KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

STORAGE AND AVAILABILITY OF DATA FOR RESEARCH

FROM INTENTIONS TO IMPLEMENTATION

ADVISORY REPORT

ROYAL NETHERLANDS ACADEMY OF ARTS AND SCIENCES
PO Box 19121, NL-1000 GC Amsterdam
T+31 (0)20 551 0700
knaw@knaw.nl
www.knaw.nl

Publication available at www.knaw.nl

Layout: Ellen Bouma
Translation: Livewords, Maastricht
Photo cover: depositphotos

ISBN 978-90-6984-747-4

Preferred citation: KNAW (2021). *Storage and availability of data for research. From intentions to implementation.* Amsterdam, KNAW.

STORAGE AND AVAILABILITY OF DATA FOR RESEARCH

FROM INTENTIONS TO IMPLEMENTATION

Royal Netherlands Academy of Arts and Sciences
May 2021

FOREWORD

'Ensure that, in accordance with the FAIR data principles, data be open and accessible to the extent possible, and remain confidential to the extent necessary'. This admonition can be found in the Netherlands Code of Conduct for Research Integrity (VSNU, KNAW, NWO, TO2, VH, 2018), an important document providing both methodological standards ('what does a good researcher do?') and ethical standards ('what should a researcher with integrity do?'). The present advisory report firmly puts researchers centre stage, with their different disciplinary needs, wishes, and practices. It has found that while researchers embrace the idea of storing data and making them available for further research, they struggle with implementation. And this in turn hinders them in the achievement of the FAIR data principles of making data findable, accessible, interoperable and reusable.

The report concludes that on the one hand the Dutch data landscape is vibrant, with key players contributing to solving researchers' data issues. On the other hand, the different institutional levels on which data issues are addressed need to align better, from the leadership of knowledge institutions to the researchers and their day-to-day practices. The report is a call for action: knowledge institutions and funding agencies are urged to take the step from good intentions to implementation. But even that is not enough. The report also recommends that the Ministers of Education, Culture and Science, of Economic Affairs and Climate Policy, and of Health, Welfare and Sport set equal and mutual requirements for data sharing between academia on the one hand and Dutch companies and government organisations on the other. Furthermore, these Departments should offer financial support to cover the costs of data storage and access. All recommendations in the report will help clarify the situation of researchers working with data. But those researchers have a responsibility as well: before starting their projects, they need to clearly identify the support and guidance they need for proper data storage and access. This is not an easy task. Therefore, I am

particularly pleased that Annex 2 provides a checklist with questions that I believe will greatly help young researchers working with data to come to grip with these matters.

It is my hope and expectation that the recommendations of this report point the way towards improved realisation of the FAIR data principles in Dutch research practice.

Ineke Sluiter
President of the Royal Netherlands Academy of Arts and Sciences

SUMMARY

The Netherlands has the ambition to play a prominent role in the quest to optimally use data for research. Dutch knowledge institutions view the sharing of data as a common goal under the umbrella of scientific exchange. Knowledge institutions and funding agencies are taking plenty of initiatives towards making data FAIR (Findable, Accessible, Interoperable, Reusable). Their intentions are good, yet parts of the data landscape remain complex and fragmented. Moving from ideas, suggestions and plans to concrete implementation is a slow process. The present advisory report focusses on the challenges and opportunities surrounding *storage and availability of data for research* from the *point of view of the researcher*, and *digital data* in particular. It aims to streamline plans and intentions to make data FAIR, in line with the needs and wishes of researchers.

The report observes that different fields of science have widely varying requirements regarding storage of and access to data. Multiple factors play a role, including the type of data (e.g. personal or non-personal), the size of the data, their usefulness for other researchers, the cost of storage, and the possibility to reproduce the data or not. In all cases data storage and access come with challenges. Specific challenges lie within the ownership of data (e.g. privacy issues), and the availability of national infrastructures for centralised or federated data storage. Common challenges concern the training of researchers, identifying proper metadata and data access methods, providing better support to researchers working with data, and creating a new culture in which data storage and access are recognised and awarded. Among researchers there is a widespread need for more guidance on a national level, to help them make data more *operable* and *interoperable*.

Progress in data storage and data sharing depends on close collaboration between researchers, data stewards in a facilitating role, management of knowledge institutions providing infrastructural support, management of knowledge institutions and funding agencies encouraging researchers to adhere to proper data storage and access practices, and ministries offering financial support.

The report urges knowledge institutions to take *action*. Management must challenge the different fields of science to make clear where the support available to researchers needs to be strengthened. Examples include removing legal and ethical hurdles, making training programmes available, and creating ICT-like services that enable researchers to easily and securely deposit, share, publish and preserve research data during all stages of their research project. The report also offers a *checklist* of questions for researchers working with data. This checklist can assist them in preparing a comprehensive Data Management Plan.

Among researchers working with data there is plenty of enthusiasm to store data and make data available for research. However, this comes with a price tag. The costs are considerable, and it must be clear from the start who picks up the bill: the ministries, the funding agencies, the knowledge institutions, or a combination. The report recommends that knowledge institutions identify how much local and national data storage and support is needed, to what extent the available data infrastructures are sufficient, and what the costs of data storage and support are. Not acting is not an option, because there are also substantial costs involved in the loss of data. Data have intrinsic value, for example, for reproducibility or for complementary studies.

Other recommendations include:

- Realise centralised or federated data storage for those fields of science where data originate from unique and expensive observation equipment.
- Incentivise researchers to follow proper data storage practices by creating a culture in which storage and access are valued. Put role models in the spotlight and reward researchers for making the data available.
- Set equal and mutual requirements for sharing of data between academia on the one hand and Dutch companies and government organisations on the other.

SAMENVATTING

Nederland wil een voortrekkersrol spelen als het gaat om optimaal gebruik van data voor onderzoek. Nederlandse kennisinstellingen zien het delen van data in het kader van wetenschappelijke uitwisseling als een belangrijk gemeenschappelijk doel. Er zijn tal van initiatieven binnen kennisinstellingen en onderzoekfinanciers om data zo FAIR (Findable, Accessible, Interoperable, Reusable) mogelijk te maken. Dat zijn goede initiatieven met goede bedoelingen, maar het datalandschap is vaak complex en gefragmenteerd. De concrete omzetting van deze ideeën, suggesties en plannen is vaak een lang en moeizaam proces. Dit adviesrapport spitst zich toe op de uitdagingen en kansen rond *opslag en beschikbaarheid van data voor onderzoek* vanuit het *oogpunt van de wetenschapper*. Daarbij gaat het met name om *digitale data*. Hierbij is vooral gekeken naar de vraag hoe plannen en ideeën om data FAIR te maken, kunnen worden gestroomlijnd en afgestemd op de wensen en behoeften van wetenschappers.

In het rapport wordt vastgesteld dat de behoeften op het gebied van opslag van en toegang tot data sterk verschillen per wetenschapsgebied. Daarbij spelen verschillende factoren een rol, waaronder het soort data (gaat het bijvoorbeeld om persoonsgegevens of niet), de omvang van de datasets, hun bruikbaarheid voor andere wetenschappers, de kosten van opslag, en de mogelijkheid om data al dan niet te reproduceren. Wat echter in alle gevallen geldt, is dat opslag en toegang altijd gepaard gaan met uitdagingen. Specifieke uitdagingen liggen op het gebied van de eigendom van gegevens (zoals privacykwesties) en de beschikbaarheid van nationale infrastructuren voor centrale of federatieve dataopslag. Dan is er nog een aantal algemene uitdagingen, zoals het opleiden van wetenschappers op dit gebied, het vaststellen van de juiste metadata en methoden voor datatoegang, betere ondersteuning van wetenschappers die met data werken en het tot stand brengen van een cultuur waarin het belang van opslag van en toegang tot data wordt erkend. Wetenschappers hebben sterke behoefte aan meer sturing op nationaal niveau om de data beter *(inter)operabel* te maken.

Wil het opslaan en delen van data werkelijk van de grond komen, dan is nauwe samenwerking essentieel: tussen wetenschappers onderling, met datastewards in een faciliterende rol, met de directies van kennisinstellingen die infrastructurele ondersteuning bieden, met de directies van kennisinstellingen en onderzoekfinanciers die wetenschappers stimuleren om te zorgen voor goede opslag van en toegang tot data, en met de ministeries die financiële steun bieden.

In het rapport worden kennisinstellingen opgeroepen *actie* te ondernemen. Zij moeten nagaan hoe wetenschappers op de verschillende wetenschapsgebieden beter ondersteund kunnen worden. Te denken valt aan het wegnemen van juridische en ethische obstakels, het faciliteren van opleidingsprogramma's en het creëren van ICT-achtige diensten waarmee wetenschappers tijdens elke fase van hun onderzoeksproject data gemakkelijk en veilig kunnen opslaan, delen en publiceren. Het adviesrapport bevat ook een *checklist* met vragen voor wetenschappers die met data werken. Deze checklist kan hen helpen om een integraal Data Management Plan op te zetten.

Onder wetenschappers die met data werken, bestaat groot enthousiasme om deze data op te slaan en beschikbaar te maken voor onderzoek. Maar daar hangt wel een (aanzienlijk) prijskaartje aan. Vanaf het begin moet duidelijk zijn wie die kosten betaalt: de ministeries, de onderzoekfinanciers, de kennisinstellingen of een combinatie van deze partijen. Een van de aanbevelingen van het rapport is dat de kennisinstellingen in kaart moeten brengen hoeveel lokale en nationale dataopslag en -ondersteuning nodig is, in hoeverre de huidige data-infrastructuren toereikend zijn, en wat de kosten van dataopslag en -ondersteuning zijn. Niets doen is geen optie: ook verlies van data brengt aanzienlijke kosten met zich mee. Data hebben een intrinsieke waarde, bijvoorbeeld voor de reproduceerbaarheid of voor aanvullende studies.

Enkele andere aanbevelingen uit het adviesrapport zijn:

- Realiseer centrale of federatieve dataopslag voor die wetenschapsgebieden waar de data afkomstig zijn van unieke en dure observatieapparatuur.
- Stimuleer wetenschappers om data op de juiste wijze op te slaan door een cultuur te creëren waar belang wordt gehecht aan een goede opslag van en toegang tot data. Zet rolmodellen in de schijnwerpers en beloon wetenschappers als zij data beschikbaar stellen.
- Stel dezelfde, wederzijdse verplichtingen aan het delen van data tussen de wetenschappelijke wereld aan de ene kant en het Nederlandse bedrijfsleven en overheidsorganisaties aan de andere kant.

CONTENTS

1. INTRODUCTION

1.1 Background

Data are a key element of academic research. In empirical research, data are collected, analysed, stored and archived. Researchers also make data available to their colleagues and to other interested people and organisations. Access to data contributes to the advancement of science, because it allows researchers to check findings and build critically on work by others. *Sharing of data is an integral part of academic knowledge exchange and Open Science.* Moreover, it triggers harvesting of data by other researchers and exchange of ideas, to mutual benefits. Over the full range of science there is a need and a demand for making data better available, where possible according to the FAIR principles (Findable, Accessible, Interoperable, Reusable). Policy makers recognise needs and opportunities for making data better available as well. In 2019 the Dutch minister of Education, Culture and Science declared Open Science in research, including reuse of research data, a priority for the coming four years (minister of Education, Culture and Science, 2019).

Over the past few decades, advances in information and communication technology in combination with ongoing globalisation have been major contributing factors to increased storage of research data in many fields of science. This aim for FAIR data and Open Science was supported by initiatives such as the National Platform Open Science[1] (NPOS) in the Netherlands and the European Open Science Cloud[2] (EOSC),

1 NPOS (in Dutch: *Nationaal Platform Open Science*) is a collaboration of 17 Dutch organisations of higher education and research that have the intent to realise Open Science. Among its members are the KNAW, NWO, the VSNU, and the National Library of the Netherlands (KB). The platform brings together parties that have initiated, formulated or supported the *National Plan Open Science*, which was presented to the Dutch government in February 2017.

