Danmarks
Grundforskningsfond
Danish National
Research Foundation

THE DANISH NATIONAL
RESEARCH FOUNDATION

OPEN ACCESS TO DATA – IT'S NOT THAT SIMPLE

OPEN ACCESS TO DATA – IT'S NOT THAT SIMPLE

More data have been created in the past two years than in all human history.

Five years ago, a cardiovascular geneticist would compare 200 DNA samples from healthy and diseased individuals to address the possible inheritance of a disease. Today, she has 60,000 complete human genomes available, free of charge, to address the same question. A few years ago, classical geographers analyzed the problem of flooding. Today, datamaticians can analyze the problem from 220 billion readings in Denmark alone, enabling them to draw up an exact potential flood map of the country. The space telescope Kepler provides the Stellar Astrophysics Centre (SAC) with hundreds of terabytes of data. Rather than making detailed studies of just a few stars, SAC can now characterize planetary systems around distant stars and probe the properties of thousands of stars. Very few scientific fields are untouched by these new scales of quantitative data.

To elucidate the advantages and challenges of open data in science, the Danish National Research Foundation (DNRF) conducted a survey among its researchers on open access to data. The results are presented here. The short message is that it is a challenge to exploit the possibilities of open data wisely.

"This digital revolution is an exceptional opportunity to excel in research, but also a challenge to exploit wisely"

– Professor Liselotte Højgaard, Chair of the Board of the DNRF

As stated by the *Ministry of Higher Education and Science*: "Open Access to research data can be of great value for researchers, citizens and businesses in the form of new knowledge and discoveries. Increased access to the underlying data can moreover contribute to efficiency of the research by reuse of data. It can also increase cross-disciplinary research."

What does "data" mean in this context? Which data are relevant to share? Who owns the data? Why would people voluntarily share their arduously collected and highly valuable raw material? And who should cover the expenses for maintaining the large data banks?

When the foundation asked survey respondents what the DNRF could do to help researchers make more data available, the answers revealed that the issue is far from simple:

> "It depends upon the type of data. There is no one-size-fits-all here."
>
> – survey respondent

> "The data is very difficult to understand and access without prior knowledge. Therefore, my biggest concern is that it will be used for something that it is not suited for."
>
> – survey respondent

> "The cost of annotating data so they can actually be used – as opposed to just do data-dumps. This is costly and currently the funding of this comes out of research grants."
>
> – survey respondent

Best practice: listen to the researchers
Increased access to data should be pursued wisely. Data need to be of high quality, quality assured, and preferably annotated to optimize reuse. We can optimize excellent research with increased access to high-quality research data by learning from the people who know what works and where the challenges and barriers lie: the researchers and the data managers, the hands-on people.

Professor Liselotte Højgaard,
Chair of the Board of the DNRF

Professor Søren-Peter Olesen,
Director of the DNRF

THE DNRF'S OPEN DATA SURVEY

The DNRF conducted a survey among researchers from all its current Centers of Excellence and among its Niels Bohr Professorships. Out of the 1,175 researchers who were invited to participate, 474 responded to the survey, a response rate of 40%. Among the Centers of Excellence leaders, the response rate was 74%. The purpose of the survey was to clarify how and how much the researchers are using open data and to illuminate the strengths of open data and the barriers to its use.

On the opposite page, opportunities and challenges are summarized. It is characteristic that the list of challenges is twice as long as the list of opportunities, although, at the same time, the DNRF researchers are generally very positive about sharing data. This report is not a thorough analysis, but a contribution to the debate on open access to data with an emphasis on the voice of the researchers.

Motivational factors
Motivation is a key parameter to develop a successful strategy for enabling increased access to high-quality open data.

Generally, it is the DNRF's experience that researchers are idealistic people who are concerned with making their research beneficial to society as a whole. They voluntarily want to share their arduously collected data to maximize public research investments, increase research integrity, and support the generation of new ideas and possible breakthroughs.

> "We want the bowl of candy out in the open, but we don't want people to steal from it."
> – Professor Bo Elberling, Center for Permafrost (CENPERM)

However, researchers aim for quality and fairness. Those who collect the data should be credited and quality must be assured.

INCREASED ACCESS TO HIGH-QUALITY RESEARCH DATA

Opportunities

- Generate collaboration
- Maximize resources
- Strengthen reputations, making it possible to attract high-profile international researchers
- Offer consistency when everybody draws conclusions from the same quality-tested data
- Allow for co-authorships
- Increase research integrity
- Promote efficiency by boosting the reuse of data
- Leverage interdisciplinary research
- Allow competition to be a positive driver

Challenges

- How to finance database maintainance
- Lack of joint infrastructure and financing across Denmark
- Barriers across technology and departments
- Continuity in long-term maintenance
- Resources involved in making quality-tested data available
- Resources involved in making sure that only quality-tested data are being made available
- Long-term storage of data: cutbacks at universities have increased this problem
- Resources involved in advertising that you have data available
- At many institutions, no permanent set-up for depositing datasets
- Belief that sharing data is bad for your competitive edge
- The balance between the pressure to publish and get patents and wanting to make data available
- The concern that data are made available often with vast delay
- Data input available for 25 years, but with no user interface
- Access to comparable data requires specialist knowledge, e.g., that of data managers
- Data validity – can you trust them?
- Concern that open data distort researchers' CVs via publications: data providers get co-authorships on papers they have not actually worked on
- Need to ensure that open data will not compromise ongoing but not-yet-completed projects
- Barriers of ethical issues, sharing of private/personal data and intellectual property rights (IPR), especially challenges during collaborations with industrial partners and universities in terms of IPR

WHAT IS “DATA”?

The DNRF supports working toward increased access to data, but always with a view to a defined and obtainable value of the effort.

Which data are relevant to share and how should they be shared?

