UiO : Universitetet i Oslo

The data explosion – a major challenge, and a great opportunity!

Report by the working group 'Storing and sharing research data' at the University of Oslo (UiO)

Figure 1 Phaistos Disk. Example of early data storage.
The Archaeological Museum of Heraklion (cc by-sa)

Abstract

Research is driven to a growing extent by the availability of large volumes of data. Data volumes will continue to increase, and data-driven research represents a major opportunity in virtually the entire scope of science. Computational science is established as the third paradigm of scientific methodology, alongside theoretical derivation and experimental studies, and data-intensive science is now viewed by many as a fourth paradigm. Gordon Bell et al (1) summarizes this in the article 'Beyond the Data Deluge'. If UiO wants to be a leading research university, we need to take advantage of the opportunities in the data-driven revolution that we are now witnessing. UiO's researchers must therefore be given the tools and expertise that are needed to be at the forefront of data-driven research. 'Riding the wave' (2), the report that really put the importance of the data explosion on the map in a European context, is now five years old.

As a public research institution, UiO must have a clear policy and effective infrastructure for the management (storing, archiving, sharing, curating and retrieving) of research data that imposes requirements on all affected parties, at the same time making it not only possible, but easy to meet these requirements. The institution must provide services, including infrastructure, competence and training, while the researchers are responsible for managing their own research data. Policy and infrastructure must take into account the dynamic nature of research, and be able to manage changes in the researchers' needs in line with technological advances. UiO must help to ensure that the institution's researchers are able to exploit the opportunities in the data explosion, and thereby contribute to solutions that allow fast and effective searches and subsequent utilization of research data, wherever the data are produced and archived.

The report gives an overarching description of the current situation nationally and at UiO, and makes specific proposals for future work. There are four main lines of thought that all future work must be based on:

i. We need to view the entire data management cycle in context; from the generation of research data to retrieval and reuse by others.

ii. We must think globally and not nationally. The solutions we create must meet the needs of researchers in other countries.

iii. We cannot solve all of the issues at UiO, but we must endeavour to resolve problems through cooperation with other national players.

iv. Our guiding principle must always be to devise systems that allow individual researchers to see more advantages than disadvantages in archiving and

sharing their own research data and utilizing the research data of other researchers. A system founded on directives and more audit follow-ups is unlikely to work, see ‘Riding the wave’:

‘Our vision is a scientific e-infrastructure that supports seamless access, use, reuse, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.’ (2)

The working group’s specific proposals are as follows:

i. Clear guidelines on data management at UiO (appendix to the report).

ii. A pilot to establish a programme for competence building and effective research support services.

iii. Clear work sharing (roles, responsibilities and authority) at institutional, national and international level, in relation to the needs for technical infrastructure, and the development of an offer for the temporary storage and sharing of research data with metadata descriptions at UiO.

iv. UiO must help to ensure that some key issues that require national coordination are identified and resolved.

In addition, the working group points out that there is a need for greater awareness of the IT challenges facing the entire organization, and refers to the reports on IT in research and education at UiO. Units should have a much more targeted strategy for IT in education and research, and this must not only be related to the technical eInfrastructure. The scope for potential at the institution and among researchers must be highlighted and safeguarded. This requires further development of IT support functions for teaching and research. UiO should also consider developing a stronger education provision that provides the knowledge society with candidates who have special expertise in handling and using research data (data scientists) in line with the recommendations in ‘Riding the Wave’ (2).

An effective system for managing research data will require broad cooperation, both internally at the institution and with external contributors. A deeply embedded institutional policy that is harmonized with the public sector and funding sources is also needed. Development and delivery of services, establishment and operation of infrastructure, communication and competence-building measures should therefore be centralized to a large extent, so that the policy entails a comprehensive and institutional service. Advisory and support functions should have strong ties to the research. It is important that the policy and the central services are harmonized with and exploit the external structures established in several fields and projects at both national and international level. These structures safeguard the management of research data in accordance with the best international standards.

Responsibilities must be clarified. We believe that a *single* body should have overall responsibility, whilst also being operationally strong enough to drive the work as a whole. We believe that the Advisory group for *eInfrastructure*, as proposed in the report on IT in research (*IT i forskning*) (3) should have this coordinating responsibility and be the driving force for the ongoing work.

