data from the long tail of research

Research data come in different forms, and this diversity may pose some challenges to responsible data stewardship. An overall – though somewhat simplified – distinction commonly made is the one between big data and small data (Borgman, 2015, pp. 8–10). The distribution of scientific work along the continuum between big and small data has been described as a head with a long tail, meaning that a small number of research projects deal with very large volumes of data, whereas the vast majority of researchers work with medium-sized or small volume data. The distribution is illustrated in Figure 5.

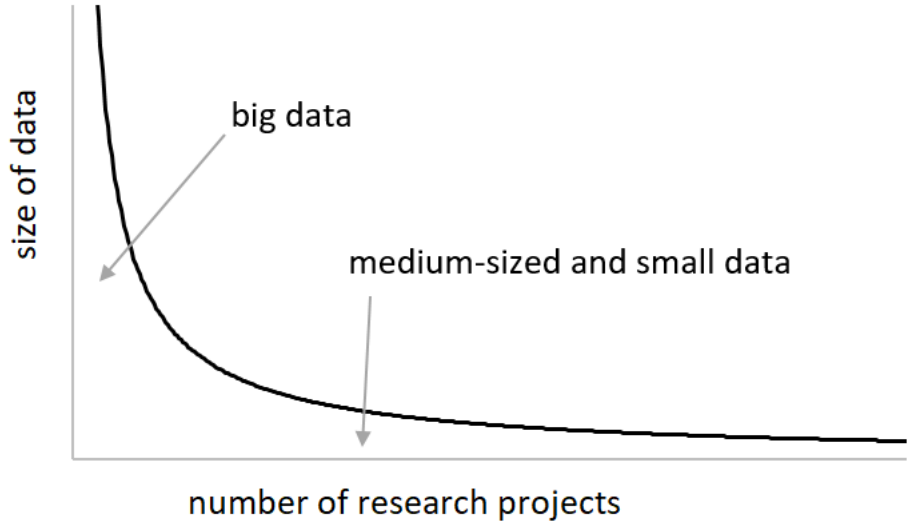

Figure 5: Distribution of research projects according to data size.

On a generalised level, the distribution shown in Figure 5 is also said to reflect properties of research data other than data volume or size (Borgman, 2015, pp. 8–10; Heidorn, 2008; The e-IRG Task Force on the Long Tail of Data, 2016). Big data tend to be homogeneous in content and form, whereas small datasets often represent a multitude of data types. Big data are often handled within large projects, which makes it more urgent to agree on common standards for data structure, file formats, documentation etc. at an early stage of the project. In such projects, it is also common to share infrastructure, tools and services like data curation. In small or less data-driven fields and/or in small or medium-sized projects, researchers can more easily adapt data management to the problems they are set to solve, without putting the progress of the project at risk. "The down-side to such flexibility is" according to Borgman (2015, p. 10) "the lack of standards on which to base shared infrastructure and the lack of critical mass to develop and sustain shared data resources". In other words, the further down the tail of the distribution in Figure 5 a research project may be placed, the more diverse the data are in content, structure and representation, and the less common are

shared standards and resources to support uniform data management (Borgman, 2015, p. 10).

With the rapid development of computing capacity and technology, big data have received much attention from research funders and from society at large during the last decade or so. However, as emphasized by Borgman (2015, pp. 8–9), "only a few fields, such as astronomy, physics, and genomics in the sciences; macroeconomics in the social sciences; and some areas of digital humanities; work with very large volumes of data in an absolute sense". Although the enormous value of small and medium-sized data is recognised by advocates of Open Science and cross-disciplinary research, the e-IRG Task Force on the Long Tail of Data (2016, p. 4) argues that the importance of long-tail data had become out of sight of funders and policy makers with the advent of Big Data. This view is supported in a 2018 report from the European Commission expert group on FAIR data: "The so-called 'long tail' of research remains poorly catered for, and vast amounts of data produced in research are not FAIR and currently lack long-term stewardship" (European Commission, 2018, p. 55). In the work to remedy this shortcoming, the design principles and organizational models of research data repositories are given a crucial role (The e-IRG Task Force on the Long Tail of Data, 2016, p. 4). Traditionally, a fair amount of the data from the long tail of research that are made available, find their home in general or generic research data repositories, one of them being DataverseNO.

4. What is DataverseNO, and how does it work?

DataverseNO (re3data.org, 2017) is a national, generic repository for open research data. The repository is owned and operated by UiT The Arctic University of Norway. The technical infrastructure of the repository is based on the open source application Dataverse (n.d.), which is developed by an international developer and user community led by Harvard University. DataverseNO supports the FAIR Data Principles and has recently been certified as a sustainable and trustworthy repository by the CoreTrustSeal (n.d.).

Established in 2017, DataverseNO has its origin in UiT's institutional research data repository, UiT Open Research Data (n.d.), which was launched in 2016. Before that, UiT had been developing and running support services for research data management for a couple of years (Conzett & Østvand, 2018). Both UiT Open Research Data and DataverseNO have grown out of the Tromsø Repository of Language and Linguistics (re3data.org, 2015), the first research data repository service established at UiT. The Tromsø Repository of Language and Linguistics (TROLLing) is a domain-specific repository that was initiated back in 2013 by linguists at UiT and has since its launch in 2014 been an open and free repository where linguists worldwide can deposit and publish their research data (Conzett, 2019; GÉANT & UNINETT, 2019). TROLLing as well as UiT Open Research Data are now part of DataverseNO constituting each its own collection.

All datasets deposited into DataverseNO are curated by research data support staff before they are published. Published data can be edited. Any changes are subject to a new curatorial round and result, on publication, in a new version of the dataset with all previously published versions still being available.

Figure 6 gives an overall outline of the organization of DataverseNO. The repository structure of DataverseNO is outlined in the middle box in the chart. Norwegian research organizations may enter into a partner agreement with UiT to use DataverseNO as an institutional repository for open research data. Datasets from researchers affiliated with DataverseNO partner institutions are published in designated institutional collections, of which there are currently nine. Individual researchers from Norwegian research organizations that are not partnering with DataverseNO can publish their data in the top-level collection of the repository (indicated with Dataset 10, 11, … in the repository structure box in Figure 6). Another type of collection within DataverseNO are special collections, which may be project-based and/or subject-based. Such collections may be open for contribution from researchers outside Norwegian organizations. TROLLing is a thematic collection, and currently the only special collection in DataverseNO. All special collections within DataverseNO are at the full responsibility of a DataverseNO partner institution.

As indicated in the top box in the DataverseNO Organization Chart, a set of organizational documents regulate the organization of the repository, including its structure, governance, data curation, and its Designated Community. At the core of these documents we find the DataverseNO Policy Framework (n.d.), which – in addition to a general part introducing the scope and mission of the repository as well as defining some core concepts – consists of four policies, namely, the DataverseNO Access and Use Policy, the DataverseNO Accession Policy, the DataverseNO Deposit Agreement, and the DataverseNO Preservation Policy.