> “If I run a Northern blot and thereby obtain data on the expression of a few genes under a specific growth condition for a specific strain of bacteria, is that the kind of data that should be made available? Or is it only large data collection efforts that I should be sharing?”
>
> – survey respondent.

In working with strategies for increased access to high-quality research data, it is imperative to be completely focused and specific about relevance. Billions of euros are being invested in open data, and open data represent opportunities for vast earnings, too. Only data of relevance and quality should be shared.

Fair research data – european open science cloud (eosc)

The European Open Science Cloud is a pan-European initiative that “aims to give Europe a global lead in scientific data infrastructures, to ensure that European scientists reap the full benefits of data-driven science. [...] The European Open Science Cloud will start by federating existing scientific data infrastructures, today scattered across disciplines and Member States.” *(Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. European Cloud Initiative - Building a competitive data and knowledge economy in Europe*, p. 6)

Participation in the EOSC requires FAIR (Findable, Accessible, Interoperable and Reusable) data, i.e., that data and meta data are machine readable, meaning that computers must be capable of accessing a data publication autonomously, unaided by human operators.

From what the DNRF has learned about the challenges with existing scientific data infrastructures in Denmark, there is a great deal of work to be done to fully reap the advantages of this pan-European initiative.

Who owns the data?

> “I am very interested in trying to use open data in future projects. I am, however, concerned that this format will continue a trend of depriving researchers of ownership of their work. It would not be right for employers to profit indefinitely from the work (data) of short-term staff whose careers they do not support.”
>
> – survey respondent.

Those working toward increased access to data should keep in mind the bigger picture and the implications for researchers, e.g., career paths, issues of publication practices, etc. The aim in all areas of this endeavor must be to increase the opportunity to excel in research, to continuously make research better. It is a delicate balance to reassure the individual researchers who took the trouble to collect the data and at the same time strengthen open access for the benefit of all researchers. This balance is dependent on the research area.

The DNRF agrees with this conclusion from *Policy Recommendations for Open Access to Research Data (www.recodeproject.eu):*

“The development of open access to research data needs to be informed by the research practices and processes in the different disciplines and characterized by a partnership approach among key stakeholders. This will help ensure the [necessary] engagement from the wide range of research communities and the embedding of open access within research practice and process.” Thus, the following pages contain five examples of how working with open data can leverage research and what the challenges are in five different research areas.

OPEN ACCESS TO RESEARCH DATA AND DATA MANAGEMENT STRATEGIES IN DENMARK

In Denmark, a number of organizations have developed initiatives to make data available for research. In 2015, the *National Strategy for Data Management*, commissioned by the Danish Rector's College, DeIC (Danish e-Infrastructure Cooperation) and Deff (Denmark's Electronic Research Library), was published. The vision was to ensure Denmark a better and more competitive research environment through the efficient collection, securing, dissemination, and re-use of relevant research data.

The Ministry of Higher Education and Science's website provides an overview of national organizations:

- The Danish Data Catalogue established by the Danish Agency for Digitisation provides an overview of and access to public data.
- The Danish Data Archive (DDA) is part of the Danish National Archives and makes research data based on questionnaires accessible to researchers and students. Grants from the Danish Council for Independent Research to research projects within the health sciences and the social sciences usually have a general obligation to deliver research data to the Danish Data Archive.
- Statens Serum Institut (SSI) is a public enterprise under the Danish Ministry of Health. SSI gathers and disseminates data about the population's state of health and data regarding activity, economy and quality in the Danish health service. The gathered data are made accessible to researchers, but researchers must pay for access.
- Statistics Denmark contains an extensive collection of register data, with data collected from the 1970s to the present. Through the Division of Research Services at Statistics Denmark, authorized research institutions can gain access to data needed to solve specific research and analytical tasks, but researchers must pay for access.

The Ministry of Higher Education and Science has not yet formulated an official strategy for open access to data, and this can – as indicated in the National Strategy for Data Management – prove to be more complex than, e.g., the development of the open access policy from 2012: "The significant differences between the subject area's conditions and challenges in the area of data will probably entail that any uniform and cross-disciplinary policy would be rather generic, and in any case it will probably be supplemented with more specific policies for the individual subject areas" (National Strategy for Data Management, p. 11).

The Ministry of Higher Education and Science supports the European Union Council's conclusions from 2016 on "*The transition towards an open science system*" and the commission's plans to establish the European Open Science Cloud (EOSC). Following this, the Danish Agency for Science and Higher Education is currently preparing an analysis of the costs and benefits and barriers and opportunities of implementing FAIR research data in Denmark.

This analysis is dedicated to the future scenario of machine readable data and meta data. While this scenario will probably rarely be fully realized, there are benefits to be reaped in pursuing it.

OPEN DATA – BRIDGING RESEARCH, MONITORING AND EDUCATION

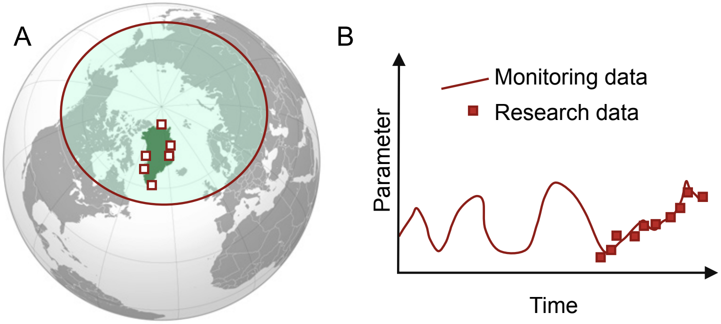
The Center for Permafrost (CENPERM) is working with soil-plant and microbial interactions and the associated ecosystem feedback processes to climate changes across sites in Greenland. One reason for focusing on Greenland was the CENPERM-affiliated researchers' previous research in Greenland but also their involvement in ecosystem monitoring since 1996. These monitoring data have been important for the science that has been completed so far in CENPERM. Similarly, it is important for CENPERM to ensure that its data are available to the public in a useful and quality-tested format.