2 The EOSC is a European Commission initiative aiming at developing an infrastructure providing its users with services promoting Open Science practices. The envisaged infrastructure is built by aggregating services provided by several providers. The EOSC was officially launched in November 2018.

and networks such as GO FAIR[3] and the Research Data Alliance[4] (RDA).

In some fields of science, researchers have developed good practices of data access and data reuse. In other fields, researchers are still struggling with making data available. Even within a single field the situation may vary. There is a need for clear strategies and for cross-fertilisation.

The present advisory report stems from an expert meeting on the improvement of storage and availability of data for research that was held on 4 December 2018, in the Trippenhuis in Amsterdam. This meeting was attended by 50 researchers. The conclusion was that a KNAW advisory report may help to give more structure and transparency to the discussion on improving the storage and availability of data for research. During the meeting a number of points were raised:

1. Academic researchers must be central in the advisory report. Attention is needed for both the principal and the practical aspects with which the researcher is confronted. From the perspective of the researcher, attention is needed for aspects such as management, legal rules, sustainability, privacy, security, training and embedding, as well as access to storage facilities. Developing a checklist that researchers can use when handling data for research may be helpful.
2. Justice must be done to different cultures in different fields of science. At the same time, workable common denominators must be found.
3. Central questions are: Which data need to be stored, and which do not? For how long? Who decides this? Who owns the data? Who shares data and who does not? In what format? Why is sharing of data useful? What are the benefits and what are the drawbacks, both scientifically and financially? Do data have an objective reality, or could their meaning or significance change over time? To answer these questions, it is not only necessary to look forward, to the data of the future, but also backward, to the data that already exist.
4. Because data sets need to be accessible whenever possible, the question arises whether a central facility for the archiving of data is desirable and feasible. What would the financial contours of such an enterprise be, and what its long-term obligations? Who has final responsibility? What are the international perspectives of such an initiative? How would this fit into the plans and initiatives of other countries?

3 GO FAIR, established in 2014, is a bottom-up, stakeholder-driven and self-governed initiative that aims to implement the FAIR data principles. It offers an open and inclusive ecosystem for individuals, institutions and organisations working together through Implementation Networks (INs). The founding member states of GO FAIR are France, Germany and the Netherlands.

4 The RDA is a research community organisation started in 2013 by the European Commission, the American National Science Foundation and National Institute of Standards and Technology, and the Australian Department of Innovation. Its mission is to build social and technical bridges to enable open sharing of data.

5. From the perspective of the researcher, the role of 'data stewards' needs to be better specified.
6. A close cooperation between public knowledge institutions, NWO and ZonMW (as funding agencies of research performed by public knowledge institutions in the Netherlands) and the NPOS was seen as desirable for the anchoring of the rapid developments around data.

The Board of the KNAW established the committee 'Storage and Availability of Data for Research' on 26 February 2019 (Annex 1). Armed with the outcome of the expert meeting on 4 December 2018, the committee set to work in April 2019. Due to COVID-19, much of the committee work was suspended from March 2020 until December 2020.

1.2 Research perspective on data

The first question the committee asked itself was: *What exactly is meant by 'data', 'storage' and 'availability' from the research point of view?*

For each of these, both generic and specific answers can be given.

DATA

Generic:
- Data comprise everything that serves as input for information, knowledge and insight.
- Data comprise everything from measuring to editing.
- Data can be raw, processed or published.[5]
- Data can be personal or non-personal, open or non-open, real or computer-generated.
- Data are observations with which theories can be tested (deductive reasoning) or built (inductive reasoning).

Specific:
- Data are input for research, output of simulations, or used for training purposes.
- Both future data and past data can be relevant.
- Data are mainstream in some fields of science, and less so in others.

5 Raw data come from, for example, an instrument, a survey or an archive. From the viewpoint of the researcher, raw data are unedited. After the raw data have been edited (correcting, calibrating, transcribing, etc.), they become processed data. The subset of processed data used by the researcher for scientific publication is referred to as published data.

STORAGE

Generic:
- Data are stored in such a way that research can be replicated.
- Data are stored in such a way that other researchers can reuse them.
- A spectrum of storage forms exists: digital, on paper, archives, samples, etc.
- Not all data need to be stored.

Specific:
- What is stored determines where and how it is stored.
- There is a difference between 'storing' and 'storing well'. Well-stored data can be reused by others without hurdles.
- Data may be stored either short-term (10-25 years) or long-term (more than 25 years).

AVAILABILITY

Generic:
- Some data are available at any time, possibly up to a certain time horizon.
- Other data are available upon request only, and under certain conditions to be decided by the owner.

Specific:
- Researchers who wish to reuse the data should be charged no more than the cost of the data provision itself (KNAW, 2019).
- The costs of storing data and making data available for research should be covered via the first and second flow of funds,[6] including back-up, access management and security.
- Governance should be provided by the home knowledge institution.

The second question the committee asked itself was: *What are the provisions and structures within the Dutch knowledge system that assist researchers working with data?*

The final report on the project *Exploring and optimising the Dutch data landscape* (2020) of the NPOS answers this question as follows in its Summary:

6 In the Netherlands, the first flow of funds relates to programming and core funding of research at universities and institutes, and the second flow of funds to programming and funding by national funding agencies NWO and ZonMW.

It proved difficult to acquire a detailed and exhaustive overview, on the one hand because the subject 'data landscape' or 'data services' is not clearly delineated and on the other because many matters are in fact organised, but not always in an obvious way. Our conclusions, in brief, are as follows. The Netherlands has an abundant but fragmented and complex data landscape, not only in terms of data services but also with regard to the development and dissemination of knowledge concerning research data management. The risk is that there will be overlaps and inefficiency and that we will miss out on opportunities for connectivity and innovation. Researchers themselves say that they lack an overview.

The NPOS report also makes clear that the Dutch knowledge institutions are moving forward in sorting out data policies, setting up data infrastructures in order to support researchers working with data, and getting the proper information directly to the researchers. Almost every knowledge institution has founded data centres (with different titles) that support researchers working with data. Nowadays, the following support structures for researchers are common within knowledge institutions in the Netherlands:

- Ethical review boards have been put into place that monitor data management and consent forms.
- Most libraries within the home knowledge institution provide in-person (via data stewards[7]) and web-based (via portals) advice on data management to researchers who need to store data and have to make data available for reuse by others.
- Researchers can ask for advice on specific topics related to research data management within the home knowledge institution. Some of that advice is aimed at setting up the *Data Management Plans* (DMPs) that are mandatory for NWO and EU research projects (source: https://www.nwo.nl/en/research-data-management).
- Handouts (source: https://www.lcrdm.nl/rdm-advies-tips) and narratives (source: https://www.uu.nl/en/research/research-data-management/rdm-stories) can help researchers navigate through the complex data landscape.

7 Data stewards are employees of knowledge institutions who are responsible for ensuring that data are properly managed and ready to be used by researchers. They are in charge of making sure that data are clean, properly stored and available for analysis. They help researchers to:

1. Create an inventory of data.
2. Know what the data can mean for the research.
3. Know which databases are available.
4. Know whether or not to trust the data, assess the data quality, and identify where the data are sourced from.
5. Know what to use the data for, and be aware of potential data regulations.

Also, the large Research Infrastructure (RI) programmes are designed to support the researcher. Examples are Health-RI[8] and ODISSEI.[9]

There are several commercial parties that offer research data services to researchers working with data. For example, all the major academic publishers active in the Netherlands furnish data services in addition to their catalogue of publications. Elsevier has the Mendeley Data service, Springer Nature cooperates with Figshare, and Wiley with Dryad. Academic publishers generally claim to support Open Science and Open Data, and say that they are offering their services to make sharing of data easy for researchers. In many cases, they do not charge researchers for using and depositing data sets in their repository (with certain limitations, such as the format and the size of the data set). Yet, researchers do not always know what happens to the data once they have deposited them (NPOS, 2020).

The Dutch data landscape is vibrant, and there are clear points of contact and pivots that contribute towards solving data issues of researchers working with data. In many ways the Netherlands takes a prominent role in the quest to optimally use data for research. For example, the Dutch participation in the EOSC is strong, the Dutch GO FAIR activities are intense, and the large Research Infrastructure (RI) programmes were initiated in the Netherlands. The report *Six recommendations for implementation of FAIR practice* (European Commission, 2020) states:

> *Overall, we found that within Europe, Western European countries, in particular, the Netherlands, UK, France and Germany, are in the lead when it comes to FAIR practice.* (p. 13)

Still, there is a gap to be bridged between the ongoing developments within the knowledge institutions and the day-to-day practices of researchers working with data. Examples of initiatives developed to bridge this gap are:

- *Further development of locally based Digital Competence Centres (DCCs)*
  National funding agency NWO provides seed money to public knowledge institutions (universities, academic medical centres, and NWO and KNAW institutes) for the establishment or further development of a locally based Digital Competence Centre (DCC). The funding is intended to stimulate knowledge

8 Health-RI, established in 2015, is the Dutch non-profit foundation supporting a public-private partnership of organisations that want to realise a national health-data infrastructure. More than 70 organisations involved in health research and health care endorse their efforts.

9 ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations), established in October 2016, is the national research infrastructure for the social sciences in the Netherlands. ODISSEI brings together researchers having the necessary data, expertise and resources to conduct ground-breaking research and embrace the computational turn in social enquiry.

institutions to further develop an existing centralised DCC, or to set up a DCC at a central level within the institution (source: https://www.nwo.nl/en/calls/local-digital-compentence-centres).

- *Further development of ICT-like services for researchers*
  Some research institutes and research groups have created an ICT-like service that enables researchers to easily and securely deposit, share, publish and preserve the data during all stages of the research project. Such services allow researchers to automatically store research data in the format of their choice and make the data fully findable, including metadata. As a consequence, the stored data can be analysed more efficiently (QR codes on labels, etc.). The service also eases the back search of older data, because the researcher no longer depends on memory sticks, the memory of former employees, etc. This motivates researchers to make use of the service. Some of these services are general, others are domain-specific and exist at the level of the research institute.

1.3 Focus and target audience

There is a certain friction between top-down implementations, enforced by funding agencies and management of knowledge institutions, and bottom-up needs and wishes of the researcher. The committee decided to voice the needs of *researchers working in public knowledge institutions in the Netherlands and working with data*, from historians to economists, from medical doctors to biologists, and from chemists to astronomers. The primary goal of the committee is *to assist them in finding their way through the maze of challenges, opportunities, restrictions and obstacles they face when dealing with storage of data and making these available to others.* Knowledge institutions and funding agencies in the Netherlands are working hard to address the needs of researchers working with data. With the present advisory report the committee hopes to contribute to a better streamlining of plans and intentions, in line with the needs and wishes of researchers.

The report is written for individual researchers, for the academic community, for policy- and decision-makers, for the main funding agencies in the Netherlands, and for the ministers of Education, Culture and Science, of Economic Affairs and Climate Policy and of Health, Welfare and Sport. The focus is on storage and availability of data for research, viewed as *operational challenges* under the larger umbrella of the goal of FAIR data and Open Science. Because the FAIR principles relate to *digital data* only, the committee decided to focus on those. That said, some of the recommendations may also apply to non-digital data (which are equally valuable for academic research), especially those that can be accessed through metadata in the digital world.

The focus of the present advisory report is on how data management in the Netherlands can be stimulated, supported and facilitated, how a better understanding of the advantages of data sharing can be attained, and how a better demarcation of roles, responsibilities and tasks can be achieved. Not all questions raised during the expert meeting on 4 December 2018, are answered.

1.4 Approach of the committee

The committee (see Annex 1) began to discuss at length the scope, breadth and limitation of the advisory report. This led to a description of what data storage and availability comprises, and what the primary tasks of researchers working with data are. This, in turn, led to the compilation of a checklist of questions for researchers, to be considered before the start of their research project (see Annex 2). This checklist can help to prepare a Data Management Plan and to seek for concrete solutions to problems that may arise during the research project.