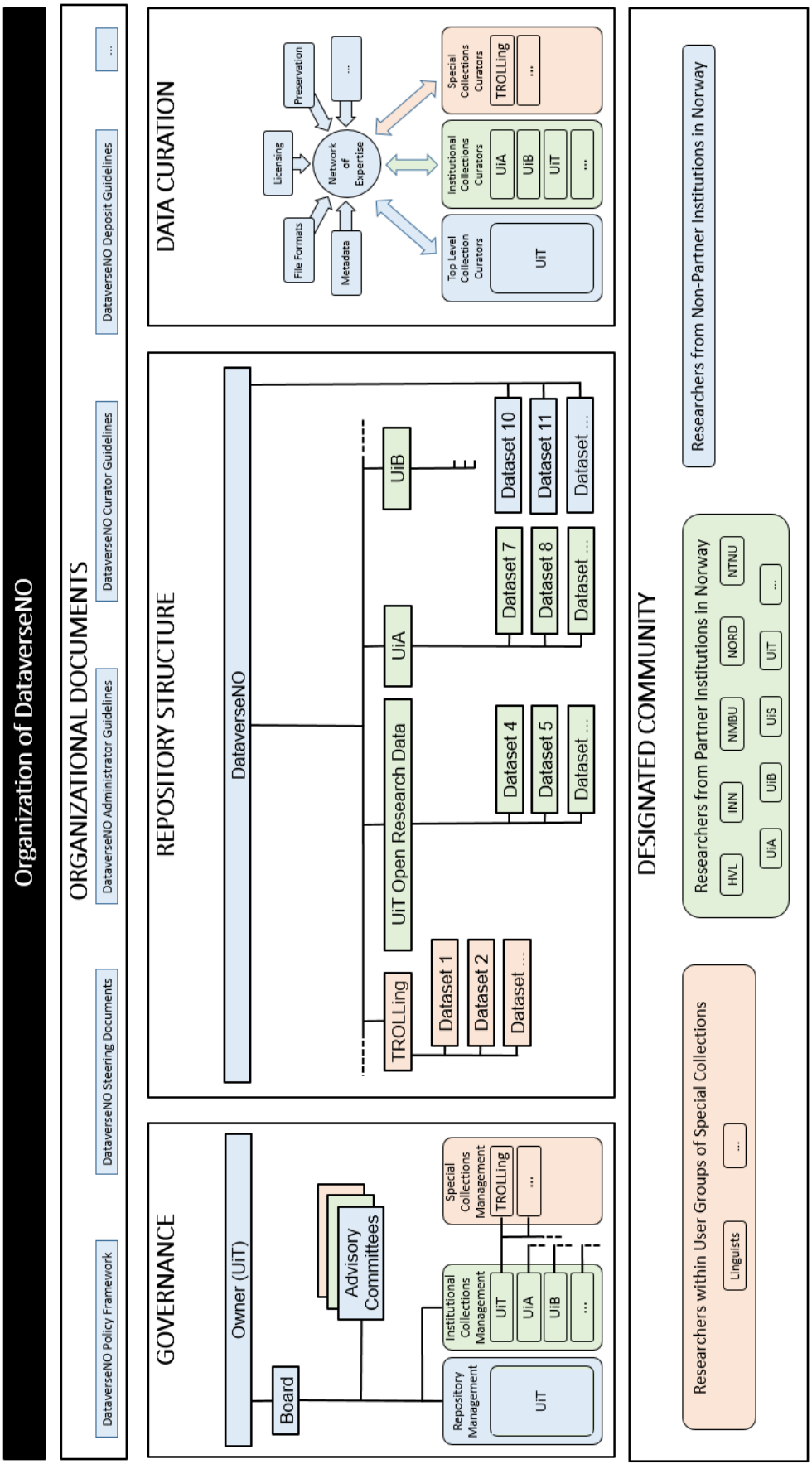

Figure 6: DataverseNO Organization Chart. From Organization of DataverseNO (n.d.), licensed under CC0 1.0 Universal.

The DataverseNO Access and Use Policy outlines DataverseNO's commitment to facilitating maximum access and use of research data published in the repository, and describes the mechanisms DataverseNO implements to fulfil this commitment, including the provision of persistent URLs and the assignment of Digital Object Identifiers (DOIs) for reliable discovery and citing of published Datasets, and facilitating the indexing of descriptive metadata by search engines. *The DataverseNO Accession Policy* explains what DataverseNO can accept for publication, which is essentially research data that can be made openly available, are appropriately documented, and are stored in preferred file formats. With the exception of special collections like TROLLing, DataverseNO accepts only data from researchers affiliated with Norwegian research organizations. At initial sign up, depositors are required to accept the *DataverseNO Deposit Agreement*, which gives DataverseNO a non-exclusive right to distribute published data on the internet, and the right to copy and convert published files to new file formats to the extent required to ensure secure storage and sustainable long-term preservation. *The DataverseNO Preservation Policy* describes DataverseNO's commitments and approaches to responsible and sustainable stewardship of published datasets in the long term, including preservation strategies and levels of preservation. The concrete preservation actions are specified in the DataverseNO Preservation Plan.

To implement the commitments stated in the DataverseNO Policy Framework, the DataverseNO Guidelines provide concrete guidance for the three main stakeholder categories involved in the operation of the repository. The DataverseNO Deposit Guidelines (n.d.) describe how depositors are required to prepare their data prior to deposit, and how to create and get datasets published in DataverseNO. The DataverseNO Curator Guidelines (n.d.) contain documentation about how curators are expected to review deposited data before publication. The DataverseNO Administrator Guidelines describe the main tasks to be carried out by repository as well as collection administrators, including user and access control management.

The DataverseNO Partner Agreement regulates the responsibilities of UiT as the owner and operator of the repository on the one hand, and of the partner institutions on the other hand. Most importantly, the agreement states that all DataverseNO policies and guidelines apply to the entire repository with all its collections. Partner institutions are thus obligated to manage their institutional collections in compliance with the DataverseNO policy framework.

Data curation is carried out by data curators at the DataverseNO partner institutions, as indicated in the middle right box in Figure 6. Datasets deposited in the top-level collection are curated by data curators at UiT. Data curators are staff members employed at the partner institutions, usually at the library of the institution. Data curators are responsible for ensuring that data published in each collection within DataverseNO (including the top-level collection) are

curated according to the DataverseNO policies and guidelines, and in line with best practice recommendations and the needs of the different user communities at stake. Curators communicate with the different user communities represented in the collection(s) they curate, e.g. during curation, but also through other channels and in other venues. Curators also communicate with the management of their collection, and with curators of other collections within DataverseNO through the DataverseNO Network of Expertise. This network of curators covers the different aspects of data curation, including metadata and documentation, file organizing and file formats, and licensing. In addition to enabling knowledge and experience exchange, this network also makes sure that curation practices across the repository are aligned with the DataverseNO policies and guidelines. The network also seeks to align curation practices across institutional collections from different partner institutions containing data from the same or similar scholarly disciplines.

The governance of DataverseNO is illustrated in the middle left box in Figure 6. As the owner and operator of DataverseNO, UiT offers the repository as a service to other research organizations and to individual researchers from research organizations in Norway.

The Board of DataverseNO has the overall responsibility for DataverseNO, with a mandate provided by the university management of UiT.

Collections within DataverseNO may have their own advisory committees which give advice to the collection managers as well as to the Board of DataverseNO on high-level aspects of the operation and development of a specific collection as well as the entire repository. The Designated Community may raise high-level or general issues with representatives from the advisory committee of the collection at stake. Currently, only TROLLing has formally established an advisory committee, the TROLLing Scientific Advisory Board, who provide their advice to the TROLLing managers.

The operation of institutional collections is embedded in the research support services and endorsed by the institutional management at the DataverseNO partner institutions. At each partner institution, there are procedures and venues in place where research support units, such as the university library, discuss issues with representatives from the different research communities at the institution. Feedback from such discussions is provided to the managers of the institutional collections. On their part, managers for institutional collections discuss advice and feedback from the user groups of their institutional collections in the Advisory Committee for DataverseNO. This committee, illustrated with the blue box in the middle of the GOVERNANCE section of the DataverseNO Organization Chart, consists of representatives from all DataverseNO partner institutions (usually the collection managers), and the managers of DataverseNO. The members of the DataverseNO Advisory Committee meet at least

twice a year to discuss issues concerning the organization of DataverseNO, including governance, policies and guidelines, repository structure and operation (including functionality), data curation, and issues raised by the Designated Community. Requests and advice from the DataverseNO Advisory Committee are communicated to the Board of DataverseNO and to the managers of the institutional collections by the Repository Management.

The DataverseNO repository is managed by staff from the University Library and the IT department at UiT. They are responsible for the management, maintenance, development and the daily operation of the repository, and they take care of the DataverseNO policies and guidelines, communication with the Board of DataverseNO, communication with and training of collection managers, the operation of the DataverseNO Advisory Committee, the configuration of the repository, establishment and configuration of institutional collections, training of collections managers, user management, the implementation of new functionality and procedures to be used in the repository, preservation planning, and the certification of the repository.

The managers of institutional collections within DataverseNO are responsible for the management and operation of the collection, including compliance of the institutional collection and underlying collections with the DataverseNO policies and guidelines, user management of collection curators, training of and communication with collection curators, establishment and configuration of underlying collections, communication with the DataverseNO repository management, communication with the management and the user groups at the partner institution as well as representing the collection in the DataverseNO Advisory Committee.

The managers of special collections have many of the same responsibilities as institutional collection managers, but limited to the scope of the collection. They communicate with the advisory committee for the collection – if applicable.