Most of the processes that CENPERM is studying in Greenland are relevant to upscale in space and time. Since better site-specific data can be extrapolated and made relevant for understanding ecosystem

> "To harvest the value of such a large international study it is crucial that comparisons can be made using similar set-ups, methods, and data format, and that the data are openly accessible."
>
> – Professor Bo Elberling, CoE leader, CENPERM

feedback mechanisms on a larger scale, more exciting results may be obtained, and a higher impact may be achieved when publishing. An example: If the center quantifies an increasing release of carbon dioxide to the atmosphere in a certain vegetation type in West Greenland due to artificial warming, a global network using similar warming for 50 sites across the Arctic is a key to understanding the variability between sites and scaling to the entire Arctic region. To harvest the value of such a large international study, it is crucial that comparisons can be made using similar set-ups, methods, and data format, and that the data are openly accessible.

Open data require resources, and funding has been allocated for that in CENPERM. The center is using a format that has been developed for PROMICE (Programme for Monitoring of the Greenland Ice Sheet) and used by others such as GEM (Greenland Ecosystem Monitoring). Both programs provide important data sources and data deposits for CENPERM. Making data open is also important for teaching. Several theses based on open data have already been completed at CENPERM, which has resulted in international collaboration that would not otherwise have been initiated. CENPERM expects to see its data used and tested, and the center welcomes future challenges to its conclusions.



Upscaling in space and with time is important for CENPERM. Open data is a key component in order to upscale. A: Beyond Greenland we are collaborating with institutions responsible for more than 20 other arctic sites. Some of the work is made in a very consistent way, as for open top chambers for artificial warming, known as ITEX chambers. B: CENPERM data collection started in 2012, but assessing long term climate responses requires access to long time series as made available by DMI and ASIAQ in Greenland.

iCourts
Centre of Excellence for
International Courts

ICOURTS' DATABASE ON INTERNATIONAL COURT DECISIONS

Over the past several years, iCourts has developed the largest currently existing database of decisions of international courts. The center's approach to data-gathering in this regard has generally been the following. First, it has obtained data from a set of specific international court websites/databases, which the center then organized in its own database. In many cases, the center has extracted additional data. The sources used for this purpose are the publicly available databases of international courts, such as HUDOC (*http://hudoc.echr.coe.int*), EUR-Lex (*http://eur-lex.europa.eu*), ICTY (*http://www.icty.org/action/cases/*), IACHR (*http://www.corteidh.or.cr/index.php/en*), etc. In some instances, the center has manually collected all decisions and then entered them in the database in Copenhagen.

An interesting feature of the database is that it organizes and digitizes content from courts, even when the courts themselves are not doing it. For example, the

"The center's ambition is to make this broader database available and to provide a set of digital analysis tools for its future users, through the sequence query-explore-download."

– Professor Mikael Rask Madsen, CoE leader, iCourt

Photo: Ditte Valente/EliteForsk

iCourts' database allows searches in the full text of rulings, even in cases where the corresponding court does not offer this functionality, for example, the International Criminal Tribunal for the Former Yugoslavia or the Inter-American Court of Human Rights. This feature is of great importance to both the legal audience and the greater public, since it facilitates access to legal sources otherwise not available.

Typical problems that the center has faced in these endeavors are data inconsistencies and data entry mistakes, or data not being available in a format that can be processed by machines. Complicating the task further is the fact that the courts' raw data are often organized in complex ways specific to the court in question or that each researcher has very specific interests. Finally, the center has faced the well-known challenge of continuously updating the datasets. This is particularly the case with regard to data that are manually coded on the basis of a qualitative reading of the cases.

The center has generally made the database available to iCourts' researchers or affiliates through the platform at *www.icourts.dk* or in specific forms requested by the researchers. The database is used both internally at iCourts and by iCourts' collaborators at, for instance, the Technical University of Denmark (DTU), the University of Copenhagen (KU), the European University Institute (EUI), Duke University, Northwestern University, and France's National Center for Scientific Research (CNRS). A fraction of the available datasets is currently made publicly available through icourts.dk. Importantly, iCourts' goal is not to replicate a data repository like Harvard's Dataverse, but rather to have a more interactive distribution of the datasets, whereby users are able to navigate and search through the data. Apart from the datasets, the center has made available a number of networks of citations to precedents, related to papers published by iCourts' researchers.

The center's ambition is to make this broader database available and to provide a set of digital analysis tools for its future users, through the sequence query-explore-download. Indicatively, these tools are: a) searches, as described above, b) self-service descriptive statistics, where users will be able to query the database, create their own charts, and save them to their computers, and c) case-law networks, where users will be able to query the database, explore the corresponding case-to-case network, and save it for their own use.

Photo: Carolina Utoft

UrbNet
CENTRE FOR URBAN NETWORK EVOLUTIONS
The Danish National Research Foundation

OPEN DATA – URBNET'S CHALLENGES AND POSSIBILITIES

The Centre for Urban Network Evolutions (UrbNet) is an interdisciplinary center with a strong base in the humanities. While open data and its challenges and possibilities have been topics of discussion in the natural sciences for quite a while, such discussions have only more recently become central to the humanities and begun to impact the way projects are designed and research structured. At an early point in its existence, UrbNet has had to engage with the possibilities and challenges posed by open data because the center works across disciplinary borders and engages with high definition methods. It has had to grapple with the fact that, in some cases, the center works with data from the natural sciences, which are considered open data, and the center needs to combine such data with data from the humanities, which are not necessarily considered open data. Rather, these data are often selected datasets and often only

"It has been possible to embed the center's own larger datasets into a more holistic regional picture, which in turn heightens the quality of the research UrbNet is able to produce."