The discussions in the committee were followed by two meetings with researchers working with data in knowledge institutions in the Netherlands. The committee further consulted experts (Annex 3 contains a full list of names) and studied available literature (Annexes 4 and 5 summarise selected reports). A preliminary draft of the report was read by a selection of KNAW members (included in Annex 3).

Prior to completion, the committee informed the KNAW Board on its progress, discussed its main findings, and took in further suggestions. At the request of the Board, a draft of the advisory report was reviewed by selected reviewers and by the Academy's advisory councils. The Board discussed an updated draft with the chair of the committee on 30 March 2021. Based on that discussion, a final update was prepared, which the Board subsequently approved.

1.5 Outline of the report

Section 2 summarises *key findings collected during two expert group meetings*: one on 'personal data' on 26 November 2020, and one on 'non-personal data' on 17 December 2020. Section 3 raises several urgent issues around storage and availability of data for research, leading to a *call for action by knowledge institutions and funding agencies*. Section 4 offers a number of recommendations for *improving* storage and availability of data for research, and for making data more *operable* and more *interoperable*. Annexes 1-5 contain supporting information (see Section 1.4).

Given that data are shared worldwide, the views, analyses and recommendations offered in this report may be valuable to researchers in other countries as well.

Especially within the broader EU perspective, there is plenty of opportunity to learn from each other, for example under the umbrella of ALLEA.[10] (Annex 4 contains a summary of selected reports addressing needs, hurdles and opportunities related to storage and availability of data for research in an international context.)

10 ALLEA is the European Federation of Academies of Sciences and Humanities, representing more than 50 academies from over 40 EU and non-EU countries. Since its foundation in 1994, ALLEA speaks out on behalf of its members on the European and international stages, promotes science as a global public good, and facilitates scientific collaboration across borders and disciplines.

2. FINDINGS FROM THE EXPERT GROUP MEETINGS

2.1 Consultation

The committee consulted the field in two expert group meetings. The first meeting took place on 26 November 2020 and was attended by researchers working with *personal data*. The second meeting took place on 17 December 2020 and was attended by researchers working with *non-personal data*. The participants had different disciplinary backgrounds and different levels of seniority within Dutch academia, but were all experienced researchers working with data. In each meeting, four groups of participants presented 'obstacles' encountered while working with data, indicated 'best practices', and made suggestions to improve storage and availability of data for research in the Netherlands. Afterwards the committee exchanged views with the participants based on a draft version of the checklist in Annex 2.

The committee and the participants discussed a number of hurdles, both principal and practical, that researchers are confronted with when storing data and making data available for reuse by others. A number of hurdles, as well as some suggestions to surmount them, are listed in Section 2.3. Section 2.2 presents the main findings of the meetings.

2.2 Main findings

One of the main findings of the meetings is that there is a *huge diversity* in the way data are shared with other researchers. There are differences between fields, and even within fields. At the same time, there is strong agreement on a number of core issues.

- All participants emphasise that FAIR (Findable, Accessible, Interoperable, Reusable) is set as the standard (for raw, processed and published data). However, making data fully FAIR is difficult, if not impossible. There is little consensus on the minimal sets of labels to be attached to the data so that they can be made FAIR. The researchers in fields of science define with their peers what the criteria for FAIR-ness are in their particular field. Across fields *different consent mechanisms and different conditions for use* exist for accessing data. In some cases, these mechanisms are based purely on scientific collaboration or are controlled by capacity considerations. In other cases, data donors put limits on how researchers can use, share and publish data, either because of commercial or privacy reasons or to protect trade secrets. Furthermore, both for primary and secondary data collection, often *consent mechanisms are governed by law*, for example because of privacy-sensitivity of the data. Storing the data for sharing purposes can clash with the General Data Protection Regulation (GDPR), for example in the medical sciences, the social sciences, the humanities and the technical sciences (where data about user or developer behaviour are exploited). Although techniques such as anonymization, pseudonymization and privacy-protected computing are being developed, *identifying the conditions of use and consent are hard nuts to crack for the individual researcher*.
- The prevailing consensus is that data storage is useful only if the data can be *located* and *accessed*. *Metadata are crucial* for finding the data, and for understanding the conditions and the settings under which the data were collected or generated. All participants agree that harmonised metadata standards do not exist and will never exist, and perhaps should not exist at all.
- Sometimes proper metadata are missing. Consequently, the *visibility* of some data sets is low, and therefore these data sets are difficult to find.
- The participants mention that data availability can be poor due to, for example, licenses and/or commercial interests. There is an inherent *asymmetry* between, on the one hand, academia striving for full sharing of data and, on the other hand, companies and government organisations not being willing or not being able to fully share data. Consequently, data cannot always be stored properly for reproducibility studies.
- Participants of the meetings who are already storing data for access by others generally agree that their activities are often experienced as auxiliary and time consuming, and that these activities may even jeopardise their main responsibilities as a researcher. Motivation is lacking and financial hurdles exist. Still, many participants think that it is important 'to start small and to start today'. In this way, researchers can come across tried-and-tested solutions of other researchers and implement these in their own data handling practice. Consulting bottom-up online communities of researchers may also help.
- Storing data and making data available for research often become critical when the results of the research project are published. That is often the right moment to safeguard the data and to make them FAIR, but this does not always have a high

priority for the researcher. Furthermore, it is not always clear which data need to be stored, and which data do not. The answer depends on the type of data (raw, processed or published; personal or non-personal; etc.).
- The various responsibilities of the individual researcher and his/her research group for data generation and data storage are unclear. There is need for *recognition and reward* of the work of academic staff involved in storage and availability of data for research.
- The participants consider storing data and making data available for research to be a part of a larger *ecosystem of scholarly output*.

2.3 Personal data versus non-personal data

During the expert group meeting with participants working with *personal data* a number of obstacles were identified:

- *Legal hurdles* prevent proper generation and storage of personal data. An example is the lack of clarity about what specific flexibilities are provided by the General Data Protection Regulation[11] (GDPR) with respect to the generating and storing of data. There is a need for simple procedures that are understood and embraced by everyone involved. The ethical review board of the home knowledge institution may be instrumental in advising the researcher on aspects of data management before the start of the research project.
- The different *levels of consent* for personal data sharing are unclear. There is a need for a national register that assigns different levels of consent to the data.
- In the handling of personal data, a number of ethical aspects are involved. However, researchers are often confronted with, and confused by, *contradictory views* on these aspects. The participants of the expert meeting noted that the many delays and uncertainties encountered along the way are frustrating. There is again a need for simple procedures as to what is allowed and what is not.
- Personal *data transfer* is becoming difficult. It is unclear who decides what is allowed. 'Central data warehousing' is not considered feasible: on the national level it will be difficult, on the international level it will be impossible. The best solution for the difficulties related to storing and transferring data will have to be

11 An important objective behind the GDPR is to promote the free flow of data across the European Union, in a standardising role. Under the research exceptions, for example, there is also considerable leeway to the way personal data can be used. The GDPR works with broad norms leaving some leeway for member states to concretise certain aspects of data processing, including the reach and practical implementation of research questions. Some of the difficulties mentioned during the expert group meeting are caused by the lack of concretisation at the national level. Moreover, privacy is contextual, as is data protection law. The challenge is to make sure that researchers have easy access to relevant expertise that can advise them on responsible data practices, for example in the form of a Code of Responsible Data Use in Research.

developed separately for each individual field of science.
- The handling of personal data can be time consuming and costly. Sometimes robust support by the home knowledge institution is lacking, and not every available infrastructure is suitable for handling the data in a proper manner.
- Often, *training* in personal data handling and storage is lacking.

Several obstacles were also identified in the expert group meeting with participants working with *non-personal data*:

- Non-personal data sets can have very large sizes, in which case it is time consuming for the researcher to organise the data for reuse. Although institutionally or nationally centralised or federated storage of data is common for non-personal data, there is a need for more *training* of researchers and more support. Storing and maintaining non-personal data come at a high cost, and it is not always clear for the researcher who picks up the bill.
- There is *lack of data-engineering expertise*. Private organisations and businesses are fully aware of the high value of research software engineering expertise, and are therefore already investing in it. Public knowledge institutions must step up the effort. Research software engineers develop, construct, test and maintain architectures, such as databases and large-scale processing systems. Research software engineers differ from data stewards and data scientists in that they clean, 'massage' and organise (big) data. Consequently, research software engineers depend on a different skill set than data stewards or data scientists.

2.4 Conclusions

The committee concludes that different fields of science have *widely varying requirements* regarding the storage of and access to data.

In some fields, data are volatile and storage on a local computer is sufficient, either because the data can be easily reproduced or regenerated (when processed data are sufficient), or because there is no strong interest from other researchers in the data. In these fields, researchers take data storage and access serious for *extrinsic reasons*, such as requirements posed by funding agencies. However, motivation and skills are lacking and financial hurdles exist. The committee calls these fields of science: *the extrinsic range of data storage and access* (or 'stragglers').

In other fields, data originate from unique and expensive observation equipment and are often hardly reproducible. Examples are astronomy, particle physics, the geosciences and, increasingly, the medical sciences. In these fields, institutionally or nationally centralised or federated storage of data is common, and cleaning, accessibility and maintenance of data are crucial. Researchers in these fields

take data storage and access serious for *intrinsic reasons*, such as the use of a common data store or the disclosure of data for reproducibility or complementary studies. However, the costs of organising and maintaining centralised or federated storage, potentially without an explicit exit scenario, can be substantial. Individual researchers cannot foot the bill for keeping data safe and accessible *in the long term*. The committee calls these fields of science: *the intrinsic range of data storage and access* (or 'speedsters').

Both ranges of data storage and access come with challenges. Specific challenges with respect to the extrinsic range are related to the *ownership* of the data (privacy issues, for example), while specific challenges with respect to the intrinsic range have to do with the national infrastructures for centralised or federated data *storage*. Common challenges, furthermore, concern the training of researchers, metadata and data access methods, the need for more (and better) support of researchers working with data, and creating a new culture in which proper data storage and access are recognised and awarded.

During both meetings, the committee noticed an acute sense of urgency to come to better storage and availability of data for research. A need for more *guidance on a national level* was expressed by many participants, leading to the identification of common principles understood and embraced by all. Some even suggested that there should be a national advisory committee for storage and reuse of data under the umbrella of the KNAW.

3. KNOWLEDGE INSTITUTIONS AND FUNDING AGENCIES: A CALL FOR ACTION

3.1 Urgency

It became clear from the expert group meetings on 26 November and 17 December 2020 that the infrastructural facilities and services that knowledge institutions and funding agencies provide are an important *contextual factor* that determines researchers' *comfort level* with regard to data storage and access. The support that researchers expect shows similarities across different fields of science, but varies widely in other ways. During the expert group meetings it further became apparent that proper data storage and access by researchers cannot be widely practiced when knowledge institutions and funding agencies do not embrace the following implications: awareness and responsibility; scientific incentives; training; data steward support; storage and maintenance; and metadata and access.

3.2 Awareness and responsibility

Especially for the fields of science in the extrinsic range, *the awareness of the importance of proper storage of and access to data is low*, both among the individual researchers and at the institutional management level. Often it is far from clear who is responsible for providing and organising proper data storage and access: the individual researcher or the institutional management level. Although funding agencies attempt to raise awareness by requesting a *Data Management Plan* (DMP), these plans are often 'paper tigers', have little impact in practice, are rarely discussed at the institutional management level, and are usually not a part of any institutional quality control process. The efforts of individual researchers that do invest time in drafting and executing a DMP often go unrecognised. Raising awareness, asking the

right questions about data storage and access, clarifying the lines of responsibility, and developing transparent data storage and access protocols: all this requires action by the management of knowledge institutions. In doing so, management will pointedly *raise awareness* among researchers that proper data storage and access must not be ignored but should be part of the normal research process. The Strategy Evaluation Protocol (VSNU *et al.*, 2020) can help raise awareness, because Open Science is one of the specific aspects that assessment committees are required to take into account during research evaluations.