The bottom section of Figure 6 illustrates the Designated Community of DataverseNO. Since the repository provides free and open access to its collections, the Designated Community of DataverseNO consists of both data contributors and data users. Data users include primarily researchers and research organizations, but also any other stakeholders in society reliant on access to knowledge, e.g. journalists, teachers, industry as well as the greater public. The interaction between data users and the repository happens primarily through direct contact with the contact person(s) for each dataset, and through the general contact information provided for each collection. In a more narrow sense, the Designated Community of DataverseNO can be described as the different user groups that in addition to being data users also are data contributors to the repository. As outlined in the DataverseNO Organization Chart, these user groups fall into three main categories:

- researchers from Norwegian research institutions that are partners of DataverseNO
- researchers working within the scope of any special collection within the DataverseNO repository
- researchers from Norwegian research institutions that are not partners of DataverseNO

Each collection is organized and managed in a way that ensures that the needs of the user group are met to the largest possible extent.

To round off this section, Table 1 gives an overview of the numbers of published datasets in Norwegian research data repositories. The numbers were retrieved from DataCite Search (n.d.) on April 6, 2020, for all repositories using Digital Object Identifiers (DOIs) provided through Unit – Directorate for ICT and joint services in higher education and research, which is the Norwegian national DOI provider.[1]

Table 1: Published datasets with DOI in Norwegian repositories as of 6 April 2020. Sources: DataCite Search and DataverseNO.[2]

Repository	2014	2015	2016	2017	2018	2019	2020	SUM
BI						1		1
DataverseNO	21	11	31	37	201	371	28	700
NIBIO				1	1			2
NILU				1	3			4
NMDC				8	29	10	7	54
NPOLAR			95	40	10	46	2	193
NSD			200	84	1 387	249	25	1 945
UNINETT	17	32	13	22	28	34	6	152

The numbers presented in Table 1 are illustrated as stacked columns in Figure 7.

[1] The table includes only DOI registrations of the resource type “dataset”. DataverseNO also assigns DOIs at file level. Unfortunately, DataCite does currently not distinguish between dataset DOIs and file DOIs. Therefore, the numbers for DataverseNO were obtained directly from the repository. The numbers for DataverseNO also include datasets that were published in TROLLing and UiT Open Research Data before these repositories were included as collections within DataverseNO in 2017.

[2] Abbreviations: BI = Norwegian Business School; NIBIO = Norwegian Institute of Bioeconomy Research; NILU = NILU – Norwegian Institute for Air Research; NMDC = Norwegian Marine Data Centre; NPOLAR = Norwegian Polar Institute; NSD = NSD - Norwegian Centre for Research Data; UNINETT = UNINETT Sigma

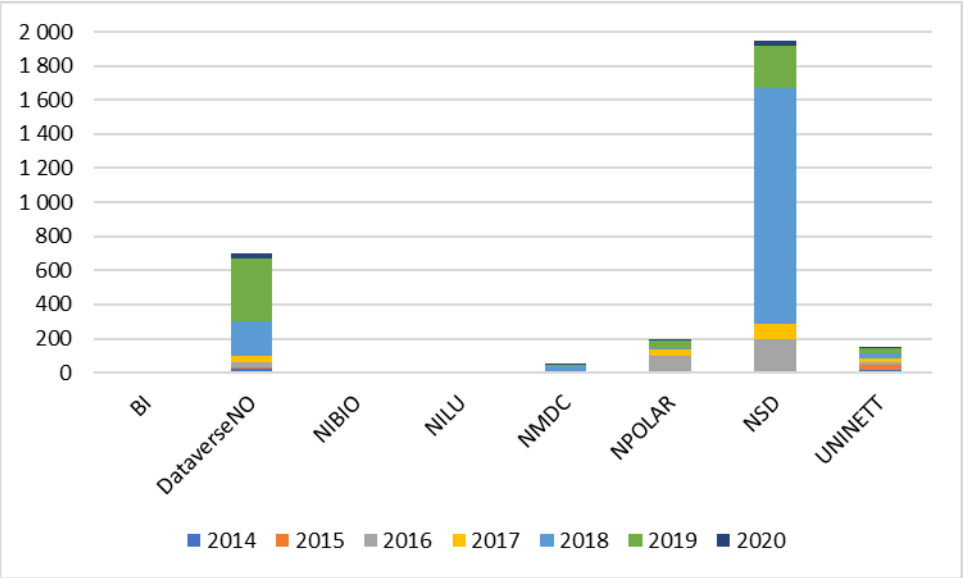

Figure 7: Published datasets with DOI in Norwegian repositories as of 6 April 2020. Sources: DataCite Search and DataverseNO.

Measured in published datasets with DOIs, DataverseNO is thus the second largest research data repository operated in Norway. This position is quite remarkable considering the brief history of the repository.

5. How FAIR are data published in DataverseNO?

Having presented the key organizational features of the DataverseNO repository, and with the overview of the FAIR Data Principles presented in section 1 in mind, let us now have a closer look at how FAIR the data published in DataverseNO are. To answer this question, I will build on a recent systematic overview of how the FAIR Data Principles are implemented in the repository application Dataverse (Crosas, 2020), supplied with information about how these implementations are applied in DataverseNO, and how DataverseNO in addition supports the FAIR Data Principles through deposit requirements and data curation.

To be Findable

Jacobsen et al. (2019, p. 14) summarize the Findable principle in FAIR as follows:

> Digital resources should be easy to find for both humans and computers. Extensive machine-actionable metadata are essential for automatic discovery of relevant datasets and services, and are therefore an essential component of the FAIRification process.

The Findable principle has four elements. According to F1, Findable (meta)data are assigned a globally unique and persistent identifier (PID). Dataverse has implemented F1 by supporting two PID systems, DOI and Handle. All datasets get assigned a PID, whereas PID assignment at file level is an optional feature in Dataverse. DataverseNO

has been using DOI since the repository was established in 2017. The repository has also adopted DOIs at file-level for all files deposited after we in 2018 upgraded the software to version 4.9 where this feature was introduced in Dataverse.

According to F2, Findable data are described with rich metadata. Dataverse supports F2 by providing a discovery metadata schema based on the following widely-used discovery metadata standards in human- and machine-readable formats: Dublin Core (n.d.), Documentation Data Initiative (DDI) (n.d.), DataCite (n.d.), and Schema.org (n.d.). Among the rich options for adding discovery metadata to datasets, DataverseNO in particular emphasizes three mandatory fields: Title, Description, and Keywords. Optionally, customized metadata schemas can easily be configured in Dataverse in addition to the built-in schemas. This option has not been utilized in DataverseNO as we whenever possible prefer to follow more standardized approaches.

According to F3, metadata about Findable data clearly and explicitly include the identifier of the data it describes. F3 is implemented in Dataverse in three ways: the dataset PID is part of the metadata record presented on the dataset landing page; the file PID is part of the metadata record presented on the file landing page; and finally, both dataset and file PIDs are included in exported metadata files. All three implementations are adopted in DataverseNO.

According to F4, Findable (meta)data are registered or indexed in a searchable resource. Datasets published with DOIs in a Dataverse-based repository are harvested and indexed by DataCite Search (n.d.). Through DataCite these metadata are made available to a number of other discovery services, including BASE (Bielefeld Academic Search Engine) (n.d.) and the discovery system used by the libraries at Norwegian universities and university colleges. Schema.org metadata are encoded in Dataverse dataset landing pages and from there indexed by Google Dataset Search (n.d.). In addition to these services, metadata from DataverseNO are also harvested by B2FIND (n.d.), and TROLLing metadata are registered and indexed by the CLARIN Virtual Language Observatory (n.d.).

The current implementation of the Findable principle in the Dataverse application and its adoption in DataverseNO are summarized in Table 2. Green shading indicates (more or less) full implementation or support, whereas orange shading indicates (more or less) lacking implementation or support.

Table 2: The implementation of Findability in Dataverse and its adoption in DataverseNO. Adapted from Crosas (2020).