– Professor Rubina Raja, CoE leader, UrbNet.

Photo: The Danish-German Jerash Northwest Quarter Project partial datasets. One core challenge in such cases is how to compare unequal datasets across disciplines and make meaningful observations.

The fact that open data are always data selected and presented by the publisher/author(s)/generator of the data needs to be underlined. Therefore, open data do not necessarily represent objective or "raw" data, especially not within the humanities. Interpretation may already have been embedded, even unconsciously, through the description of data presented in an open dataset within the humanities. For example, this is the case with datasets of ceramics (one of the core material groups for establishing relative chronologies through typological developments within archaeology), where dating is relative and based exclusively on the excavator's stratigraphy and his or her interpretation of this (if not combined with high precision dating methods). Another challenge with open data within archaeology and the empirical material stemming from archaeological excavations is presented by the sheer amount of material and the resources for processing that it takes to prepare the raw data for presentation. This does not lessen the importance of making data more broadly available. However, it does present challenges to the ways in which such data can be and are made available.

UrbNet has ongoing discussions about best practices within the field of open data, both as users and providers, within the humanities. Within the center, the researchers, among other projects, are working with the Carlsberg Foundation, which has funded a collective research project called "Ceramics in Context," on developing a best practice scenario drawn from an ongoing fieldwork project in Jerash, Jordan. This project provides a full quantification approach, making all data from a six-year excavation project available online and in print.

Furthermore, through mining datasets from other published projects and encouraging colleagues working in related fields to participate and publish in volumes edited by UrbNet and related projects, it has also been possible to embed the center's own larger datasets into a more holistic regional picture, which in turn heightens the quality of the research UrbNet is able to produce. Such projects call for resources involving a large number of man-hours for manual work on databases and data interpretation in general. Therefore, such open data projects present challenges within the humanities where many projects do not have such resources available.

Locally produced pottery from Gerasa/Jerash, Jordan. Every excavated sherd, more than a million, has been registered over five years of excavation and will be made available together with the final publication. The Danish-German Jerash Northwest Quarter Project.

STELLAR ASTROPHYSICS CENTRE

OPEN DATA AT THE STELLAR ASTROPHYSICIS CENTRE

The Stellar Astrophysics Centre's (SAC) open data policies are becoming the norm within astrophysics, supported by the key role played by SAC in making data available to the asteroseismic community.

Hosting data for the international community
Not only is SAC a heavy user of open data, it is also supplying data and facilitating the open sharing of data between international researchers. Beginning with NASA's Kepler mission, SAC set up the Kepler Asteroseismic Science Operations Centre, where one of the prime objectives is distributing the original scientific data from the mission and facilitating the sharing of derived data products through custom-made websites. An example of the standard search interface of the Kepler project can be seen in Figure 1. The set-up consists of a number of different fields to

> "The Stellar Astrophysics Centre's (SAC) open data policies are becoming the norm within astrophysics, supported by the key role played by SAC in making data available to the asteroseismic community."
>
> – Professor Jørgen Christensen-Dalsgaard,
> CoE leader, SAC.

provide search constraints, sorted into a number of tabs. Several parameters have easy-to-use sliders and other helpful graphical user interfaces.

This work has continued with the Stellar Observations Network Group and Transiting Exoplanet Survey Satellite projects, where the center is also creating data archives to be shared with the community. In all cases, an interested user needs only to register as a user (which is free for all) and can download data within seconds.

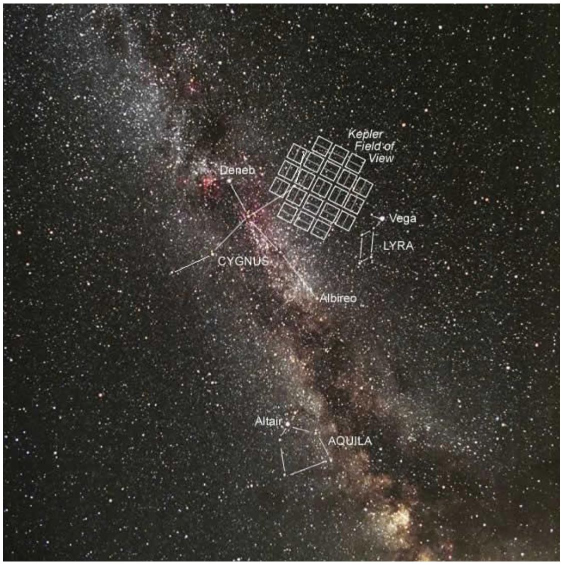
In addition to providing reliable and efficient data access through the Kepler web interface, SAC researchers have also implemented an additional interface for the Seismic Plus Portal, developed as part of the EU-funded SpaceInn project, which uses the Virtual Observatory (VO) format, an open standard of sharing astronomical datasets. Meta data are made available through the VO Table Access Protocol using a dedicated VO server at Aarhus University.

Access to data via general astronomical archive facilities

SAC is using data from a series of available archives. The most used is the European Southern Observatory Archive, containing all raw and calibrated data from these telescopes, available via *http://archive.eso.org/cms/eso-data.html*. Particularly important have been data from the Very Large Telescope and the La Silla facilities. The center also uses archive data from the Nordic Optical Telescope via *http://www.not.iac.es/archive/*, especially the FIES spectrograph. Finally, SAC uses data from NASA facilities (*https://archive.stsci.edu/*), which provide access to data from the Hubble Space Telescope, Kepler, and other NASA/ESA space missions.

Long-term storage of scientific data

In the context of the "Data Management in Practice" project, SAC researchers have been deeply involved in a project with the Danish Royal Library to investigate how scientific data can be stored and archived for long-term preservation (>50 years) while still being useful for active scientific research. An inter-departmental collaboration has led to a prototype archive that could potentially be used for long-term archiving of digital scientific data. The analysis was published in a joint paper (under the Digital Curation Centre). The political aspects of who would be responsible for paying for the maintenance and running costs of such an archive are, however, still unclear.