3.3 Scientific incentives

Raising awareness is crucial but will come to nothing if incentives are lacking. Some funding agencies only fund research projects when data are readily made available, while some fields of science thrive only because data are shared. However, for the majority of researchers, proper data storage and access just takes up a lot of their time, with no apparent incentives, appreciation or added value for their academic career. Academic careers are heavily influenced by publications and citations, and often only lip service is paid to those making data available to others. Therefore, researchers might actually benefit from not sharing data freely, or from putting restrictions on sharing data (such as demanding authorship).

Knowledge institutions and funding agencies have a role to play by *incentivising researchers in all fields of science to practice proper data storage and access*. Knowledge institutions must devise ways not only to facilitate proper data storage and access, but also to create a culture in which proper data storage and access are valued, and role models are put in the spotlight, in the same way as is done with excellent researchers and teachers. In their review process of grant applications, funding agencies should assess and take into account researchers' past data sharing records.

3.4 Training

In general, researchers in all fields of science have *insufficient legal or technical training* themselves to ensure proper data storage and access. Nevertheless, in many cases they find that the home knowledge institution does not offer support in the form of advice or training. Knowledge institutions therefore need to *step up their effort to provide training* for researchers working with data.

3.5 Data steward support

Because researchers in all fields of science may fear that their activities with respect to storing data for access by others could jeopardise their main responsibilities as a researcher, knowledge institutions should *step up their efforts to provide support* (and possibly the level of support) to researchers working with data. Researchers must be supported by *data stewards*, in much the same way as they are supported by hardware specialists or laboratory technicians. These data stewards must be flexible in the type and level of support they provide, as this will depend on the level of expertise of the researchers they are supporting, and on the customs of the scientific discipline concerned. The specific task of data stewards is to support researchers in making data more accessible and reusable (Jetten *et al*., 2021). On top of having expertise in the field of data management, data stewards must also have a thorough knowledge of privacy issues and copyright law (KNAW, 2019). For some fields of science data storage and access might be outsourced to the data stewards, whereas for other fields technical support for accessing massive volumes of data via software interfaces might be in order. Additionally to the training for researchers working with data as mentioned above, *training programmes* must be made available to the researchers so as to ensure that they fully leverage the support that the data stewards can provide.

Attracting data stewards will not be easy for academic organisations, as these specialists are in *high demand* in virtually any sector of the economy and in society at large. Companies offer attractive packages to data stewards in support of their data researchers. For that reason, attracting data stewards on temporary contracts – as currently often occurs with research project funding – will not suffice. Knowledge institutions should therefore develop clear and attractive career paths for data stewards and offer permanent contracts that are not dependent on research project funding.

3.6 Storage and maintenance

For the fields of science in the *intrinsic range*, knowledge institutions and national funding agencies must play a sustained role in realising *national infrastructures for centralised or federated data storage.* In these fields, infrastructures are typically well organised, so no immediate action is needed.

At the same time, for the fields of science in the *extrinsic range*, data are volatile and therefore reproducibility of these data can be problematic. These fields could benefit from coordinated storage and access. Individually or collaboratively, knowledge institutions should act to *challenge* the different fields to answer the question *how much* they would benefit from coordinated data storage and access.

In order to prevent academic research from becoming part of a data landscape in which storage of and access to data is controlled by *commercial players*, the committee feels that:

- Data collected by others (i.e., secondary data) should be stored properly for reproducibility studies. In general, researchers manage to come to a good understanding with data donors on the terms for storage and accessibility. Knowledge institutions, however, must be aware that such negotiations with data donors are conducted by researchers whose main talents lie in performing research. Insofar as researchers indicate to the home knowledge institution that they are in need of legal and ethical support in negotiating with data donors (especially with large and powerful companies and government organisations), the institutions should provide them with the necessary type of expertise without delay. This can lead to improved storage of data, and to enhanced data sharing as well.
- With respect to the storage of published data, the knowledge institutions should take a closer look at the role that academic publishers can play in the data landscape. On the one hand, academic publishers provide researchers with facilities and support for data storage and accessibility, which makes data accessible for reproducibility studies. On the other hand, by putting their money where their mouth is, they have become an important player in the data landscape. The knowledge institutions should weigh the pros and cons of this situation from the viewpoint of researchers working with data. When the pros outweigh the cons, knowledge institutions and academic publishers should join forces.

3.7 Metadata and access

Under the general principles of FAIR data, there is already a wide variety of definitions for *metadata*, and this diversity will only increase. The implication for knowledge institutions is that they must brace themselves for a diversity of data storage and access methods, especially where metadata are concerned, and that institutional attempts to harmonise metadata because of cost considerations will be opposed by researchers. However, if their perspective is to help researchers store data and make data available for research, then *clear user guidelines* for metadata are needed.

4. RECOMMENDATIONS

Based on the expert group meetings, the interviews and the available literature, the committee concludes that some parts of the Dutch data landscape are vibrant, while other parts are complex and fragmented. Implementation of ideas, suggestions and plans is slow. At the same time, knowledge institutions and funding agencies in the Netherlands are working hard to bridge gaps. They collaborate with other Dutch organisations of higher education and research, including the KNAW, towards realising Open Science and Open Data. In view of the latter, the committee does not follow the suggestion made during the export group meetings that there should be a national advisory committee for storage and reuse of data under the umbrella of the KNAW.

The committee comes to the following recommendations, addressed at the management of knowledge institutions (policymakers and decisionmakers), national funding agencies (NWO and ZonMW), and the ministers of Education, Culture and Science, of Economic Affairs and Climate Policy, and of Health, Welfare and Sport. *All recommendations put the researcher centre stage.* But researchers have a responsibility as well: before starting their project they must make it clear to their management which support and guidance they need for proper data storage and access. The checklist included in Annex 2 can help them become aware of the various steps that need to be made along the path of data generation and analysis.

Recommendations to the management of knowledge institutions

PRIMARY

Provide support

Challenge the different fields of science to make clear to what extent the support that is offered to researchers needs to be improved. Support researchers by deploying

data stewards in a facilitating role, by streamlining data acquisition and storage, by removing legal and ethical hurdles, by simplifying rules and by reducing overhead, etc. Make training programmes available to researchers to ensure that they fully leverage the support that the data stewards can provide. Develop clear and attractive career paths for data stewards, including training facilities, and offer permanent contracts that are not dependent on research project funding. Create ICT-like services to enable researchers to easily and securely deposit, share, publish and preserve research data during all stages of the research project.

Identify the costs of data storage and support

Data storage and support costs are underestimated and rising. Challenge the different fields of science to make clear how much local and national data storage and support is needed. Investigate to what extent the available data infrastructures are sufficient. Investigate the feasibility and the costs of new data storage and support. Once researchers are convinced of the benefits of coordinated data storage and access, the costs have to be borne by the institutions in the first and second flow of funds, not by the researchers. Due to the wide variety in types of data that are needed for research performed by individual researchers and research groups at Dutch knowledge institutions, several national data storage centres may be needed. Explore in what way these centres can cooperate with each other and/or with international storage centres via exchange agreements, joint protocols and facilities for distributed analysis. Investigate how existing public facilities can be integrated. Identify the computational challenges surrounding the analysis of remote data.

SECONDARY

Raise awareness

Data management is in many respects a 'people problem' as well as a 'technical problem'. Researchers want basic assistance with the day-to-day issues they are faced with when creating and managing the data. Addressing these issues need not be overly expensive or depend on the implementation of highly technical solutions (Ward *et al.*, 2010). Much is already available. Challenge the different fields of science to define how much they would benefit from coordinated data storage and access. Clarify the lines of responsibility and the different roles played by everyone involved, and develop transparent data storage and access protocols. Give guarantees to researchers that data will remain available even after research projects have been terminated. These actions will raise awareness among researchers about the need, the advantages and the non-optional nature of proper data storage and access.[12] In doing so, build the expertise to deal with the highly diverse questions that individual researchers will ask in response, ranging from the legal and ethical aspects of data storage to technical support and who picks up the bill.

12 Organisations such as SURF or DANS might be instrumental in taking these actions.

Provide expertise to overcome legal and ethical hurdles

Legal and ethical hurdles may prevent proper generation and storage of data. There is a need for simple procedures that are understood and embraced by everyone involved. In case of personal data, identify what flexibilities the General Data Protection Regulation (GDPR) offers to the generation and storage of the data that are needed for research performed by knowledge institutions in the Netherlands, and advise knowledge institutions on responsible data practices, for example in the form of a Code of Responsible Data Use in Research. Knowledge institutions should provide the necessary type of expertise, and without delay, to researchers in need of legal and ethical support when negotiating with data donors (especially with large and powerful companies and government organisations).

In doing so, knowledge institutions will build up the expertise to help researchers find proper ways of storing and sharing data within the limits of contract or law, and ultimately will have sufficient seniority and power to persuade researchers to renegotiate the legal and ethical conditions with the data donors. The ethical review boards of the knowledge institutions should include one or more experts in data management. At the same time, a culture of 'audit and compliance' must be avoided.

Provide training

For some fields of science, the level of knowledge should be such that the researchers can work with data stewards and legal experts, whereas for other fields the researchers need to have hands-on expertise themselves in storing and retrieving data. Offer training in data management and data storage in all fields of science. It is important that training is not a 'one size fits all'. National cooperation of knowledge institutions is helpful, for example via national research schools for PhD students.[13]

Value the additional workload

As an employer, recognise and reward the work of the academic staff involved, also in the light of international initiatives such as DORA and the Hong Kong Principles, and national initiatives such as the position paper *Room for everyone's talent: towards a new balance in the recognition and rewards for academics* (VSNU *et al*., 2019). Differentiate between individual versus team in the needs and the responsibilities for data generation and data storage for research (see also PNN, 2020). Create a culture in which proper data storage and access is valued and in which role models are put in the spotlight. The quality, execution and impact of Data Management Plans (DMPs) can be part of the annual performance evaluation of members of staff. Institutional assessments can include an assessment of the quality, relevance and implementation of data storage and access.

13 Organisations such as the Netherlands eScience Center and the local Digital Competence Centres (DCCs) can play an important role in developing training and curriculum.

Monitor security
Make sure that available software is safe enough for data protection.

Weigh the pros and cons of academic publishers storing published data
Take a closer look at the role that academic publishers can play in the data landscape. Weigh the pros and cons of their involvement from the viewpoint of researchers working with data. When the pros outweigh the cons, knowledge institutions and academic publishers can join forces.

Recommendations to the management of knowledge institutions and to national funding agencies

Realise centralised or federated data storage
Play a sustained role in realising national infrastructures for centralised or federated data storage for those fields of science where data originate from unique and expensive observation equipment and are often hardly reproducible, and where researchers take data storage and access serious for *intrinsic reasons*. The key challenge is to make data compatible and interoperable.

Incentivise researchers to practice proper data storage and access
Actively reward researchers for making data available, for example when considering promotions and grant awards. This can lead to improved data storage and sharing.

Recommendations to national funding agencies

Evaluate the realisation of past data storage proposals
Evaluate a sample of research proposals funded in the past in which substantial funding for data storage was involved. Has the proposed data storage been established?

Recommendations to the ministers of Education, Culture and Science, of Economic Affairs and Climate Policy, and of Health, Welfare and Sport

Set requirements for sharing of data
Set equal and mutual requirements for sharing of data between academia on the one hand and Dutch companies and government organisations on the other. Some data donors put limits on how researchers can use, share, publish, verify and reproduce data. Create a fair and efficient playing field. This could be done by means of a covenant agreed between academia, Dutch companies and government organisations. The ministries should initiate discussions between these parties and control the course of the discussions.