Principle	Implementation in Dataverse	Applied in DataverseNO
F1	Support for DOI and Handle	Yes (DOI)
	Always at the dataset level	Yes
	Optionally at file level	Yes
F2	Metadata standards in human- and machine-readable formats: Dublin Core; Documentation Data Initiative (DDI); DataCite; Schema.org	Yes
	Optional custom metadata	No
F3	Dataset PID is part of metadata record presented on Dataset landing page.	Yes
	File PID is part of metadata record presented on File landing page.	Yes
	PIDs are included in exported metadata files.	Yes
F4	DataCite metadata is harvested and indexed by DataCite Search.	Yes. In addition: B2FIND and VLO.
	Schema.org metadata is indexed by Google Dataset Search.	Yes

To be Accessible

The gist of Accessibility is according to Jacobsen et al. (2019, p. 14) the following:

> Protocols for retrieving digital resources should be made explicit, for both humans and machines, including well-defined mechanisms to obtain authorization for access to protected data.

The A part in FAIR consists of two elements, A1 and A2. Being Accessible implies according to A1 that (meta)data are retrievable by their identifier using a standardized communications protocol, i.e. a system of rules that allow information to be transmitted between communication systems. The properties of such protocols are further specified in two sub-principles. First, the protocol is open, free and universally implementable (sub-principle A1.1). Data and metadata stored in Dataverse may be accessed through a number of protocols that are in line with A1.1, including Hypertext Transfer Protocol (HTTP) (2020), rsync (2020) over Secure Shell (SSH) (2020), and Representational state transfer (REST) (2020) via Application programming interface (API) (2020), which provides access through e.g. cURL (2020). All these methods are available for users in DataverseNO. HTTP is the default protocol used when users access (meta)data in DataverseNO, whereas access through the other protocols needs to be clarified with the repository management in advance. According to sub-principle A1.2, Accessible data use a protocol that allows for an authentication and authorization procedure, where necessary. For data access via API, Dataverse supports both session- and API key-based authentication. Both methods are

available in DataverseNO. File access in Dataverse can be restricted, either permanently or during an embargo period. In either case, Dataverse allows depositors to handle authorization by deciding whether potential users should be allowed to request access to restricted files, as well as defining possible terms for access. Also these features are available in DataverseNO, however only for embargoed files. Permanent access restriction is currently not accepted in DataverseNO as the repository by default only accepts data that are intended to be made openly available.

Another important aspect of Accessibility is described in A2, namely metadata being accessible, even when the data for some reason no longer are available. By default, datasets – including the files they contain – cannot be deleted in Dataverse. As previously described, any change applied to a dataset results in a new version of the dataset, leaving all previously published versions still being findable and accessible. There may however occur situations where data have been published that for compelling reasons (e.g. legal issues) should not be openly available. In such cases, the files of a dataset may be deaccessioned, meaning that access to these files is removed. Deaccessioning does not affect the citation metadata of the dataset; the data are thus still findable and citable. After deaccessioning, the metadata include information about why the data are no longer available. The deaccession procedure is available in DataverseNO and has been applied a handful of times.

The current implementation of the Accessible principle in the Dataverse application and its adoption in DataverseNO are summarized in Table 3. As in the previous summary, green shading indicates (more or less) full implementation or support.

Table 3: The implementation of Accessibility in Dataverse and its adoption in DataverseNO. Adapted from Crosas (2020).

Principle	Implementation in Dataverse	Applied in DataverseNO
A1	Yes	Yes
A1.1	Support for HTTP (W3C), Rsync over ssh (GNU General Public license) RESTful API (e.g., access through cURL)	Yes
A1.2	Authentication API Tokens	Yes
	Authorization service	Yes, but only for embargo. By default, DataverseNO only accepts open data.
A2	A deaccessioned dataset (data not available) is still findable and citable.	Yes
	Metadata includes information about why the data are not available.	Yes

To be interoperable

Interoperability as used in FAIR implies according to Jacobsen et al. (2019, p. 14) that digital resources can be used in the following way:

> When two or more digital resources are related to the same topic or entity, it should be possible for machines to merge the information into a richer, unified view of that entity. Similarly, when a digital entity is capable of being processed by an online service, a machine should be capable of automatically detecting this compliance and facilitating the interaction between the data and that tool.

Interoperable data have three main characteristics, as specified in I1 to I3. First (I1), Interoperable (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. On a general level, this is implemented in Dataverse with Linked Data (2020) support through JSON-LD (2020) for Schema.org, meaning that the general metadata of a dataset and its files are represented in a format that allows the information to be searched for and processed together with other Linked Data supported data. Although available in DataverseNO, this feature would need more systematic adaptation to make published data Interoperable within a Linked Data approach on a larger scale. Currently, Interoperability may work well for general attributes that are provided through metadata automatically encoded in files (e.g. information about file type), but not for more content and domain-specific attributes. Supporting the latter level of Interoperability is a challenge for most general data repositories, given the heterogeneity of long-tail data they host. However, within certain domains, Dataverse offers more elaborate domain-specific Interoperability support. For example, this is true for quantitative social sciences data that may benefit from a DDI (XML) based schema supporting extensive variable metadata. These

possibilities have been applied to some datasets published in DataverseNO.

Another aspect of Interoperability is the use of (meta)data vocabularies that follow the FAIR principles, as stated in I2. FAIR controlled vocabularies and data models may be deployed manually in Dataverse, e.g. as keywords in the general metadata section, usually requiring some guidance from professional data curators. In DataverseNO, there are a few datasets where this possibility has been explored; see e.g. Gammeltoft (2019). FAIR controlled vocabularies may also be implemented in Dataverse through customized metadata schemas, or be specified as prefilled or suggested values in metadata templates (which in Dataverse are called dataset templates). DataverseNO has been making use of the latter approach, particularly in cases where a single project produces multiple datasets with related content, so that vocabulary values to a large extent are common for all datasets (e.g. datasets covering data from time series). By default, however, Dataverse does not support controlled vocabularies and complex ontologies yet. There is ongoing work to implement controlled vocabulary support for several domains in Dataverse, e.g. through the Social Sciences & Humanities Open Cloud (SSHOC) (n.d.). By default, however, Dataverse does not support controlled vocabularies and complex ontologies yet.

The last property of Interoperability – specified in sub-principle I3 – is that Interoperable (meta)data include qualified references to other (meta)data. Such references may be added in two fields of the general metadata schema in Dataverse, one for related data, and another for related materials (other research objects). Currently, information may only be entered as free text into these two fields, and the information is not exported to DataCite. This shortcoming will be remedied in a future version of the Dataverse application.

The current implementation of the Interoperable principle in the Dataverse application and its adoption in DataverseNO are summarized in Table 4. As in the previous summaries, green shading indicates (more or less) full implementation or support, whereas orange shading indicates (more or less) lacking implementation or support. In addition, yellow shading indicates partial implementation or support.

Table 4: The implementation of Interoperability in Dataverse and its adoption in DataverseNO. Adapted from Crosas (2020).

Principle	Implementation in Dataverse	Applied in DataverseNO
I1	Linked data support with JSON-LD for Schema.org	Partially, for general attributes such as file type
	DDI (XML) as a rich schema to support extensive variable metadata	Partially/in some datasets
I2	FAIR controlled vocabularies and data models may be deployed manually, e.g. in well-curated datasets	Partially/in some datasets
	Custom metadata and metadata template can help.	Partially/in some datasets (metadata template)
	Controlled vocabularies and ontologies not supported by default. But, cf. ongoing work on support for some domains (e.g. SSHOC).	No
I3	DDI schema supports references to other data.	Yes, where applicable
	Not yet supported: structured metadata about related objects included in exported DataCite metadata (coming soon)	No

To be Reusable

Digital resources supporting the last part of FAIR, Reusability, can be described as follows according to Jacobsen et al. (2019, p. 14):

> Digital resources are sufficiently well described for both humans and computers, such that a machine is capable of deciding: if a digital resource should be reused (i.e., is it relevant to the task at-hand?); if a digital resource can be reused, and under what conditions (i.e., do I fulfill the conditions of reuse?); and who to credit if it is reused.

Reusable (meta)data are defined in the FAIR Data Principles as (meta)data that are richly described with a plurality of accurate and relevant attributes. The principle is further elaborated in three sub-principles. Sub-principle R1.1 describes the first characteristic of Reusability as (meta)data being released with a clear and accessible data usage license. Information about data use license or waiver as well as – where applicable – information about data access and terms of use are by default included in the metadata of datasets published in Dataverse. Licenses other than the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication (n.d.) are not predefined, and they are by default not machine-readable. This latter shortcoming has so far not had unbearable consequences for DataverseNO as the repository uses CC0 as its default license, which has been applied to almost all published datasets. Currently, the Dataverse application has no support for explicit and machine-readable licenses for metadata. Therefore, DataverseNO has specified the terms for reuse of metadata on the DataverseNO Metadata Harvesting webpage (n.d.), though only in a human-readable format.