CENTER FOR PERSONALIZED MEDICINE IN IMMUNE DEFICIENCY (PERSIMUNE)

Health data: How to ensure access without compromising privacy

Given the unique person identifier number, it is possible to combine legacy and contemporary data generated by the health system as part of routine care activities for the entire Danish population. Selection of the population is nil and the completeness of data is high. These data can be combined with more complex data using modern -omics technologies to characterize the host genome and that of invading microorganisms, the profile of immune cells, and the circulating proteins and metabolites, in order to create a systemic biological characterization of the host. In doing so, researchers are able to identify novel pathways for health. As a result, Denmark is in a strong position to productively contribute to the development of personalized care, the new paradigm in modern medicine.

"The creation of the data warehouse with analyzable data elements is the result of a collective and extensive combined effort by medical and programming experts over several years; maintenance of optimal data quality requires continued investment in data sources, since their quality and electronic formats are continuously changing."

– Professor Jens Lundgren, CoE leader, PERSIMUNE.

The combined dataset – stored in a data warehouse – with access to comprehensive computer power can be used for research purposes as well as for real-time support for clinical decision-making as part of routine care.

There are strict regulations for gaining access to data within the data warehouse. These rules help to ensure that an individual's expectations that her/his health data are kept secret and are accessible only to relevant health-care professionals involved in her/his care are met. For research purposes, data can be partly anonymized, but they should still be kept in a secure and controlled space governed by the institution responsible for the data.

Hence, open access to health data is not possible. Rather, the institution responsible for the data warehouse must establish a governance structure that ensures access while not compromising privacy.

These considerations have been resolved for the formation and operations of the PERSIMUNE data warehouse at the Rigshospitalet. PERSIMUNE is able to collate data from a widely diverse set of data sources while still ensuring accessibility to the data for relevant and legitimate interested groups. By preserving ownership and the continued involvement of those responsible for generating the data (clinicians and researchers), PERSIMUNE has made it possible to export data for transcending research projects using a defined governance structure. The creation of the data warehouse with analyzable data elements is the result of a collective and extensive combined effort by medical and programming experts over several years; maintenance of optimal data quality requires continued investment in data sources, since their quality and electronic formats are continuously changing. Therefore, this effort is best centralized as Denmark engages in unfolding a national vision for implementing personalized medicine.

THE DNRF OPEN DATA SURVEY

The DNRF conducted a survey among researchers from all of its current Centers of Excellence and among its Niels Bohr Professorships. Out of the 1,175 researchers who were invited to participate, 474 responded to the survey, a response rate of 40%. All academic positions are represented, with Ph.D. students and post-docs accounting for 61% of the respondents. Approximately 60% of the respondents have Denmark as their home country. The purpose of the survey was to clarify how and how much the researchers are using open data and to illuminate the strengths of open data and the barriers to its use.

FIGURE 1
DISTRIBUTION OF SCIENTIFIC AREAS AMONG RESPONDENTS

All areas are represented. Approximately 50% have backgrounds in the natural sciences and 30% in the life sciences.

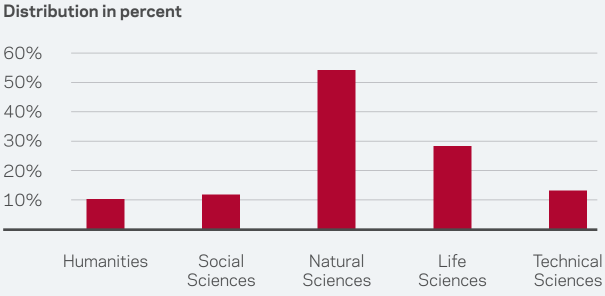


FIGURE 2
DISTRIBUTION OF OPENNESS OF DATA OUTSIDE OF RESEARCHERS' OWN RESEARCH GROUP ACCORDING TO SCIENTIFIC AREA

- ● Humanities
- ● Social Sciences
- ● Natural Sciences
- ● Life Sciences
- ● Technical Sciences

It is most common to make data available within the Natural Sciences and the Humanities.

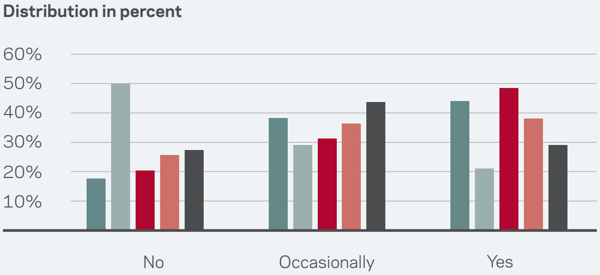


FIGURE 3
DISTRIBUTION OF OPENNESS OF DATA OUTSIDE OF RESEARCHERS' OWN RESEARCH GROUP ACCORDING TO ACADEMIC POSITION

- ● PhD student
- ● Postdoc or assistant professor
- ● Associate professor
- ● Professor MSO
- ● Professor

Professors are most likely and Ph.D. students are less likely to make data available.

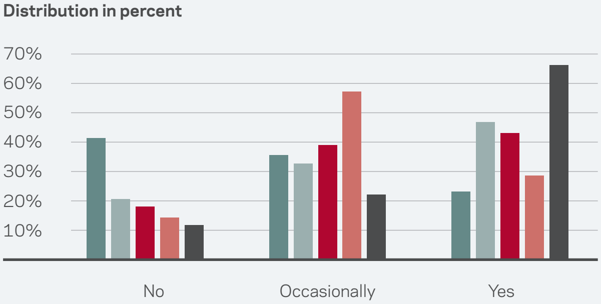


FIGURE 4
NUMBER OF DATA SETS MADE AVAILABLE BY THE RESEARCHERS AMONG THOSE WHO MAKE THEIR DATA AVAILABLE

- ● 0
- ● 1-5
- ● 6-20
- ● 20+

Of those who are making their data available to others, the most common number of shared data is in the 1-5 range.