Offer financial support

It is clear that considerable costs are involved in reaching the ambitious national goals with respect to FAIR data. The present budgets in the first and second flow are not sufficient to reach these goals. Consult the management of the knowledge institutions and the national funding agencies on the matter of their financial needs with respect to data storing and making data available for research. The costs of data storing and making data available for research are considerable, and it should be made clear to the researchers who picks up the bill. The costs of data storage and access could be split between the ministries, the knowledge institutions and the funding agencies in the Netherlands. 'Not acting' is not an option, because there are substantial costs involved in data loss, for example if tests and surveys must be carried out again. Data have intrinsic value as well, for example for reproducibility or complementary studies.

RELEVANT LITERATURE

AWTI (2016). Dare to share. Open access and data sharing in science

Ayris, P. (2017). Challenges and opportunities for RDM in the arts, humanities and social sciences. A practitioners' viewpoint

Borghi, J., Abrams, S., Lowenberg, D., Simms, S., Chodacki, J. (2018). Support Your Data: A Research Data Management Guide for Researchers. Research Ideas and Outcomes 4

EASA/European Association of Social Anthropologists (unknown year). Statement on data governance in ethnographic projects: pdf

Elsevier CWTS (2017). Open data: The research perspective

Ember, C., Hanisch, R., Alter, G., Berman, H., Hedstrom, M. & Vardigan, M. (2013). Sustaining Domain Repositories for Digital Data: A White Paper

European Commission (2018). Facts and figures for open research data

European Commission, Directorate-General for Research and Innovation (2016). Prompting an EOSC in practice. Final report and recommendations of the Commission 2nd High level expert group on the European Open Science Cloud (EOSC)

European Commission, Directorate-General for Research and Innovation (2018). Turning FAIR into reality, First report and recommendations of the Commission high level expert group on the European Open Science Cloud

European Commission, Directorate-General for Research and Innovation (2018). Cost of not having FAIR research data in the EU. Written by PwC EU Services

European Commission, Directorate-General for Research and Innovation (2018). Realising the European Open Science Cloud. Final report and action plan from the European Commission expert group on FAIR data

European Commission, Directorate-General for Research and Innovation (2020). Six recommendations for implementation of FAIR practice

European Research Council (2018). Open research data and data management plans

Gabella, C., Durinx, C. & Appel, R. (2018). Funding knowledgebases: Towards a sustainable funding model for the UniProt use case [version 2; peer review: 3 approved]. F1000Research 2018, 6(ELIXIR):2051

Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M. & Van Gelder, C.W.G. (2021). Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship. Zenodo

KNAW (2012). Responsible research data management and the prevention of scientific misconduct
KNAW (2016). Opening the book on open access. Interviewbundel
KNAW (2018). Replication studies
KNAW (2018). Big data in scientific research with personal data
KNAW (2019). Reuse of public data. More academic research and better government policy
Landelijk decanenoverleg faculteiten Maatschappij- en Gedragswetenschappen (2018). Richtlijn archivering wetenschappelijk onderzoek
LEARN (2017). 20 RDM Best-Practice Recommendations
Minister of Education, Culture and Science (2019). Nieuwsgierig en betrokken. De waarde van wetenschap
Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature* | Vol 578 | 27 February 2020
NPOS (2020). Final report. Exploring and optimising the Dutch data landscape
NWO (2017). Topwetenschap vereist topinfrastructuur. Adviesrapport nationale digitale infrastructuur. Permanente Commissie voor Grootschalige Wetenschappelijke Infrastructuur. ICTsubcommissie: https://www.nwo.nl/nieuws/nwo-pleit-voor-structurele-financiering-van-nationale-digitale-infrastructuur
O'Doherty, K.C., Mahsa, S., Dove, E.S., Bentzen, H.B., Borry, P., Burgess, M.M., Chalmers, D., De Vries, J., Eckstein, L., Fullerton, S.M., Juengst, E., Kato, K., Kaye, J., Knoppers, B.M., Koenig, B.A., Manson, S.M., McGrail, K.M., McGuire, A.L., Meslin, E.M., Nicol, D., Prainsack, B., Terry, S.F., Thorogood, A. & Burke, W. (2021). Toward better governance of human genomic data. *Nature Genetics* | VOL 53 | January 2021 | 2–8 |
OECD (2017). Business models for sustainable research data repositories. *OECD Science, Technology and Industry Policy Papers*, No. 47, Paris: OECD Publishing
Pels, P., Boog, I., Florusbosch, H.J., Kripe, Z., Minter, T., Postma, M., Sleeboom-Faulkner, M., Simpson, B., Dilger, H., Schönhuth, M., Poser, A., Castillo, R.C., Lederman, R., Richards-Rissetto, H. (2018). Data management in anthropology. The next phase in ethics governance?
PNN (2020). PhD Survey: Asking the relevant questions. PhD criteria, Open Science, Recognition and Rewards, Career
Powell, K. (2021). How a field built on data sharing became a tower of Babel. *Nature* | Vol 590 | 11 February 2021
RDA (2014). The Data Harvest Report – sharing data for knowledge, jobs and growth
Science Europe (2021). Practical Guide to the International Alignment of Research Data Management
SIEF / Société Internationale d'Ethnologie et de Folklore (year unknown). Statement on data management in ethnology and folklore
The British Academy & The Royal Society (June 2017). Data management and use: Governance in the 21st century
Von der Heyde, M. (April 2019). Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future. Zenedo
VSNU, KNAW, NWO, TO2 and VH (2018). Netherlands code of conduct for research integrity
VSNU, NFU, KNAW, NWO and ZonMW (November 2019). Room for everyone's talent: towards a new balance in the recognition and rewards for academics
VSNU, NWO and KNAW (2020). Strategy Evaluation Protocol 2021-2027

Ward, C., Freiman, L., Molloy, L., Jones, S., Snow, K. (2010). Making sense: talking data management with researchers. *International Journal of Digital Curation*, 6 (2). ISSN 1746-8256

Wilms, K.L., Stieglitz, S., Buchholz, A., Vogl, R., Rudolph, D. (2018). Do researchers dream of research data management? Proceedings of the 51st Hawaii International Conference on System Sciences | 2018

Woeber, Catherine Anne (2017, master thesis). Towards best practice in RDM in the humanities

LIST OF ABBREVIATIONS

ALLEA	ALL European Academies
CBS	Statistics Netherlands (in Dutch: Centraal Bureau voor de Statistiek)
DANS	Data Archiving Networked Services. Dutch National Centre of Expertise and Repository for Research Data
DCC	Digital Competence Centre
DPA	Dutch Data Protection Authority (in Dutch: Autoriteit Persoonsgegevens)
DMP	Data Management Plan
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
Health-RI	Health Research Infrastructure
LCRDM	National Coordination Point Research Data Management (in Dutch: Landelijk Coördinatiepunt Research Data Management)
NPOS	National Platform Open Science
NWO	The Dutch Research Council (in Dutch: Nederlandse Organisatie voor Wetenschappelijk Onderzoek)
ODISSEI	Open Data Infrastructure for Social Science and Economic Innovations
RDA	Research Data Alliance
RI	Research Infrastructure
SURF	Organisation for ICT-cooperation of educational and research institutions in the Netherlands
VSNU	The Association of Universities in the Netherlands (in Dutch: Vereniging van Universiteiten)
ZonMW	The Netherlands Organisation for Health Research and Development (in Dutch: Nederlandse Organisatie voor Gezondheidsonderzoek en Zorginnovatie)

REVIEW

At the request of the Academy's Board, a draft of this report was reviewed by the following reviewers:

- Prof. Hans van Duijn, Chair of Permanent Committee Large-scale Scientific Research Infrastructures, The Dutch Research Council
- Prof. Nathali Helberger, University professor in Law and Digital Technology, with a special emphasis on Artificial Intelligence, University of Amsterdam
- Prof. Maarten Krol, Associate professor in Air Quality and Atmospheric Chemistry, Utrecht University; Netherlands Institute for Space Research *SRON*
- Prof. Karin Verweij, Extraordinary professor Genetics in Psychiatry, Amsterdam UMC
- Henk Wals, Director DANS

In addition, the report was reviewed by:

- The Academy's Council for the Humanities
- The Academy's Council for Medical Sciences
- The Academy's Council for Natural Sciences and Engineering
- The Academy's Social Sciences Council

The reviewers are not responsible for the final report.

ANNEXES

Annex 1. Resolution establishing a 'Storage and Availability of Data for Research Committee'

Remark: small edits have been made in the original resolution. The edits are added in footnotes.

The Academy Board has resolved to establish a 'Storage and Availability of Data for Research Committee' (referred to below as 'the Committee'), doing so on the basis of the following considerations:

- Across the full breadth of science and scholarship, there is a trend towards making research data more readily available for reuse in accordance with the FAIR principles (*Findable, Accessible, Interoperable, Reusable*).
- Researchers – at various organisational levels and in the light of various funding conditions – find themselves facing a range of policy measures and facilities intended to encourage and support the storage and availability of data for research. Many have concluded that such initiatives often result in additional costs and/or extra work for researchers without it being clear what the benefits will be and for whom they are intended. This makes it especially urgent for the Academy to issue advice on this matter.
- On 4 December 2018, the Academy organised a meeting of experts, which concluded that the Academy should issue advice in order to bring about greater clarity and structure in discussions regarding improving the storage and availability of data for research. (A summary of the matters dealt with at the meeting is attached as an appendix.)
- The Academy wishes to ensure that the opinions of researchers are heard within the process of implementing the National Open Science Plan (NOSP), which the Academy cosigned in early 2017.

- The Academy aims to ensure that the Netherlands plays a leading role in international initiatives towards making data more accessible for reuse and for the country to contribute actively to such initiatives – for example the European Open Science Cloud (EOSC) – designed to support and encourage similar development. The perspective of the researcher needs to be central in this regard.

Section 1. Task of the Committee

The Committee will carry out its work from the perspective of the researcher. The advice provided is intended to bring about greater clarity and structure in discussions regarding improving the storage and availability of data for research, with a view to arriving at a coordinated national approach and for the Netherlands to play a leading role internationally.

The tasks of the Committee are twofold:

1. to carry out a wide-ranging exploration of the possibilities for making data more accessible for research and for storing it in a sustainable manner;
2. to prepare an Academy advisory report with specific recommendations for making those possibilities feasible in various fields of science and scholarship.

1.1. Exploratory study

The Committee's first task will be to carry out a wide-ranging exploratory study, guided by the following questions:

1. Why should researchers store research data and make it available? Why should they *not* do so? Are there differences between the various fields of science and scholarship and between research methods? Are there any cross-discipline considerations that are common to the various fields of science and scholarship?
2. What research data should researchers retain or not retain, for how long and in what form should they do so, and who should decide this?
3. What research data should be available for use by researchers, and who should decide this?
4. What research data should be available for use by non-academic parties, and who should decide this? To what extent should private parties and government bodies benefit from research data that researchers share with them, and what conditions and responsibilities should apply?

The exploratory study will bring greater clarity and structure to initiatives – across the full breadth of science and scholarship – towards making research data more readily available for research and for storing it sustainably.

1.2. Advisory report

The Committee's second task will be to prepare an advisory report to be issued by

the Academy. In drawing up the report, the Committee will consider what similarities and differences there are between the various fields of science and scholarship, and what kind of generic national and/or international data infrastructure, support, and guidelines researchers need. Relevant aspects in this regard include ownership, division of tasks, funding, researcher training, legal requirements, the role of publishers, and appreciation for making research data reusable for research.

1.3. Procedure

The Committee's exploratory report will be based in part on one or more familiarisation meetings, which will form the basis for surveying the wishes expressed within a number of scientific and scholarly fields regarding the storage and availability of data for research. A meeting can also be held with a view to identifying national and international data initiatives undertaken by scientists and scholars and industry.