The second characteristic of reusable (meta)data is according to sub-principle R1.2 that they are associated with detailed provenance. Provenance is information about the origin of the data, e.g. how they were obtained and processed, as well as who has been involved in the management of the data. Dataverse has support for rich metadata including information about data authors and other contributors, data providers, data distributors, as well as related data (e.g. used as input data). Any changes made to published datasets are automatically documented through version control in Dataverse. In addition, Dataverse supports the registration of provenance information in a more formal way based on the W3C PROV data model for the interchange of provenance information on the Web (W3C, n.d.). This feature has so far not been made use of in DataverseNO.

The last characteristic of Reusability included in FAIR is described in sub-principle R1.3, which says that reusable (meta)data meet domain-relevant community standards. There is a multitude of (meta)data standards within the different domains of science. Dataverse currently supports a few of them. In addition to the discovery metadata schema, there is quite a substantial schema for social sciences, compliant with parts of the Documentation Data Initiative (DDI). Less rich metadata schemas are available for data from astronomy and astrophysics, and for data from life sciences (Dataverse Metadata References, n.d.). In DataverseNO, these metadata schemas are used in some of the datasets from relevant domains.

There is ongoing work in the Dataverse community to establish metadata schemas including controlled vocabularies for more domains.

In addition to the metadata schemas provided in Dataverse by default, domain-relevant community standards may be implemented by creating custom metadata blocks. As mentioned previously, this option has not been utilized in DataverseNO as we whenever possible prefer to follow standardized approaches.

On file-level, Dataverse automatically extracts metadata from FITS files used in astronomy (FITS (Flexible Image Transport System), n.d.). Currently, DataverseNO does not contain any FITS files.

Also on file-level, Dataverse automatically converts certain tabular file formats (e.g. R frames, Stata, and Excel files) into tab-separated plain text format, which is a more preferred file format for long-term preservation. For these tabular files, Dataverse also extracts upon ingest the variable metadata for each column in the table. This feature has been applied to some of the files in DataverseNO.

In addition to the Reusability implementations that are available by default in the Dataverse repository application, the DataverseNO deposit guidelines include two measures that are essential for increasing the reusability of data. First, data must be submitted in

preferred file formats that are suited for long-term preservation. The guidelines contain a list of preferred file formats for some common file types; for other file types, the file format is assessed during curation. If necessary, file formats are discussed in the DataverseNO curation network and the DataverseNO Advisory Committee before the repository management decide whether the format should be included on the list of preferred file formats.

Second, datasets to be published in DataverseNO must include a ReadMe file containing human-readable information about how to reuse the data, including what in some fields is called a data dictionary (Tierney & Ram, 2020, p. 7). Although ReadMe files are not machine-readable and thus do not support one of the basic intentions of the FAIR Data Principles, they contribute substantially to making research data more reusable by humans. In particular, this is true for data from the long tail of research, which includes fields where domain-specific data and metadata standards still are either non-existent or not established. For this type of long-tail research data, I argue that Reusability for humans has higher priority than Reusability for machines in the implementation phase of the FAIR Data Principles. Paraphrasing Tierney & Ram (2020, p. 1), I also argue that making data reusable falls on a continuum, and entering it should come with feasible barriers.

The current implementation of the Reusability principle in the Dataverse application and its adoption in DataverseNO are summarized in Table 5. As in the previous summaries, green shading indicates (more or less) full implementation or support, yellow shading indicates partial implementation or support, whereas orange shading indicates (more or less) lacking implementation or support.

Table 5: The implementation of Reusability in Dataverse and its adoption in DataverseNO. Adapted from Crosas (2020).

Principle	Implementation in Dataverse	Applied in DataverseNO
R1		
R1.1	Included in metadata: data use license/waiver; data access and use terms. But, licenses other than CC0 are not predefined and by default not machine-readable.	Yes. Almost all datasets are published under default license CC0.
	By default no support for explicit information about metadata license	Terms for reuse of metadata described on website
R1.2	Rich citation metadata including information about data authors and other contributors, providers, distributors, related data (input data)	Yes
	Versions with changes documented automatically	Yes
	W3C PROV support	No
R1.3	DDI for social science data	Partially/in some datasets
	Metadata blocks for other community standards	Partially/in some datasets
	Ongoing work on support for more domains.	No
	Custom metadata	No
	FITS for astronomy data	N/A (so far)
	File format conversion to reusable formats (tabular)	Partially/in some datasets
		Data in preferred file formats
		Datasets include ReadMe file.

It must be noted that the FAIR Data Principles were not designed as fully operationalized criteria or a checklist, and there currently does not exist a commonly adopted approach or method for how to assess the FAIRness of digital resources or infrastructures. Nevertheless, I argue that the FAIRness evaluation presented in this section may prove useful for the continuous FAIRification efforts in the Dataverse community and beyond.

Table 6 summarizes the results from the FAIRness evaluation of Dataverse and DataverseNO presented in this section. Note that the size of the different elements in the table are not claimed to represent their exact contribution to the overall FAIRness, but they may give a rough indication of the FAIRness of Dataverse and DataverseNO as discussed in this section.

Table 6: Summary of current FAIR implementation or support in Dataverse and DataverseNO

	F				A			I			R		
	1	2	3	4	1.1	1.2	2	1	2	3	1.1	1.2	1.3
Dataverse													
DataverseNO													

= (more or less) full implementation or support
= partial implementation or support
= (more or less) lacking implementation or support
= not applicable

Bearing in mind the mentioned limitations of the FAIRness assessment carried out above, the main conclusions from this section may be summarized as follows: Dataverse and DataverseNO provide strong support for Findability and Accessibility, somewhat weaker support for Reusability and rather weak support for Interoperability. Dataverse and DataverseNO are continuously working to increase their FAIRness support. For instance, DataverseNO has recently been selected as one of 12 repositories to get support from FAIRsFAIR (n.d.) to improve its level of Interoperability.

The FAIRness support provided by applications like Dataverse and by repositories like DataverseNO is fundamentally important for the realization of FAIR. In addition, we need mechanisms which ensure that such support can be provided in a stable way also in the long term. This brings us to the last topic to be discussed in this article, the sustainability and trustworthiness of DataverseNO.

6. How sustainable and trustworthy is DataverseNO?

The realization of FAIR depends to a large extent on an ecosystem of federated infrastructures. In such a system, data are managed in different infrastructures which however all follow common standards to make data discoverable and reusable across infrastructures and scientific domains. Although not explicitly addressed by the FAIR Data Principles, the sustainability of infrastructures enabling FAIR data is recognized as a core issue. In order to be sustainable, infrastructures and services must have appropriate funding. In their report on how to turn FAIR into reality, the European Commission expert group on FAIR data make two priority recommendations on funding. The first one is about providing strategic and coordinated funding (European Commission, 2018, p. 55):

> Funders should adopt a coordinated approach to supporting core infrastructure and services, building on existing investments where appropriate. Funding should be tied to certification schemes, sustainable business models and other community-vetted indicators that demonstrate viability.

The second one is about providing sustainable funding (European Commission, 2018, p. 57):

> Funders who issue requirements on FAIR must provide support to ensure the components of the FAIR ecosystem are maintained at a professional service level with sustainable funding. Service providers should explore multiple business models and diverse income streams.

How does DataverseNO relate to these recommendations? Let us briefly discuss the different elements in turn.

1. Funders should adopt a coordinated approach to supporting core infrastructure and services, building on existing investments where appropriate.

As mentioned earlier, repositories are considered core elements in the realization of FAIR. Given the volume of published datasets, it seems appropriate to say that DataverseNO is a core infrastructure for research data in Norway. Building on existing resources is a key principle of the European Open Science Cloud (EOSC) and should also be applied to national investment strategies. DataverseNO has taken this key principle to heart by – among other things – making use of existing open-source software (Dataverse) and extensive support from the Dataverse developer and user community (to be further discussed below) as well as by drawing on existing human resources and organizational infrastructure at the owner institution and the partner institutions.