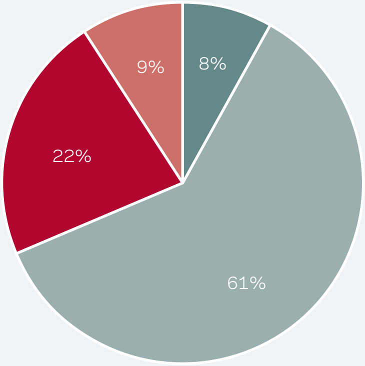


FIGURE 5
THE DISTRIBUTION OF HOW RESPONDENTS MAKE THEIR DATA AVAILABLE

- ● As open access
- ● Upon request via an application procedure
- ● Restrict access to immediate collaborators
- ● Restict access to registered users
- ● Other

The main portion of respondents (45%) make their data available through open access.

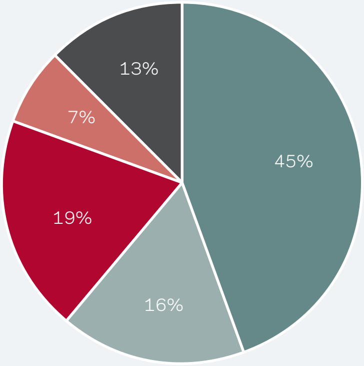


FIGURE 6
DISTRIBUTION OF WHERE THE RESEARCHERS MAKE DATA AVAILABLE

- ● Repository
- ● As supplementary data in journals
- ● Data archive
- ● Project website
- ● Online database
- ● Other online form
- ● Other

Approximately 32% make data available through supplementary data in journals and approximately 20% make data available in a repository.

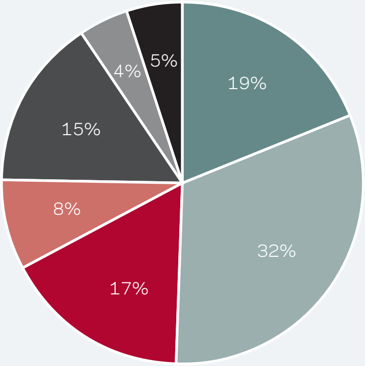


FIGURE 7
THE DISTRIBUTION OF IMPORTANCE AMONG DIFFERENT REASONS FOR MAKING DATA AVAILABLE

The three most important reasons for making data available are: 1. It enables validation and/or replication of data (81% find it very or extremely important). 2. It is good practice to share research data (77% find it very or extremely important). 3. It enables collaboration and contribution by others (73% find it very or extremely important).

● Not at all important
● Slightly important
● Moderately important
● Very important
● Extremely important

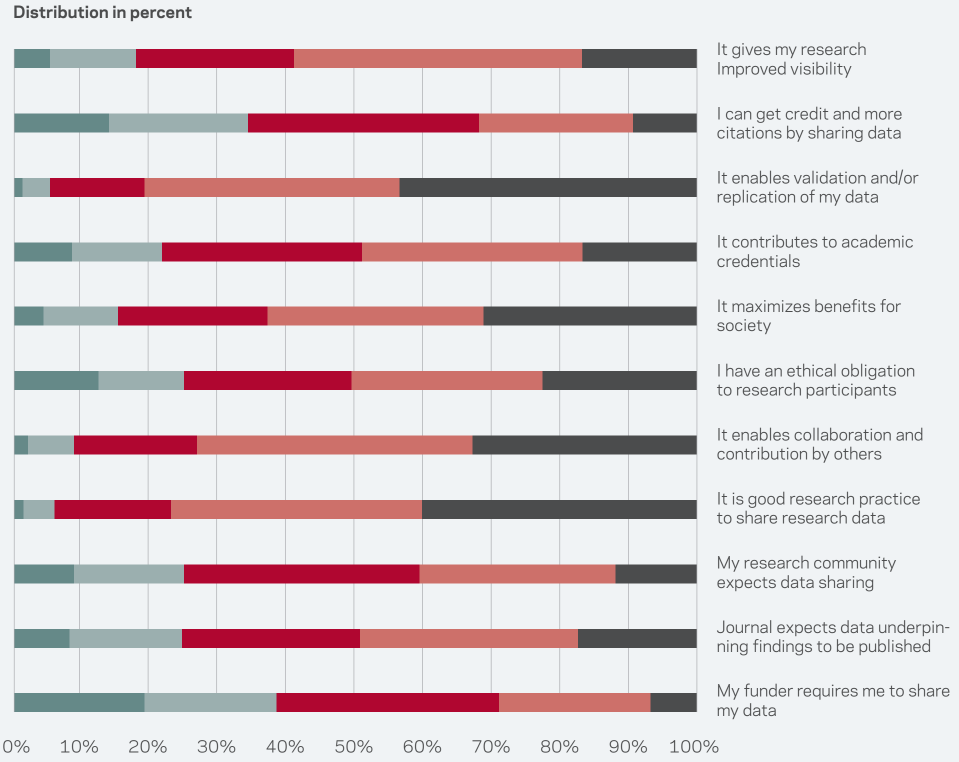


FIGURE 8

The distribution of importance of different barriers to making data available. The two most important barriers to making data available are: “it requires time/effort to prepare data” (46% find this reason very or extremely important) and “loss of publication opportunities” (41% find this reason very or extremely important). Other important barriers are “data contain confidential/sensitive information” (35% find it very or extremely important) and “others may misuse or misinterpret my data” (30% find it very or extremely important).