The Committee's advisory report will also be based on one or more meetings of experts that will be organised. The Committee will submit the draft report to the Academy Board before 1 July 2020.

The advisory report will be drawn up in English.

Section 2. Composition of Committee and Appointment Period

The following individuals will be appointed to the Committee in their personal capacity:

Chair:

- Prof. Frank den Hollander, Professor of Probability, Leiden University

Other members:

- Prof. Dorret Boomsma, Professor of Biological Psychology, VU University Amsterdam
- Prof. Antal van den Bosch, Professor of Language and Artificial Intelligence, University of Amsterdam, and director of the Meertens Institute
- Prof. Inald Lagendijk, Professor in Computer Science, Delft University of Technology
- Prof. Ariana Need, Professor in Sociology and Public Policy, University of Twente
- Prof. Julia Noordegraaf, Professor of Digital Heritage, University of Amsterdam[14]
- Prof. Frits Rosendaal, professor in Clinical Epidemiology, Leiden University Medical Centre
- Morris Swertz, Head of Genomics Coordination Centre, University Medical Centre Groningen

14 In April 2020 prof. Julia Noordegraaf stepped down as a member of the Committee.

- Prof. Jeannot Trampert, Professor of Geophysics, Utrecht University

The Committee's term will run until mid-2020.
Prof. Maarten Prak will serve as portfolio manager on behalf of the Academy Board[15].
The Committee will be assisted by Ans Vollering (Academy Bureau)[16].

Section 3. Integrity and Quality

Prior to the first meeting of the Committee, the members took note of the *Code to Prevent Improper Influence due to Conflicting Interests* [*Code ter voorkoming van oneigenlijke beïnvloeding door belangenverstrengeling*]; they confirmed having done so in a written statement. The Committee members familiarised themselves with the 'Manual Concerning Academy Advisory Reports' [*Handleiding adviezen KNAW*], adopted by the Academy Board on 18 September 2017. The policy set out in that Manual will be followed when assessing the draft advisory report.

Section 4. Work Plan

The Committee will draw up a work plan specifying its working methods and its communication and implementation strategy.

Section 5. Travel Allowance

The Academy will reimburse the Committee members for their travel expenses but will not make any other payment to them.

Section 6. Confidentiality

The Committee members will treat as confidential all information to which they become privy while implementing this resolution and which can be assumed to be confidential.

Adopted in Amsterdam on 26 October 2019 by the Board of the Royal Netherlands Academy of Arts and Sciences.

On behalf of the Board of the Royal Netherlands Academy of Arts and Sciences,

Mieke Zaanen
Director General of the Royal Netherlands Academy of Arts and Sciences

15 In September 2020 prof. Maarten Prak was succeeded as portfolio manager by prof. Sjaak Neefjes.

16 On 1 January 2021 the support allocated to the Committee was expanded by Hanneke van Doorn (Academy Bureau).

Appendix to resolution establishing a 'storage and availability of data for research committee': summary of proceedings at the meeting of experts on 4 December 2018

Drawn up by the Preparatory Committee for the meeting of experts on 'Storage and Availability of Data for Research', comprising Frank den Hollander (chair) and members Kees Aarts, Anna Akhmanova, Antal van den Bosch and Frits Rosendaal. The Committee was assisted by Hanneke van Doorn and Ans Vollering.

Purpose of the meeting of experts

During the meeting of experts on 4 December 2018 the Preparatory Committee, acting at the request of the Academy Board, explored what the focus of an advisory report by the Academy should be in order to provide the maximum added value within the process of implementing the National Platform Open Science (NPOS), which the Academy co-signed in early 2017.

The subject of the storage and availability of research data raises a wide range of questions, regarding both practical matters and principles. During the meeting of experts organised by the Academy, the Preparatory Committee focussed, from the perspective of the researcher, on the questions regarding principles, an approach that is an appropriate one for the Academy.

The purpose of the meeting was to determine whether the Academy should advise on this matter, and if so, what problem(s) or question(s) should be central to the remit of the Academy committee that will prepare the advisory report. The purpose of the meeting of experts was not, therefore, to already begin answering questions but to formulate the main questions.

Approach at the meeting of experts

Prior to the meeting of experts, participants had received a discussion paper with some preliminary key questions that should be addressed in an advisory report. The chairman, Frank den Hollander, opened the meeting with a brief introduction. There were then three short pitches, by Dorret Boomsma (VU University Amsterdam), by Femius Koenderink (AMOLF: Physics of functional complex matter), and by Karel Luyben (National Platform Open Science). After an explanation by Stan Gielen of the current data policy pursued by the national funding agency NWO, the participants split into three breakout groups to discuss the preliminary key questions. The groups then reported back on their findings, followed by a concluding plenary discussion of the added value of an advisory report to be issued by the Academy and the possible statement of the problem or question(s) for the Academy committee that will prepare that report.

Points to consider arising from the concluding plenary discussion

At the end of the meeting, the chairman, Frank den Hollander, concluded that there was a great deal of support for an Academy advisory report on the storage and availability of data for research. It was suggested that the following points be considered when drawing up that report:

1. The perspective of the researcher should be central to the advisory report, with attention being paid to both practical matters and matters of principle.
2. The challenge is to do justice to the differences between the various fields of science and scholarship while at the same time finding viable common denominators.
3. What data should be retained and what data should not? And if they should be retained, for how long? Who will decide that? Who owns the data? Who is to share data and who isn't? In what form? Regenerating data can sometimes be a meaningful solution.
4. What is the value of sharing data? What are the pros and cons? 'Sharing data' is not the same thing as 'sharing data effectively'. Do data possess objective reality or are they constantly changing?
5. Is a central data archiving facility a viable option? Who will have final responsibility? What are the financial aspects of this and what are the long-term obligations? The role of commercial publishers needs to be looked at critically.
6. What are the international aspects and responsibilities regarding data storage and availability that researchers have to deal with? How will all these matters be dealt with as regards consultation with other countries?
7. From the perspective of the researcher, consideration also needs to be given to such aspects as management, law, sustainability, privacy, training, and international embedding. An advisory report should not only look forward but also backward, to the data that already exists.
8. The role of 'data stewards' needs to be repositioned. They should serve the researcher and not just the institution. Here, too, national arrangements must be made. This is especially difficult where multidisciplinary data is concerned.
9. The Academy's role in this process is seen as being positive. The Academy is ideally suited for placing the focus on the position of science and scholarship and the researcher. This is very welcome within the major policy discussions that are currently taking place regarding data. The advisory report can help structure the debate about data. Cooperation with NWO and NPOS can ensure that the report is embedded within the rapid developments regarding data.
10. The advisory report could also lead to the compilation of an 'instruction manual' to assist researchers in dealing with data from a broad perspective. This can be done with full attention being paid to the specific needs of researchers' own particular fields.

Annex 2. Primary tasks and checklist of questions for researchers working with data

This Annex aims to help researchers navigate through the maze of challenges, opportunities, restrictions and obstacles, and to help them secure the necessary support from the home knowledge institution.

I. Primary tasks

The committee listed a number of *primary tasks* that each researcher working with data is confronted with in each research project:

Data

- Decide what is proper collection or generation of data.
- Identify who owns the data and who has control.
- Address issues of intellectual property.
- Address privacy issues, including pseudonymization and anonymization. When in doubt, ask the Dutch Data Protection Authority (DPA), which supervises processing of personal data in order to ensure compliance with laws that regulate the use of personal data.

Storage

- Decide what is needed at a minimum and at best achievable regarding proper storage.
- Identify who is responsible for the storage, what the costs of storage are, and who picks up the bill.
- Determine whether or not the data can potentially be modified or altered by users.
- Secure safe storage, possibly via back-up.

Availability

- Clarify what the advantage or disadvantage is of sharing data.
- Data can be shared always, sometimes or never. Identify possible reasons for not making data available during a certain window of time.
- Distinguish between data that relate to people and data that do not.

Based on the above tasks the committee compiled a checklist of questions for researchers working with data, in order to help them develop a Data Management Plan (DMP) before the start of the research project and in the final stage of the research project to properly store data and make data available for reuse by others. The checklist can help researchers find practical solutions for problems with data storage and making data available for research. These solutions usually depend on the research project involved.

II. Checklist of questions

By following the checklist below, the researcher becomes *aware of the importance of proper storage of and access to data,* and of the steps to be made along the path of data generation and analysis. In some of these steps, the researcher needs support from his/her employer. Making clear what support is needed and how it is realised can be a major challenge. It must also be clear in advance of the research project who is responsible for providing and organising proper data storage and access.

In the checklist below, data are taken to be *digital data*.

FRAME

Data come in different categories:

- Data that are indispensable for other researchers.
- Data that are useful for other researchers, but not indispensable.
- Data that are potentially useful in the future.
- Data that do not need to be stored and for which it suffices to describe how the data were generated.
- Data for which reproducibility or complementary studies are the main reason for storage.

Important aspects of storage and availability of data for research are:

- Data are useless without metadata. Metadata are necessary to make data findable and (re)usable for other researchers. Metadata contain information about how the data are generated, organised, processed and stored, including software and analysis scripts.
- Data may be eliminated in the course of the research process, or thereafter. In case of personal data, this may happen at the request of the data donor and on the basis of his/her legal right to be forgotten.

QUESTIONS

The following questions need to be addressed:

(A) Which data are allowed to be generated, which data are allowed to be accessed, and by whom?
Sub-questions are:

1. Both the generation of data and the access to data may be restricted. What restrictions apply at the home institution and nationally (and possibly internationally)?
2. How do users authenticate themselves before they are granted access?
3. Are there any legal or ethical hurdles to be taken?
4. What agreements need to be made beforehand and with whom? Are there contradictory requirements to be bridged?

(B) Which data and support will be needed for a proper execution of the research? Sub-questions are:

1. Of the necessary data, which will be primary data collection and which secondary data collection? Data can be collected either by the researcher (primary data collection) or by others (secondary data collection). In case of personal data, the collection of primary data should comply with the General Data Protection Regulation (GDPR). Does the collection of secondary data comply with the conditions of the data source, including possible financial compensation and possible sharing of responsibility for the data?
2. Are the data made FAIR? How is this measured? What are the standards?
3. Does the home knowledge institution offer enough facilities and support to acquire the necessary data? If no, then where can this be secured? If yes, then what are the local rules for data storage, the local support points and the local support tools? What are the protocols? Who is responsible for what? Are training and workshops being offered? How do local initiatives fit with national plans? Who are the local legal experts, ethical experts, data stewards and software engineers?
4. Which preparations need to be made prior to the collection of primary data? What needs to be put in a consent form?
5. Who needs to be contacted prior to the collection of secondary data?

(C) Which data will be stored and kept for future access, and which will not? Sub-questions are:

1. If data are stored and kept for future access, then why is this so? What are the costs and the benefits involved? What if the data volume is so large that storage costs are prohibitive? How long will the data be stored and why? Who is the owner of the stored data? What are the roles of the researcher, the home knowledge institution, the granting agency and, in case of personal data, the data donors? Is there sufficient agreement on ownership and roles? What are the standards for storing and keeping data for future access? Is there coordination? Is there transparency? Will the data be stored and kept for future access with a *Digital Object Identifier* (DOI) and with metadata? Is there an Open Science community with whom experiences can be shared and discussed? Are there national or international data infrastructures to make use of? Are there colleagues with relevant expertise who can help, for example legal experts, ethical experts, data stewards and software engineers?
2. If the data are not stored and kept for future access, then why is this so? In case of personal data, are there privacy issues? Is there a lack of storage space? Are the costs involved too high? Are the data difficult to classify?

(D) Will data be made accessible for reuse by other researchers?
Sub-questions:
If the answer is 'no', then:

1. What are the motives for not making data accessible for reuse by other researchers? For instance, can data not be made accessible because of privacy issues, ethical matters or restrictive contracts? Can these obstacles be removed? What to do when circumstances change along the way?
2. Will data be made accessible later? Perhaps data can be made accessible after some pre-specified delay or embargo.