2. Funding should be tied to certification schemes, sustainable business models and other community-vetted indicators that demonstrate viability.

a) Funding should be tied to certification schemes

DataverseNO realized shortly after its establishment that certification through a recognized organization is an important and useful way to provide evidence of good quality for depositors, partner institutions and potential users of data published in the repository. After we had worked with our application over a period of two years, DataverseNO earned the CoreTrustSeal in March 2020. The CoreTrustSeal (n.d.) is a quality seal of approval for sustainable and trustworthy research data repositories. Through the CoreTrustSeal certification, data repositories and other infrastructures and services demonstrate that they meet a number of requirements for both technical infrastructure and stewardship model and routines. In total, 15 qualification requirements are included in the CoreTrustSeal certification, and DataverseNO meets 14 of them at the highest level ("The guideline has been fully implemented in the repository"). As part of the certification process, we have developed a comprehensive set of guidelines for the repository, the DataverseNO Policy Framework (n.d.).

b) Funding should be tied to sustainable business models

One of the aspects assessed in CoreTrustSeal certification is the business model which a repository operates on. The business model of DataverseNO builds on common approaches to cooperation between higher education institutions in Norway. The multi-institutional model of DataverseNO has grown out of an institutional service at UiT, realizing the advantage of the network effect as described in Arlitsch & Grant (2018). Instead of establishing similar or even (near-)identical repositories at each and every institution, the repository service first built at one institution was rearranged to meet the demands from a larger community. The repository has been established entirely within existing budget limits at the owner institution and the partner institutions, involving no project-based funding. Instead, funding allocations have been (re-)prioritized, and resources have been gradually scaled in line with increasing demand. Deploying an open-source and free application (Dataverse) to run the main technical infrastructure of the repository renders the costs for technical operation and maintenance at a feasible level. Together with the fee for DOI minting, these costs are shared between the partner institutions. The largest contribution from each partner institution is their investment in institutional curation support for datasets to be published in their institutional collections, as well as their investment in research data management training for support staff and researchers. However, this institutional burden is considerably reduced by the collaborative approach to curational training and knowledge exchange between DataverseNO partner institutions. Furthermore, support from the international Dataverse community is an additional contribution to leverage more powerful research data management support services at the DataverseNO partner institutions.

The Organisation for Economic Co-operation and Development (OECD) has explored the income streams, costs, value propositions, and business models for 48 research data repositories. This survey is summarized in a report from 2017, including a set of recommendations on how to develop sustainable business models for repositories (OECD, 2017a). The authors of the report identify the following broad categories of business models being employed by research data repositories (OECD, 2017a, pp. 38–39):

- Substantially structurally funded (i.e. central funding or contract from a research or infrastructure funder that is in the form of a longer-term, multi-year contract)
- Substantially supported by the host institution (i.e. direct or indirect support from the host institution)
- Substantially depending on data deposit fees (i.e. in the form of annual contracts with depositing institutions or per-deposit fees)

- Substantially funded from access charges (i.e. charging for access to standard data or to value-added services and facilities)
- Substantially supported through contract services or project funding (i.e. charges for contract services to other parties or for research contracts)
- Business models based on a combination of revenue sources

The business model of DataverseNO may be said to be a combination and adaptation of three of these models: 1) host or institutional support; 2) deposit-side contract; and 3) possible future project funding. The OECD report gives an analysis of possible advantages and disadvantages associated with these models. Let us briefly evaluate the DataverseNO business model based on these findings.

Host or institutional support

The main funding source of DataverseNO is institutional support in the form of investment in technical infrastructure at the owner institution, and investment in human resources to provide support at both the owner institution and all partner institutions.

Among the major strengths of host or institutional support models, the OECD report points out the following four (OECD, 2017a, p. 43). The first strength is longer-term sustainability, because universities and other research organizations tend to be long-lived and have robust funding. Both the owner institution and the partner institutions of DataverseNO are state-owned universities and thus part of the national, governmental higher education and research system and under the ultimate responsibility of the Norwegian Ministry of Education and Research. They are all reputable institutions that have existed for many decades – though in some cases not under their current name. Thus, they all are organized and funded in a way that ensures the operation of sustainable services for higher education and research in an enduring perspective.

The second strength of host or institutional support models is convergence of interest between host and repositories. This is typically the case where the data repository aligns with the research strategy of the hosting institution (OECD, 2017a, p. 43). All institutions involved in DataverseNO have recognized Open Science as an important issue in their missions. As a general rule, these institutions retain ownership of the research data produced by their employees, and they thus have a genuine interest in preserving the long-term value of these assets. Therefore, their investment in curational support and participating as partner institutions in DataverseNO is highly convergent with the long-term commitment of the repository.

A third strength of host or institutional models that the report mentions is cost optimisation that may be obtained by “sharing services within the institution” (OECD, 2017a, p. 43). In the case of DataverseNO, cost optimisation is above all a result of the repository

being shared between institutions. This applies above all to costs associated with the development and maintenance of the technical infrastructure, and to a somewhat lesser extent to costs involved in the development of human skills and expertise.

As a fourth major strength of host or institutional models the report considers that "repository support staff can be close to the researchers for on-hand support" (OECD, 2017a, p. 43). The operation – including curation – of institutional collections in DataverseNO is part of the onsite research support services and the institutional management at the DataverseNO partner institutions. This enables repository support staff to provide customized local support for data depositors.

As the major possible weakness of host or institutional models the report reckons that "the focus on the local institutional community […] may lead to fragmentation of domain data, lower levels of curation, and lower interoperability" (OECD, 2017a, p. 43). In my view, this possible weakness is not so much related to the choice of business model, but has rather to do with the question of what type of repository is best suited to meet the different needs of different research communities. DataverseNO is a generic repository and thus contains mostly long-tail data that do not easily fit into trusted domain-specific repositories. However, despite its generic mission, DataverseNO strives to provide domain-specific expertise as far as possible. At the larger partner institutions, data deposited into institutional collections are curated by research data support staff who are subject specialists in addition to being trained in research data management. Special collections of DataverseNO are without exception managed and curated by specialists within the subject in question. As mentioned earlier, the Interoperability support of the repository application is being continuously improved for an increasing number of scientific domains. Overall, DataverseNO has thus taken substantial actions to mitigate the potential risk of fragmentation of domain data, lower levels of curation, and lower interoperability.

Deposit-side contract

In addition to the institutional support as described above the business model of DataverseNO includes a deposit-side contract which is part of the partner agreement. The annual fees from partner institutions cover the shared costs for technical maintenance as well as for training of collections managers, and other support given by the owner institution.

The main strength of business models relying on deposit-side contracts is the potential such repositories might have to achieve economies of scale and cost optimisation (OECD, 2017a, p. 44). Although the institutions currently involved in DataverseNO are not-for-profit organizations, they may profit from effects of economies of scale, and they are of course interested in reducing their costs. As

compared to the establishment of multiple repositories at each of the currently nine partner institutions, the multi-institutional model of DataverseNO obviously contributes to cost optimisation. It is however somewhat uncertain to what extent DataverseNO may be said to have achieved economies of scale.

The OECD report identifies several possible weaknesses of deposit-side contracts (OECD, 2017a, p. 44). Contract-based funding can be unpredictable as commitments are time-limited, e.g. from year to year. DataverseNO partner agreements are terminable. The DataverseNO partner agreement takes into account the consequences of a possible withdrawal of partnership. Datasets from leaving partner institutions may be transferred to other trustworthy repositories on request from the partner institution. In case datasets are to remain in DataverseNO, the leaving institution will have to provide the necessary financial means to ensure the long-term preservation of these datasets. Also, according to the DOI agreement with DataCite, DataverseNO commits to providing access to published datasets for at least ten years after DOI assignment. Given these precautions, the withdrawal of (a substantial part of) partner institutions would of course be unfortunate as it would reduce access to shared resources such as knowledge exchange, as well as reduced opportunities for cost optimisation. It would however not threaten the sustainability of the stewardship of research data still being published in the repository, as data curation still would be the responsibility of the remaining partner institutions. In fact, the basic costs associated with repository maintenance were by and large the same also when UiT operated their own single-institution repository, which was the predecessor of DataverseNO.

Another possible weakness of business models relying on deposit-side contracts is that they may involve relatively high transaction costs in managing contracts (OECD, 2017a, p. 44). DataverseNO keeps administrative costs as low as possible. In the usual case, a partner agreement is signed once, when the partnership is established. After that, partner institutions are invoiced for annual fees. The annual fee is calculated based on a straightforward allocation key. This procedure has proven to be very functional in similar models for collaboration within Norwegian higher education institutions.