● Not at all important
● Slightly important
● Moderately important
● Very important
● Extremely important

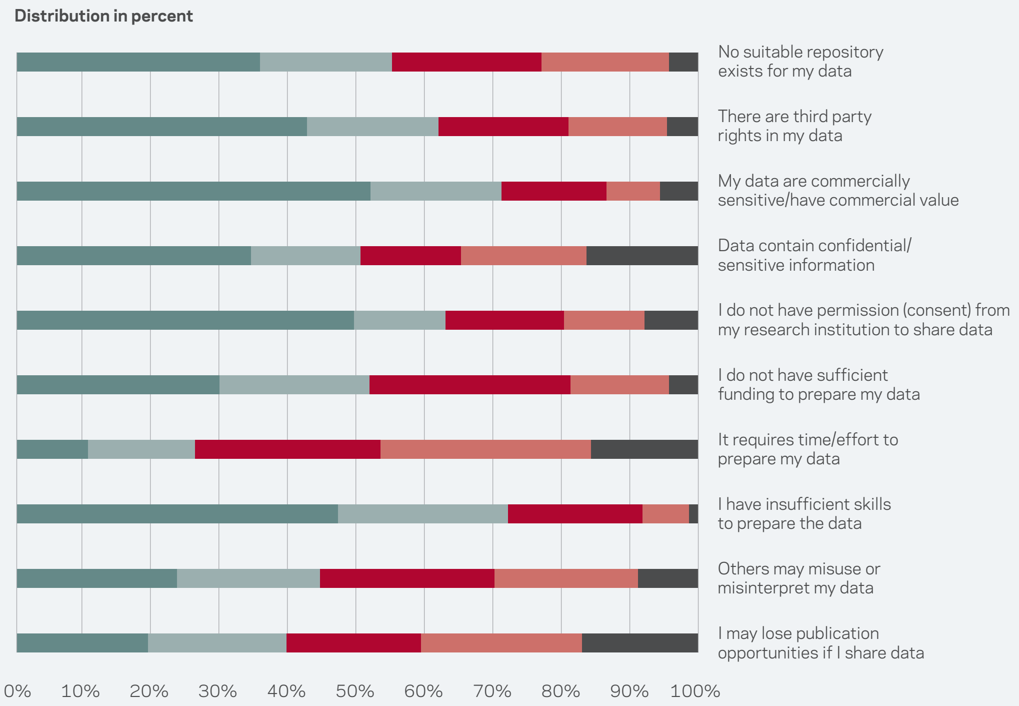


FIGURE 9

Distribution of motivations for making data available. There are five motivators that over 30% of the respondents think are important: Extra funding to cover the cost (35%), Co-authorship on papers resulting from reuse (32%), Knowing how others use the data (32%), Evidence of data citation (32%), and Enhanced academic reputation (31%).

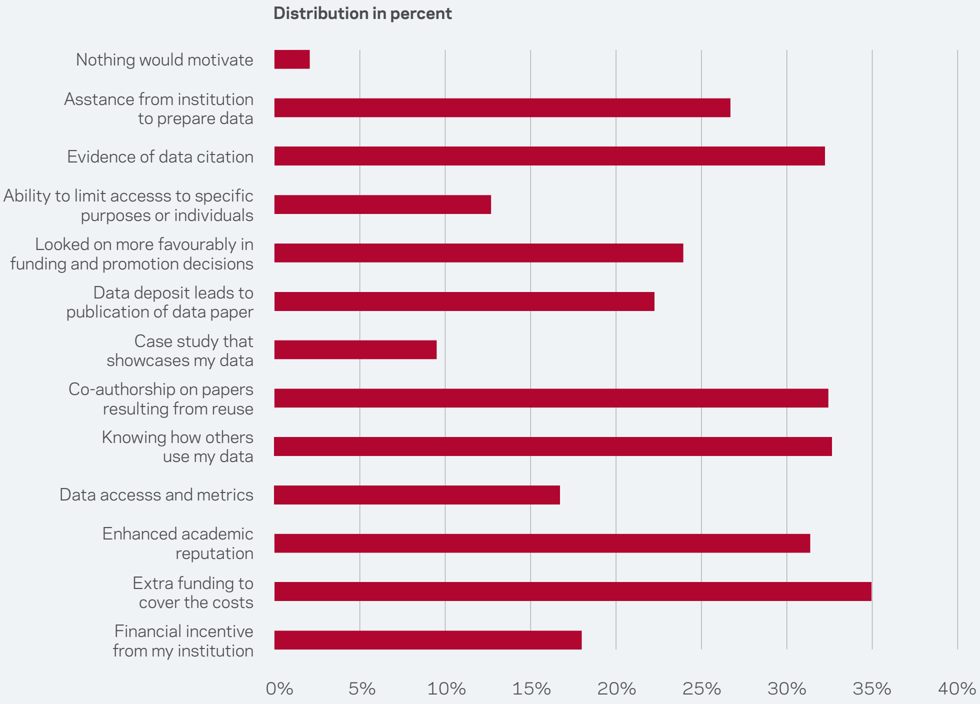
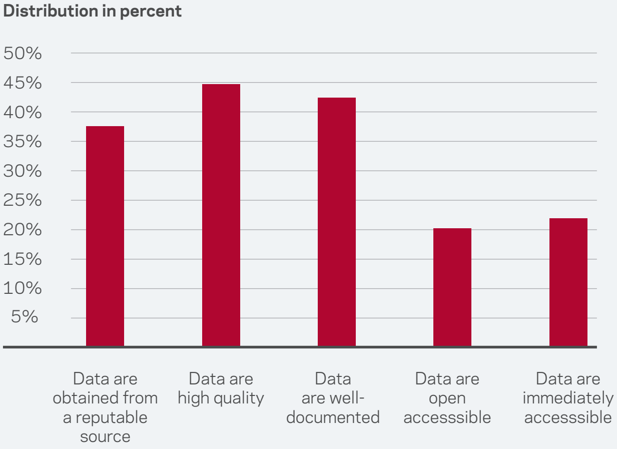


FIGURE 10

Distribution of what is important to the researchers when they themselves use existing data. The three most important reasons are: Data are of high quality (45%), Data are well-documented (42%), and Data are obtained from a reputable source (38%).