If the answer is 'yes', then:

1. What are the motives for making data accessible? Possible motives are:
   - Funding agencies (such as the home knowledge institution, NWO, ERC) require that data are stored.
   - There are incentives for the researcher to make data accessible, such as long-term financial support.
   - Other researchers are more willing to collaborate after data have been shared. Collaboration can improve data.
   - Other researchers are more willing to quote research publications or refer to data that are made accessible.
   - Data are indispensable for other researchers.
   - Data have potential value in terms of reuse (quality, originality, innovation, size, scale, cost), uniqueness (non-repeatable observations), or history (in particular, history of science).
   - While not indispensable for other researchers, data are indispensable for complementary studies.
   - Data are essential or useful for non-scientific purposes, for example, cultural heritage or museums.
   - The Netherlands code of conduct for research integrity (VSNU *et al*., 2018) states that data must be stored and made accessible for reuse, so that in the future research can be verified and reproduced. From the viewpoint of science integrity, others should be able to verify the research findings, and to check the data to avoid fraud. Not only raw and published data must be stored properly, but processed data as well.
   - Academic publishers require that the published data are made accessible for reproducibility studies. They provide researchers with facilities and support for storage and accessibility of the data.
2. Which data will be made accessible? Possible forms are:
   - Make all or part of the data accessible. Processing of personal data must comply with the GDPR.
   - Data include suitable metadata necessary for analysis.
   - Software and analysis scripts are also data.
   - In case of controlled access to data, the researcher adds metadata to search catalogues.

3. To whom will the data be made accessible? Possible target groups are:
   - All researchers.
   - Specific groups of researchers.
4. At what moment will the data be made accessible? Possible moments are:
   - Immediately upon completion of the research project.
   - At some later time that has been agreed on beforehand.
5. In what form and under what conditions will the data be made accessible? Possible forms are:
   - Decide on file formats and code systems to represent the data in such a way that other researchers can easily understand and process the data.
   - Decide on the conditions for making data accessible, including financial compensation for relevant costs. These can include the costs of the provision of data, data storage, and data transferral. The potential re-user shares responsibility for the data.
6. What is the best form to transfer the data? Possible forms are:
   - Transfer data either directly, via a data journal, or via a local, national or international repository or archive.
   - In case of personal data, data analysis (for example, meta-analysis) may be carried out within the walls of the home institution of the owner of the data (so that the data do not need to go out), either automatically or manually.

Annex 3. Individuals consulted

Kees Aarts, University of Groningen
Albert-Jan Boonstra, ASTRON
Susan Branje, Utrecht University
Marcel Broersma, University of Groningen
Rense Corten, Utrecht University
André Dekker, Maastricht UMC+ and MAASTRO
Gijs van Dijck, Maastricht University
Alastair Dunning, Delft University of Technology
Pearl Dykstra, Erasmus University Rotterdam
Joris van Eijnatten, Utrecht University and Netherlands eScience Center
Katrien Helmerhorst, Erasmus University Rotterdam
Roy Hessels, Utrecht University
Ignace Hooge, Utrecht University
Bart Jacobs, Radboud University
Maurice van Keulen, University of Twente
Martijn Kleppe, KB/National Library of the Netherlands
Willem Jan Knibbe, Wageningen University & Research
Vianney Koelman, Eindhoven University of Technology
Jan Kok, Radboud University
Ruben Kok, Dutch Techcentre for Lifesciences (DTL)
Frans de Liagre Böhl, Utrecht University
Willem Lijfering, LUMC
Detlef Lohse, University of Twente
Karel Luyben, European Open Science Cloud (EOSC)
Enrico Mastrobattista, Utrecht University
Barend Mons, LUMC
Renée de Mutsert, LUMC
André Niemeijer, Utrecht University
Dong Nguyen, Utrecht University
Karin van der Pal-de Bruin, LUMC
Brenda Penninx, Amsterdam UMC
Corné Pieterse, Utrecht University
Oliver Plümper, Utrecht University
Simon Portegies Zwart, Leiden University
Mirjam van Praag, VU Amsterdam
Menno Rasch, Utrecht University
Marcel Reinders, Delft University of Technology
Jeroen de Ridder, UMC Utrecht
Mirko Schäfer, Utrecht University
Stefan Schouten, Utrecht University and Royal Netherlands Institute for Sea Research (NIOZ)
Eline Slagboom, LUMC
Marco Spruit, Utrecht University
Marc Steen, TNO innovation for life
Ingmar Swart, Utrecht University
Federico Toschi, Eindhoven University of Technology
Damian Trilling, University of Amsterdam
Daniël Vanmaekelbergh, Utrecht University
Ivan Vasconcelos, Utrecht University
Peter Visscher, University of Queensland, Australië
Arjen de Vries, Radboud University
Henk Wals, Dutch National Centre of Expertise and Repository for Research Data (DANS)
Lieke Welling, LUMC
Mariette Wolthers, Utrecht University
Raúl Zurita Milla, University of Twente

Annex 4. Some selected reports, a summary

This Annex contains a summary of five reports addressing various needs, hurdles and opportunities related to the storage and availability of data for research. Part of their content has been re-voiced by the experts that were consulted by the committee and has found its way into the various sections of the present advisory report.

Knowledge institutions and funding agencies in the Netherlands are working hard to realise improvements in the data landscape. The committee takes note of the slow pace of implementation. With the present advisory report the committee aims to give *more profile* to the discussion regarding the storage and availability of data for research and raise *the sense of urgency* with respect to making data more operable and more interoperable. The advisory report is a *call for action* to progress from intentions to implementation.

I. KNAW (2012). Responsible research data management and the prevention of scientific misconduct

In the development towards Open Science and Open Data there is a growing need to clarify how research data are being generated and used. The report focusses on publicly funded research only.

Recommendations:

1. Let each field of science decide for itself what is best, within pre-set boundaries.
2. There is no need for more rules, but rather for more activities that enhance the vitality of existing rules.
3. Include existing rules in research evaluation schemes.
4. Organise random sampling at an institutional level to check running practices (also a role for peer review journals).
5. Make data management part of the Standard Evaluation Protocol.
6. Make existing research practices more visible to the public.

Key questions addressed by the report:

- How to identify good and bad practices regarding the handling of scientific data?
- How does the balance between collaboration and competition affect the handling of data?
- How to identify existing routines towards enhancing scientific integrity?
- Who is responsible for explaining the norms of scientific integrity to junior researchers?
- How can misbehaviour be prevented?
- How can careful handling of scientific data at all stages of research be enhanced?

The report makes a clear distinction between 'uncareful' research and 'fraudulent' research. It raises the question whether modern science has enough self-cleansing capacity. Is the image that researchers have of themselves realistic when it comes to proper handling of scientific data?

The report argues in favor of allowing for flexibility during the initial creative stages of research, where there needs to be room for speculation.

II. Elsevier CWTS (2017). Open data. The research perspective

The report is a joint study by CWTS Leiden and Elsevier on how researchers perceive Open Data. The aim is to bridge the gap between policy making on the one hand and daily practices on the other. One problem is that researchers can expect few rewards for making data available to others, while doing so involves considerable complexities. Another problem is that a one-size-fits-all approach is not effective because of varying cultures. The report concludes that more effort is needed at the interface between policy development and the desire to make Open Data a responsible, effective and sustainable action. Open Data mandates would benefit from a better alignment with researcher incentive and revaluation structures.

Data sharing practices vary by field of science: there is no general approach. Many researchers feel they personally own the data they have collected. Less than 15% of researchers share data in a data repository. Of those who do share data, 80% share them with direct collaborators (trust apparently is an important aspect). Key obstacles to sharing of data concern questions of ownership, responsibility and control.

Further observations:

1. There is lack of training in data sharing, which complicates data-storage preparation. There is also a lack of resources.
2. Sharing of data is not associated with credit or reward. Open Data management is often perceived as a burden and not as a responsibility.
3. Privacy as well as ethical, financial and legal issues are hurdles. Use and reuse is hampered by cultural differences between nations.
4. Research *Data Management Plans* by funding agencies are not considered a strong incentive.
5. Guidance on implementation of Open Data mandates is needed, and sharing should be incentivised.

Open Data refers to data that can be freely used, reused and redistributed. Data come in different forms: raw, processed, published, summarized, aggregated, human versus machine readable.

Open Data need to be embedded in the research process from start to finish. There are different needs for 'big science' and for 'little science'. Key questions are:
1. How are researchers sharing data?
2. Do they want to share data?
3. Why would they be reluctant?
4. What are the effects of new practices and new infrastructures?

Data journals (journals exclusively devoted to data) are a recent phenomenon. They are growing rapidly in number. Most researchers prefer to publish the data alongside an article in a data journal rather than in a repository, because they expect that this will lead to more collaborations and more citations, encourages others to reciprocate, and allows for better reproducibility of completed research.

Computer science, physics and astronomy have a well-established tradition of sharing data. In some fields there is no tradition at all. About 65% of the researchers consider themselves the owner of the data. About 50% take the data with them when they transfer to a different institute. Many think (incorrectly) that by publishing the data they transfer their ownership to the academic publisher. There is fear for misuse and misinterpretation of data (and for taking false credit).

Data management requires a lot of effort, time and resources. Some 75% of the researchers do the archiving themselves. Monitoring is done in various ways, with scattered practices across different fields. There is little activity towards standardization, training, organisation, infrastructure. Funding is lagging behind. Most researchers do not perceive their institutions, funding agencies or academic publishers as being in charge.

III. European Commission (2018). Facts and figures for open research data

The survey addresses:
1. The attitude of researchers towards data sharing.
2. The availability of data repositories.
3. The policies of funding agencies and academic publishers.

Ad 1. Researchers have varying reasons for sharing or not sharing data. The current level of sharing is low.

Ad 2. The largest number of data repositories are in the life sciences and the natural sciences, the smallest number in the engineering sciences. The majority are based in the United States, Germany, the United Kingdom, Canada and France, with Australia, Switzerland, Japan, the Netherlands and India being the runners-up.

Ad 3. Funding agencies formulate data management requirements, but these fail to address the needs of the researchers and therefore do not have the desired effect. About 57.5% of the data journals have no policy for Open Data. About 15% encourage Open Data, while only 27.5% require Open Data.

The NPOS (National Platform Open Science) in the Netherlands has 3 key ambitions:
1. Aim for 100% open access to publications.
2. Make research data optimally suitable for reuse.
3. Set up evaluation and valuation schemes to recognise and reward researchers who share data.

IV. European Research Council (2018). Open research data and data management plans

In February 2018, the European Research Council (ERC) issued an *Open research data and data management plan*, which is being updated regularly. For all ERC-funded research, open access of publications is mandatory. The aim is to make research data publicly available when possible, according to the FAIR principles.

Not all data can be made fully open (privacy, security, required delay, intellectual property). Any restrictions should be made explicit and should be justified within FAIR.

Researchers are required to draw up a *Data Management Plan* and to identify a suitable data repository. The ERC provides earmarked funding for this.

Both not-for-profit and for-profit data repositories exist. The ERC offers a list of guidelines and pointers, tuned to the different fields of science.

V. KNAW (2018). Big data in scientific research with personal data

The report focusses on the identification of opportunities for the strengthening of scientific research with the help of big data. Big data differ from normal data in volume, velocity and variety. Big data have innovation potential and economic potential.

The use of big data leads to a refinement of research questions, the development of new research methods, enhanced hypothesis-posing, critical interpretation and qualitative analysis, experimentation with digital methods, algorithmic and computational thinking, application of non-standard statistics (such as data mining, machine learning and deep learning). Key concerns are how to guarantee sufficient quality of big data, how to monitor that big data are 'constructed' (i.e., are the result of many choices made along the way), how to determine in what way big data really alters the research, and clarify the restrictions associated with big data.