The OECD report mentions a third possible weakness of business models based on deposit-side contracts: “In contrast to discipline, subject, or institutional repositories, there may be limited engagement with researchers, as users or depositors” (OECD, 2017a, p. 44). As DataverseNO is a multi-institutional repository, there is no real contrast to institutional repositories with regard to researcher engagement. Research support services at partner institutions are highly involved in the operation of the collections within DataverseNO, particularly through their work with data curation. Partner institutions have well-established venues in place where research data support staff involved in the operation of DataverseNO,

e.g. university library staff discuss data management-related issues with researchers from the different research communities at the institution.

Again, we may note that DataverseNO has taken precautions to mitigate potential risks of the deposit-side contract element of its business model.

Possible future project funding

DataverseNO has been developed and established without any project funding whatsoever. However, we do not rule out the possibility to apply for project-based funding to further develop the repository. A prerequisite for such projects will be that they do not have a negative impact on the long-term sustainability of the repository.

Project funding used in this way will take advantage of the main strength of business models based on this type of funding, namely "support for innovation and development, and opportunities for further developments in the future" (OECD, 2017a, p. 47).

The primary disadvantages of project-based funding identified in the OECD report include "the short-term nature of such funding, lack of flexibility in its spending, and possible diversion of staff and effort from the core tasks and functions of repository operation" (OECD, 2017a, p. 47).

The first two shortcomings are not relevant in the case of DataverseNO, as project-based funding will not be used for operational purposes. The risk of diversion of staff and effort may to a large extent be prevented by hiring temporary staff, a procedure that is commonly practised at Norwegian higher education institutions.

Finally, it should be mentioned that all three funding sources combined in the business model of DataverseNO are compatible with open data principles, as they do not imply any data access charges for users of datasets published in the repository.

Table 7 summarizes the main strengths and weaknesses of the three funding sources combined in the business model of DataverseNO. Green shading indicates strengths, whereas orange shading indicates weaknesses. The rightmost column shows whether the strengths and weaknesses are applicable to DataverseNO. Here, identical shading indicates full applicability, opposite shading indicates lacking applicability, whereas yellow shading indicates partial applicability.

Table 7: Main strengths and weaknesses of the DataverseNO business model

Funding source			Applicable to DataverseNO
Institutional funding	Pros	Compatible with open data principles	Yes
		Longer-term stability	Yes
		Convergence of interest between host and repository	Yes
		Cost optimisation	Yes
		Close to researchers	Yes
	Cons	Fragmentation of domain data	To some extent; but considerable support for domain data
		Lower levels of curation	No
		Lower interoperability	To some extent; but continuous improvement
Deposit-side contract	Pros	Compatible with open data principles	Yes
		Economies of scale	Not relevant
		Cost optimisation	Yes
	Cons	Unpredictable funding	To some extent; but no threat to sustainability
		High administrative costs	No
		Limited engagement with users	No
Project funding	Pros	Support for innovation and development	Yes
		Compatible with open data principles	Yes
	Cons	Short-term nature	No
		Lack of flexibility	No
		Diversion of staff and effort	To some extent; but may be prevented by hiring of temporary staff

This overview shows that the organizational model of DataverseNO has managed to take advantage of all major strengths, and to mitigate or even eliminate the major weaknesses of the funding sources combined in its business model. Notably, by embedding data curation and other support services in the partner institutions, the model combines the scaling advantages of multi-institutional collaboration with the closeness of onsite support for researchers.

Let us now turn to the last part of the strategic and coordinated funding recommendation from the European Commission expert group on FAIR data:

c) Funding should be tied to community-vetted indicators that demonstrate viability

In addition to business models, there is another aspect that should be considered when evaluating the sustainability of research infrastructures, and that is their relation to international networks. The role of such networks is emphasized in another report commissioned by the OECD Global Science Forum (GSF):

> [A]ccess to research data […] is dependent on individual data repositories at the institutional, national and disciplinary levels and on co-ordinated international networks of these repositories (OECD, 2017b, p. 3).

One of the main characteristics of successful networks is that they "define their scope well and avoid the "not-invented-here" or "try-to-solve-it-all" mentality" (OECD, 2017b, p. 23). One such network that has managed to steer clear of both these pitfalls is the Dataverse developer and user community. The most obvious contribution of this network to the success of DataverseNO is of course the Dataverse repository application, which is the technical backbone of DataverseNO. When we started to develop TROLLing back in 2013, Dataverse was virtually the only repository application designed for research data that could be used out-of-the-box. As is the case with any product, the software had (and still has) its shortcomings, but it has always provided solid core functionality for archiving, publishing and citing research data. The last few years, Dataverse has been further developed to meet more and more of the requirements of a FAIR-aligned repository software. As of 11 April 2020, there are at least 55 repositories worldwide that are run on Dataverse (n.d.).

The popularity of Dataverse is reflected in a continuously growing developer and user community. This community consists of hundreds of members, including software developers, researchers, librarians and data scientists. Over 100 people have been and are still contributing to the software code, the main distribution of which is led by Harvard University. Other members of the community contribute with user interface (UI) and user experience (UX) testing and training; presenting and discussing issues in online discussion fora, community calls and the yearly Dataverse Community Meeting; as well as organizing workshops and training sessions (Durand, 2020). Several members of the community are involved in national and international infrastructure projects that aim at developing further different features of the application and/or tools that can be integrated with Dataverse. In 2018, the Global Dataverse Community Consortium (n.d.) was established to provide international organization to existing community efforts and to provide a collaborative venue for institutions to leverage economies of scale in support of Dataverse repositories around the world.

Being embedded in the global Dataverse community has contributed significantly to improve both the technical infrastructure and the operation of DataverseNO, and it has also advanced the further alignment of the repository with international recommendations for sustainable research data stewardship.

To summarize this somewhat longish discussion we may conclude that the way in which DataverseNO is funded and operated is well in line with the first recommendation from the European Commission expert group on FAIR data on strategic and coordinated funding.

The second funding-related recommendation from this expert group concerns sustainable funding. The recommendation contains two parts. Here is the first one (European Commission, 2018, p. 57):

3. Funders who issue requirements on FAIR must provide support to ensure the components of the FAIR ecosystem are maintained at a professional service level with sustainable funding.

This recommendation is being implemented at DataverseNO partner institutions. These institutions require or at least expect their researchers to manage their data in line with the FAIR Data Principles. As consequence, they commit to provide sufficient support to handle these data in a responsible way, e.g. with the services they offer through their institutional collections in DataverseNO. So far, these efforts have been enabled mostly by reprioritizing job tasks and redefining responsibilities at these institutions. As the demand for research data support services increases, institutions will need to consider major reallocations of existing funding to ensure the sustainable operation of these services.

Most of the long-tail research projects carried out in Norway are funded by the host institutions of researchers through ordinary funding. In addition to this ordinary institutional funding, the recommendation above should also be implemented by the Research Council of Norway and other research funding bodies. This can be done in different ways. At the researcher side, funders should cover costs associated with research data management and make it mandatory to include these costs in grant applications for research projects. On the infrastructure side, funders should contribute to the development and operation of not only domain-specific research data infrastructures, but also support services for data from the long tail of research.

The second part of the sustainable funding recommendation from the European Commission expert group on FAIR data is repeated here (European Commission, 2018, p. 57):

4. Service providers should explore multiple business models and diverse income streams.

The business model of DataverseNO already combines different funding sources. As it may become more and more common for institution-external research project grants to cover costs for research data management (see suggestion above), this funding will be an additional income stream for DataverseNO, given that the partner institutions allocate adequate shares of this income to the operation of their institutional collections in DataverseNO.

In the possible case of minor research organisations joining DataverseNO, we also may consider charging data deposit fees from these institutions in exchange for data curation services, at least in an initial phase. However, as a general rule, the organisational model

of DataverseNO requires partner institutions to take responsibility for such core support services for their researchers.