CONTENTS

1. A great opportunity

The term ‘data explosion’ refers to the enormous volumes of digital data we generate globally, either directly for research purposes or as a result of our digital lives. The data explosion is immense, and perhaps we do not realize just how pervasive it is; how it is in the process of changing not only research but society as a whole. In one day, a modern DNA sequencing machine can read several billion parts of the human genetic code. In the course of a year, such a machine generates several terabytes of data (trillions of data units). It is not easy to relate to such figures, but we can get an idea when we consider that the annual output from such a machine is equivalent to the information currently found in 20 libraries on the scale of the US Library of Congress (2). A single specialized instrument in one sub-field of science and in just one year – all from an instrument type that does not in any way produce very much research data. At the ‘Swedish Solar Telescope’ at La Palma, the Institute of Theoretical Astrophysics receives 2.5 terabytes of data per day under good observational conditions. Now imagine the ‘big picture’ across all disciplines, over decades, and globally. Perhaps we then get more than just an idea – we see the scale of the data management challenge, but we should also recognize the great opportunity that lies in these enormous volumes of data.

The above example focuses on technology, science and medicine, but the situation is no less important for the humanities and social sciences. ‘Large volumes of data’ can be interpreted in various ways in the different disciplines, and the term ‘long-tail data’ refers to the enormous number of small diversified datasets that are produced globally within different disciplines, collections of various kinds that either represent isolated studies or series of studies over many years. One such example is the ongoing digitization of the National Library of Norway’s collections, which is paving the way for new important research projects in a variety of humanities and social science disciplines. The data volumes involved here are also large and increasing rapidly, and the potential to utilize the total dataset is immense. We must, at an institutional, national and global level, meet the needs of different disciplines through diversity in the types of research data and variation in perspectives.

The data explosion represents a paradigm shift for research (4), and the global scientific community must be proactive and facilitate the conditions for development. Research is a spiral in which new knowledge is constantly being added to existing knowledge, which then gives rise to new research, which in turn generates more new knowledge. If we want to be at the forefront of research, we must keep abreast of the steadily increasing rate of development and the ever-growing volume of valuable research data. This, however, requires the research community to be able to find and retrieve the required research data quickly and effectively, and subsequently compile the information and start the research on the basis of existing information and knowledge.

We are thus facing a variety of challenges. How do we exploit the explosive increase in volumes of global research data? Can we, with the right framework, verify other researchers’ results and interpretations to a greater extent, and thereby force a general quality improvement? And not least, if we avoid unnecessarily duplicating studies that have already been conducted, how should we utilize the resources this will save? The scope for potential is huge. The development will support, for example, interdisciplinary and large-scale studies related to key societal challenges, such as poverty, energy and global warming, and will facilitate convergence between research issues and disciplines.

2. Storing, archiving, sharing and curating data – a brief introduction

A brief introduction to the key concepts is required in order to consider the measures proposed in this report. Research data is defined here as ‘*registration/recording/reporting of numerical scores, textual records, images and sounds that are generated by or arise during research projects*’ (5). The Research Council of Norway chooses to make a distinction between input data and output data. Input data is research data that are retrieved from external sources and exist independently of the research project in which they are to be used. Output data are described as ‘data generated through research’. When output data are archived, they thereby become input data.

We have used a schematic research data flow as our basis (Figure 1), and for most purposes this will serve as a good model. In other cases, this simple schematic illustration will not describe the problem quite as well. Nevertheless, we believe it is useful to set the scene.

During the start-up/planning phase of a research project, the researcher gathers input data (marked with 0 in Figure 1). What has been done previously and what research data exists? Even at this early stage, the researcher should decide how and when new research data (output data) are to be shared and archived. Several funding sources already require a data management plan to be described in project applications. The researcher should cover the following in the plan:

i. Whether research data can be shared or whether they are exempt (see Chapter 3).

ii. The terms for sharing data (including *when*).

iii. Which channels the research data should be published in.

iv. How long is it desirable/appropriate to store research data.

The researchers who have generated the research data will often need a reasonable period of time to exploit them. This is known as the right of first use, and must be reflected in the data management plan.