The report calls for more public-private cooperation around big data, with a special role for the ministry of Education, Culture and Science and the NPOS (National Platform Open Science); for setting up a national infrastructure that can support local infrastructures; and for the Netherlands taking a leading role internationally. The report argues that researchers are in ever greater need of support from data experts that transcend fields of science. Training possibilities need to be enhanced, and awareness of privacy issues concerning personal data must be strengthened.

The report discusses several examples for personal big data:
1. Personalized health.
2. Geo-privacy.
3. Online marketing.
4. Digital humanities.

It raises the problem of 'journey of data' (data moving from one user group to another) and 'function creep' (data being used for other purposes than originally intended).

The report addresses policy makers and management, but from the perspective of the researcher and his/her environment. It focusses on personal data only.

Recommendations:
1. Researcher: Big data make the proper formulation of hypotheses and theory only more important.
2. Institute: Provide a solid local infrastructure. Provide researchers with proper support on a permanent basis.
3. Universities and UMCs: Provide training for researchers working with personal data. Include data specialists in steering committees.
4. Ministry of Education, Culture and Science: Take the initiative to build an overarching infrastructure that supports local infrastructures (including judicial and ethical experts). Extra investments are needed (together with NWO, SURF, VSNU).

Annex 5. Costs and benefits of storage and availability of data for research

Costs

There appears to be widely shared misconception that the *costs* of storage and availability of data for research are only a small fraction of the total costs of scientific research. In reality, however, the costs are substantial, and they are rising. Clearly, figures vary greatly between fields of science: the costs may be high for astronomy and medicine and low for theology and classical languages. Estimates range from as low as 1% to as high as 10%. Here are some quotes:

- *Agencies supporting research must back up the new open access requirements with funding to ensure their success. Overall, the funding model that provides the highest level of stability, best access for both ingest and retrieval, and greatest equity amongst organisations, is the infrastructure model. The percentage of the total research budget needed to support this approach is likely to be domain specific. We estimate that successful domain repositories can be operated at funding levels of less than 5% of the total research budget. (Some fields might be as low as 1%; the cost might rise to 10% in fields with high data rates or particularly diverse and complex metadata.) These are modest costs to assure a strong return on public investments in the research and to enable uses of data unanticipated by the original investigators.* (Ember *et al*., 2013, page 11)
- *Our informal estimate is that the infrastructure and operation of a truly effective data-sharing system could cost on the order of 5 per cent of total research budgets. For the Commission, which spends over €10 billion a year through its Horizon 2020 programme, that would amount to half a billion euros.* (RDA, 2014, page 33)
- *In addition, well budgeted data stewardship plans should be made mandatory and we expect that on average about 5% of research expenditure should be spent on properly managing and stewarding data.* (European Commission, Directorate-General for Research and Innovation, 2016, page 17)
- *We have estimated that less than 1% of the total amount dedicated to research grants in the life sciences would be sufficient to cover the costs of the core data resources worldwide, including both knowledge bases and deposition databases.* (Gabella *et al*., 2018, Summary)

Benefits

When one talks about costs one should also talk about *benefits*. What are the benefits of storing data and making them available for research? Who benefits, when and how?

The available literature on this matter is clear: the benefits are difficult to estimate. A (complicated) calculation is laid down in the report of the European Commission, Directorate-General for Research and Innovation (2018), written by PwC EU Services. The report states in its Section 4:

- *Interpreting the overall cost of not having FAIR research data as a single value overlooks many non-quantifiable benefits of FAIR. Nonetheless, at €10.2bn per year in Europe, the measurable cost of not having FAIR research data makes an overwhelming case in favour of the implementation of the FAIR principles.*
- *To put this into perspective, research expenditures in Europe amounted to €302.9bn in 2016. While the minimum true cost of not having FAIR can be seen as only 3% of all research expenditures, €10.2bn per year is 78% of the Horizon 2020 budget per year and ~ 400% of what the European Research Council and European research infrastructures receive combined.*
- *To top this, figures for the open data economy suggest that the impact on innovation of FAIR could add another €16bn to the minimum cost we estimated.*
- *While this study does not account for the cost of implementing FAIR, if we assume that the additional costs allocated for data management are up to 2.5% of all research expenditures, this would leave a positive balance of ~ €2.6bn per year from the implementation of the FAIR principles. Moreover, not all the costs for implementing the FAIR principles would be recurrent. Once the proper infrastructure is in place, one could expect the net benefits from the FAIR principles to increase.*
- *Our study presents results for the EU research economy as a whole, however, the cost of not having FAIR data varies strongly from discipline to discipline. In some data intensive disciplines such as genomics or crystallography, (some of) the FAIR principles have already been implemented without the need for a quantified cost-benefit analysis.*

An elaboration of the value and benefits of research data repositories can also be found in OECD (2017) on its pages 15–17:

- *There are many benefits to preserving and making research data openly available. There has been a steady move towards openness in research over the past two decades that has accelerated in the last few years, with some significant changes in the global policy environment in which research is conducted (Bicarregui, 2016). National and international funders of research are increasingly likely to mandate open data and demand data management policies that call for the long-term stewardship of research data.*
- *This trend is consistent with many recommendations and reports by both the members of the OECD and the organisation itself. In fact, in many countries research data are increasingly viewed as an essential part of the research infrastructure. At the same time, adapting research practices to fully implement new open data policy requirements will require a reinvigorated or new infrastructure to support it (Bicarregui, 2016).*
- *Even under the best of circumstances, however, in which governments have a default rule of open data, not all research data can or should be made broadly available. Data may be subject to various restrictions, such as the protection of personal privacy, national security, proprietary concerns, or other forms of confidentiality,*

complicating the decisions to save them and then to make them available. Many datasets are not of requisite quality, are not adequately documented or organised, or are of insufficient (or no) interest for use by others.

- *Nonetheless, there are various reasons, summarised below, why data that are generated in or for research should follow a default rule of openness. Open data and their organisation and curation in research data repositories can generate multiple benefits (Uhlir, 2006).*
- *Economic benefits*
  - *There has been an increasing awareness across governments of the key role that government data or public-sector information (PSI) plays in supporting the goals of research and innovation, specifically, and in economic terms more broadly. Among the studies looking at the value of PSI and its current or potential wider economic impacts, there are a few canonical reports that have gained widespread attention (PIRA, 2000; Weiss, 2001; Dekkers et al., 2006; DotEcon, 2006; Pollock, 2009; Vickery, 2010). While these and other studies document the economic benefits derived from PSI, including open data produced or used in research, they are generally beyond the immediate scope of this report.*
  - *With regard to research data repositories, a recent series of UK-based studies combined qualitative and quantitative approaches to measure the value and impact of research data curation and sharing (Beagrie et al., 2012; Beagrie and Houghton, 2016, 2014, 2013a, 2013b). These studies have covered a wide range of research fields and practices, looking at the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS), the British Atmospheric Data Centre (BADC), and the European Bioinformatics Institute (EBI).*
  - *Two outcomes from these studies stand out. First, there are substantial and positive efficiency impacts, not only reducing the cost of conducting research, but also enabling more research to be done, to the benefit of researchers, research organisations, their funders, and society more widely. Second, there is substantial additional reuse of the stored data, with between 44% and 58% of surveyed users across the studies saying they could neither have created the data for themselves nor obtained them elsewhere.*
  - *While these studies tend to provide a snapshot of the repository's value, which can be affected by the scale, age and prominence of the data repository concerned, it is important to note that in most cases, data archives are appreciating rather than depreciating assets. Most of the economic impact is cumulative and it grows in value over time, whereas most infrastructure (such as ships or buildings) has a declining value as it ages. Like libraries, data collections become more valuable as they grow and as one longer invests in them, provided that the data remain accessible, usable, and used.*
  - *The users of these data repositories come from all sectors and all fields – close to 20% of respondents to the ESDS and EBI user surveys were from the government, non-profit and commercial sectors (i.e., non-academic), as were around 40% of*

respondents to the BADC user survey, and close to 70% of respondents to the ADS users survey. Consequently, value is realised and impacts felt well beyond the research sector alone.

- *Both the scale and extent of these impacts are reflected in citation analyses. Such analyses show widespread and increasing dataset and repository citation spanning a number of years in both academic publications and patent applications, also attesting to both research and industry use. (Bousfield et al., 2016)*

- *Other benefits*
  - *There are many other values, in addition to the economic benefits, that are promoted through the long term stewardship and open availability of research data. These include better research, enhanced educational opportunities, and improved governance. (CODATA, 2015)*
  - *Among the most important benefits are those for research itself - both enabling new research and reproduction of completed research. A fundamental principle for research quality and integrity is the ability of others to verify the results by checking the data used to derive the research findings and to avoid fraud. The underlying data need to be broadly available for verification and reproducibility of results, sometimes even many years after their publication. (Doorn, 2013; NRC, 2009b, 2004, 2003, 1999, 1997)*
  - *Interdisciplinary and international research, including participation of less-developed countries, can be enhanced. Much research now is data intensive and access to many different kinds of data is an essential part of the research process (Hey et al., 2009). Well-curated and open data allow unhindered data mining and automated knowledge discovery; that is, to have machines find, extract, combine, and disseminate the data with minimal or no human intervention (NRC, 2012a). The rapidly expanding area of artificial intelligence (AI) relies to a great extent on saved data. Open data also permit legal interoperability, which is necessary for the generation of new datasets. (RDA-CODATA, 2016)*
  - *Downstream applications and commercial innovation are stimulated by open access to upstream data resources, leading to the creation of new wealth from research, … However, new opportunities are also emerging through collaborative innovation based on data access and 'data driven innovation for growth and well-being'. The beneficial economic effects of open data extend to the research process itself by reducing inefficiencies, especially in the avoidance of research duplication. (CODATA, 2015)*
  - *Furthermore, new types of research are promoted that are important in their own right and can lead also to serendipitous breakthroughs (Arzberger et al., 2004). For example, the collection and open sharing of all kinds of data are fundamental to the rise of citizen science and crowd sourcing approaches, with the data made available through public repositories (Benkler, 2006; Uhlir, 2006). Such approaches have been adopted in many domains, including the space sciences, ornithology (Lauro, 2014; Robbins, 2013), environmental studies, and*

even search and rescue missions (Barrington, 2014). Moreover, greater openness supports the reputational benefits of those who compiled the datasets by making such information more broadly available and generally democratising research (CODATA, 2015). Education, across all ages and disciplines, can be enhanced by access to data from open repositories. At the secondary and even primary education levels students can use open data repositories to further their scientific understanding and skills. University students need open data to experiment with or to learn the latest data management techniques (CODATA, 2015). And of particular interest to government policymakers, data management and curation skills, which require a good educational foundation, are a growth area for employment in an era of shrinking job opportunities. (NRC, 2015)

– *Finally, the role of open repositories of research data in supporting good governance should not be overlooked. Openness of public information strengthens freedom and democrat ic institutions by empowering citizens, and supporting transparency of political decision-making and trust in governance. It is no coincidence that the most repressive regimes have the most secretive institutions and activities (Uhlir, 2004). Open factual datasets also enhance public decision-making from the national to the local levels (Nelson, 2011), and open data policies demonstrate confidence of leadership and generally can broaden the influence of governments (Uhlir and Schröder, 2007). Countries that may be lagging behind socioeconomically frequently can benefit even more from access to public data resources. (NRC, 2012b, 2002)*

There are several other reports that elaborate on the costs and benefits of research data repositories, for example *Turning FAIR into reality* (European Commission, Directorate-General for Research and Innovation, 2018).

Who carries the costs and who reaps the benefits?

The costs of storage and availability of data for research often weigh fully on the local research budget. The question who carries the costs and who derives the benefits is very relevant, but this is beyond the scope of the present advisory report. Is it the individual researcher, the research team, the home knowledge institution, or someone else? Unfortunately, the available literature on the costs and benefits does not examine this question in in detail.