7. Summary and outlook

DataverseNO is a national, generic repository for open research data. This article presents the organization and operation of DataverseNO and gives an evaluation of the repository along three sets of recommendations: the FAIR Data Principles, recommendations on how to turn FAIR into reality, as well as recommendations for sustainable business models for data repositories. DataverseNO provides strong support for Findability and Accessibility, somewhat weaker support for Reusability, and rather weak support for Interoperability. The business model of DataverseNO takes advantage of the strengths and mitigates or eliminates the risks of the funding sources that are combined in the model. As attested by the CoreTrustSeal, DataverseNO has proven to be a sustainable and trustworthy research data repository, primarily for data from the long tail of research.

DataverseNO is continuously working to increase its FAIRness and its sustainability. Other future activities include the improvement of domain-specific support, e.g. by implementing metadata schemas for more domains. As the Dataverse application soon will offer support for sensitive data, we also may want to consider extending the scope of the repository to include this type of data. Finally, DataverseNO will of course remain open for new organizations to join in as partners.

Acknowledgements

Thanks to Helene N. Andreassen, Mercè Crosas, and the editors of this issue of *Ravnetrykk* for useful comments on earlier versions of this article.

Disclaimer

The author is one of the repository managers of DataverseNO.

References

Application programming interface. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Application_programming_interface&oldid=958345761

Arlitsch, K., & Grant, C. (2018). Why So Many Repositories? Examining the Limitations and Possibilities of the Institutional Repositories Landscape. *Journal of Library Administration*, *58*(3), 264–281. https://doi.org/10.1080/01930826.2018.1436778

B2FIND. (n.d.). Retrieved 21 May 2020, from http://b2find.eudat.eu/

BASE (Bielefeld Academic Search Engine). (n.d.). Retrieved 21 May 2020, from https://www.base-search.net/

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world* (pp. XXV, 383). The MIT Press.

CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. (n.d.). Retrieved 23 May 2020, from https://creativecommons.org/publicdomain/zero/1.0/

Christian, T.-M., Gooch, A., Vision, T., & Hull, E. (2020). Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLOS ONE*, *15*(3), e0230281. https://doi.org/10.1371/journal.pone.0230281

CLARIN Virtual Language Observatory. (n.d.). Retrieved 21 May 2020, from https://vlo.clarin.eu/

Conzett, P. (2019). *Disciplinary Case Study: The Tromsø Repository of Language and Linguistics (TROLLing)*. https://doi.org/10.5281/zenodo.2668775

Conzett, P. (2020). *Research Data Publishing at UiT The Arctic University of Norway* (Version 1) [Dataset]. DataverseNO. https://doi.org/10.18710/JWTJJB

Conzett, P., & Østvand, L. (2018). Støttetenester for forskingsdatahandtering på UiT Noregs arktiske universitet – erfaringar og forslag til beste praksis. *Nordic Journal of Information Literacy in Higher Education*, *10*(1), 65–80. https://doi.org/10.15845/noril.v10i1.283

CoreTrustSeal. (n.d.). Retrieved 21 May 2020, from https://www.coretrustseal.org/

Crosas, M. (2020). Fair Principles and Beyond: Implementation in Dataverse. *Septentrio Conference Series*, *2*, Article 2. https://doi.org/10.7557/5.5334

Crosas, M., Gautier, J., Karcher, S., Kirilova, D., Otalora, G., & Schwartz, A. (2018). *Data policies of highly-ranked social science journals* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/9h7ay

CURL. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=CURL&oldid=954043706

Data Documentation Initiative (DDI). (n.d.). Retrieved 23 May 2020, from https://ddialliance.org/

DataCite. (n.d.). [Website]. Retrieved 23 May 2020, from https://schema.datacite.org/

DataCite Search. (n.d.). Retrieved 21 May 2020, from https://search.datacite.org/

Dataverse. (n.d.). Retrieved 21 May 2020, from https://dataverse.org/home

Dataverse Metadata References. (n.d.). Dataverse. Retrieved 23 May 2020, from http://guides.dataverse.org/en/latest/user/appendix.html

DataverseNO Curator Guidelines. (n.d.). Info: DataverseNO. Retrieved 21 May 2020, from https://site.uit.no/dataverseno/admin-en/curatorguide/

DataverseNO Deposit Guidelines. (n.d.). Info: DataverseNO. Retrieved 21 May 2020, from https://site.uit.no/dataverseno/deposit/

DataverseNO Metadata Harvesting. (n.d.). Info: DataverseNO. Retrieved 21 May 2020, from https://site.uit.no/dataverseno/about/#metadata-harvesting

DataverseNO Policy Framework. (n.d.). Info: DataverseNO. Retrieved 21 May 2020, from https://site.uit.no/dataverseno/about/policy-framework/

Dubline Core. (n.d.). Retrieved 23 May 2020, from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Durand, G. (2020). Dataverse's Approach to Technical Community Engagement. *Septentrio Conference Series*, *2*, Article 2. https://doi.org/10.7557/5.5424

European Commission. (n.d.). *European Open Science Cloud (EOSC)*. Retrieved 4 April 2020, from https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

European Commission. (2018). *Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data*. Publications Office of the European Union. https://op.europa.eu/s/n1Yo

FAIRsFAIR. (n.d.). Retrieved 21 May 2020, from https://www.fairsfair.eu/

FITS (Flexible Image Transport System). (n.d.). Retrieved 21 May 2020, from https://fits.gsfc.nasa.gov/

Gammeltoft, P. (2019). *The place-name Elverhøy in Norway* (Version 1) [Dataset]. DataverseNO. https://doi.org/10.18710/OG9ARD

GÉANT, & UNINETT. (2019, May). *Why TROLLing is the thing to do for linguists*. In The Field. https://www.inthefieldstories.net/why-trolling-is-the-thing-to-do-for-linguists/

Google Dataset Search. (n.d.). Retrieved 21 May 2020, from https://datasetsearch.research.google.com/

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Hypertext Transfer Protocol. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Hypertext_Transfer_Protocol&oldid=957536773

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2019). FAIR Principles: Interpretations and

Implementation Considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

JSON-LD. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=JSON-LD&oldid=956136847

Linked data. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Linked_data&oldid=951149328

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396. https://doi.org/10.1037/h0054346

Neylon, C. (2017). Compliance Culture or Culture Change? The role of funders in improving data management and sharing practice amongst researchers. *Research Ideas and Outcomes*, *3*, e14673. https://doi.org/10.3897/rio.3.e14673

Nosek, B. (n.d.). *Shifting Incentives from Getting It Published to Getting it Right*. Retrieved 4 April 2020, from https://osf.io/bxjta/

OECD. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD Publishing. https://doi.org/10.1787/9789264034020-en-fr.

OECD. (2017a). Business models for sustainable research data repositories. *OECD Science, Technology and Industry Policy Papers*, *47*. https://doi.org/10.1787/302b12bb-en

OECD. (2017b). Co-ordination and support of international research data networks. *OECD Science, Technology and Industry Policy Papers*, *51*. https://doi.org/10.1787/e92fa89e-en

re3data.org. (2015). TROLLing; editing status 2020-04-07. *Re3data.Org - Registry of Research Data Repositories*. https://doi.org/10.17616/R3834T

re3data.org. (2017). DataverseNO; editing status 2020-04-07. *Re3data.Org - Registry of Research Data Repositories*. https://doi.org/10.17616/R3TV17

Representational state transfer. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Representational_state_transfer&oldid=956443795

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed., pp. XXI, 551). Free Press.

Rsync. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Rsync&oldid=956572441

Schema.org. (n.d.). Retrieved 23 May 2020, from https://schema.org/

Secure Shell. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Secure_Shell&oldid=957079117

The e-IRG Task Force on the Long Tail of Data. (2016). *Long Tail of Data* (Version 1.74, E-IRG Task Force Document). e-IRG.

http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf

The Global Dataverse Community Consortium. (n.d.). Retrieved 21 May 2020, from http://dataversecommunity.global/home

The Social Sciences & Humanities Open Cloud. (n.d.). Retrieved 21 May 2020, from https://www.sshopencloud.eu/

Tierney, N. J., & Ram, K. (2020). A Realistic Guide to Making Data Available Alongside Code to Improve Reproducibility. *ArXiv:2002.11626 [Cs]*. http://arxiv.org/abs/2002.11626

UiT Open Research Data. (n.d.). DataverseNO. Retrieved 21 May 2020, from https://opendata.uit.no/

W3C. (n.d.). *PROV-Overview*. Retrieved 21 May 2020, from https://www.w3.org/TR/prov-overview/

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18