SUMMARY OF SURVEY RESULTS

The majority of respondents (app. 60%) make their data available to researchers other than those in their research group. When asked how they make the data available, 45% answered as open access. Whether researchers make their data available seems to be associated with scientific area. Half of the researchers within the social sciences do not make their data available. Making data available was associated with academic position: professors share data more than Ph.D. students. The results of the survey agree with the interviews when looking at reasons, barriers, and motivators for making data available. Validation, good practice, collaboration, and visibility of one's own data are at the top of the range of reasons for making data available. However, the researchers do find some barriers to making data available, including it is time consuming, the data contain sensitive information, someone may misuse the data, or the researcher may lose publication opportunities. When asked what motivates researchers to make data available, survey respondents said that the reasons include extra funding to cover the costs, opportunities for co-authorship, knowledge of how others use the available data, evidence of data citation, and enhanced academic reputation. When researchers themselves use existing data, quality assurance, a reputable source, and data documentation are important.

COMMENTS AND SUGGESTIONS

Research is, to a large extent, carried out at universities. In this sense, the universities play a significant role in promoting the open data agenda wisely. Some departments, researchers, and research centers, such as the Stellar Astrophysics Centre, are front runners in making data available and in developing policies that are becoming the international norm and thereby making Danish research environments attractive to international researchers. However, there is still a great potential for leveraging Danish research in furthering open access to research data. Universities can help this agenda by raising awareness of the opportunities and limitations surrounding open data, by promoting a culture of open science generally, and by developing policies from the researchers' and data managers' best practice scenarios from specific research areas/fields. Open data policies should be dynamic and continuously reflect researchers' needs as well as technological development.

Cross-institutional collaboration

If Danish research is to reap the full benefits of open data, collaborations between and within institutions are essential. Benefits could also be harvested from a centralized organ that systematically addresses legal and ethical issues arising from open access to research data, and institutions could cross-develop training programs on how to make data available, how to reuse data, how to acquire data management skills, etc.

The individual institutions should work toward establishing joint infrastructure and a sustainable long-term system for the curation and preservation of data. Data are, in many cases, very diverse, and it takes a great deal of coordination at the national level to harvest the opportunities embedded in the data. The universities should work together to raise awareness of the potential opportunities in maintaining a joint infrastructure and financing it nationally. In some areas, increased access to open research data has great transformative potential. To exploit the transformative potential, universities could consider linking the promotion of the open data agenda to career development. For example, universities could adopt open access to research data as one criterion, among others, for career progression, but at the same time, universities are encouraged to consider any negative implications for researchers, e.g., career paths, issues involving publication practices, etc., when preparing policies.

The Danish open access agenda

There is no question that funding bodies are key stakeholders in the promotion and implementation of open access to research data. Policies and earmarked funding could affect how data are managed, shared, curated, and preserved.

None of the largest Danish public and private research funding bodies - the Independent Research Fund Denmark, the Innovation Fund Denmark, the DNRF, the Carlsberg Foundation, the Novo Nordisk Foundation – have a policy regarding open access to research data, with the Lundbeck Foundation as an exception. This is fairly common around Europe, with the UK as an exception, probably the fact that it is a difficult area.

Making data available imposes heavy demands on resources. The DNRF recommends that policies and initiatives should be developed by all stakeholders together as a national long-term strategy that ensures comprehensive funding for scalable infrastructure, e.g., establishing joint public databases, and respects the voice of the researchers.

Open access to high-quality research data is an issue of both technical quality and scientific quality that calls for a long-term strategy and the willingness to back the strategy financially and with state-of-the-art data management services and infrastructure for sustainable curation and preservation.

DeIC (Danish e-Infrastructure Cooperation) and Deff (Denmark's Electronic Research Library) commissioned and approved the thoroughly prepared National Strategy for Data Management in 2015, and the National Forum for Research Data Management has brought together people from the Danish universities, the Royal Danish Library, Aarhus University Library, and the Danish National Archives to advance discipline- and researcher-informed initiatives within research data management at universities, and to link these initiatives in a national and international collaboration focusing on knowledge sharing and activities across institutions. The DNRF supports the continuation and further development of this work at a national level.

Suggestions for the researchers
Scientists usually share data either in smaller collaborations where they take advantage of complementary approaches to a problem or in larger consortia where the aim is to obtain many observations in multicenter studies. In both cases, it is a closed club and the upside is co-authorship on higher-impact publications.

When you submit your newly discovered gene to the open GenBank, the upside is that you get the credit for being the first to identify this particular gene. Likewise, when iCourts creates a database that every scientist in the field refers to, iCourts gets the credit. Thus, scientists should make themselves aware of how they can get credit for moving out into the open. Scientists may use databases established by others, which is preferable from an administrative point of view. If they want to create a data bank of their own, they should think long term and explain their need to funders.

When embarking on a database project, scientists should be aware of the legal issues and intellectual property rights (IPR). They should make contracts for IPR upfront, including determining ownership of the data and deciding how to handle sensitive data, and they should seek legal advice.

"It is the recommendation of the DNRF that all stakeholders should collaborate to develop a wise open access policy for data, with the aim of strengthening research and, at the same time, securing the individual researchers' opportunities."

"It is the recommendation of the DNRF that all stakeholders should collaborate to develop a wise open access policy for data, with the aim of strengthening research and, at the same time, securing the individual researchers' opportunities."

Danmarks
Grundforskningsfond
Danish National
Research Foundation

Danish National Research Foundation
Holbergsgade 14, 1
DK-1057 Copenhagen K

T +45 3318 1954
F +45 3315 0626
E dg@dg.dk

www.dg.dk

November 2017