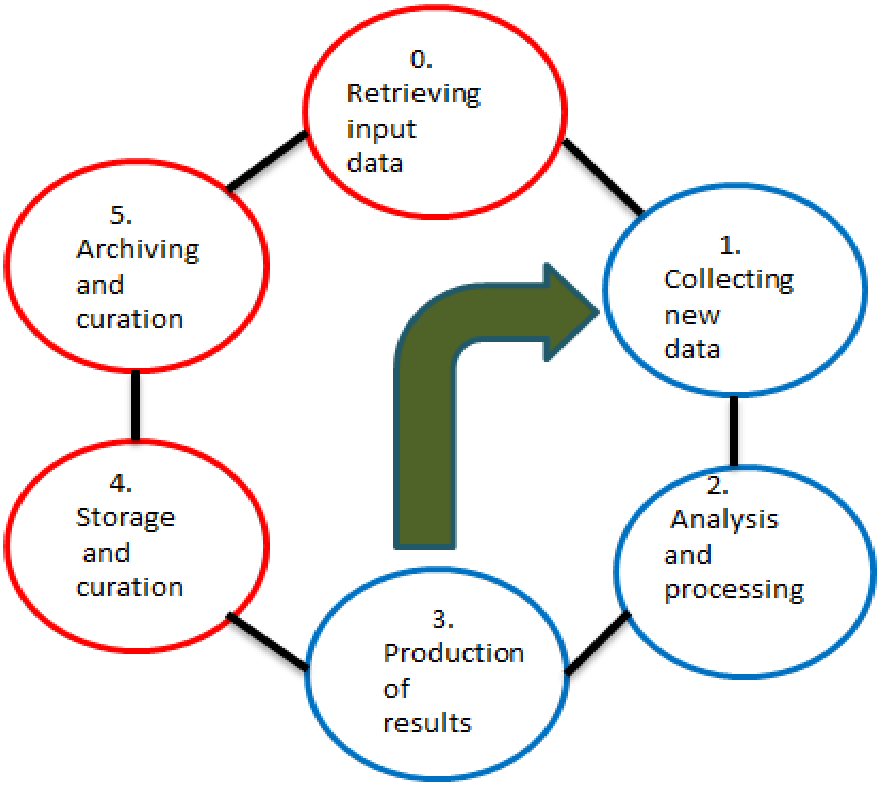

Figure 1. Schematic research data flow. The report deals with the red circles, while the blue circles are covered in the report on IT in research (IT i forskning)*.*

Using this method, the researcher/research team is ready to collect/produce new research data (1). Collected raw data often need to be processed before they can be analysed (2). The analyses will often reveal that the results (3) at this point do not give answers to the questions raised, and a need arises for a new/different data collection (1). An internal research flow will thus be established in order to generate results that give answers to the original questions. In this research flow, there will be a need to store research data, which must be curated on an ongoing basis (4). The research data will typically be shared internally in research groups, often with external participation. Storage solutions for sharing research data and co-authorship for the entire research team (including external contributors) are therefore highly desirable. In the final phase in this simple schematic flow, the research data are archived for future use (5). The research data are now ready to share with the world at large, for example via what we refer to in this report as *retrieval services*. These are tools/services that researchers can use to retrieve and access relevant global research data.

Archiving and sharing entail research data being utilized by others, for new purposes. *Metadata*, the information that is necessary to enable the research data to be used by others, must therefore accompany the research data. Metadata can include:

- Information on data formats, and on where, when and with what instruments the research data were collected.
- Information on, for example, technical equipment and software used, and any discipline-related standards that have been followed.
- Information that makes it possible to retrieve the research data, and to understand whether there are any restrictions on reuse.

Unique identifiers, often called digital object identifiers (DOIs), must also be linked to the research data and often also to the researcher(s) responsible for the collection(s). Archiving also helps to ensure the *reproducibility* of the study, where the results are verifiable. This is important in a time when various quarters are reporting that significant parts of the studies published globally are not reproducible.

Curation describes the entire process of collecting, archiving, maintaining and preserving research data for concurrent and future use. A research data archive without curation is like a library where the books are not described and classified. Retrieval of research data would therefore eventually become impossible. The curation also entails considering what should and should not be preserved, as well as ensuring that the preserved material is readable for the future. The data curator/data archivist's task is to quality assure metadata and manage the research data collections. Much of this data curation must take place in, or in cooperation with, the research communities that generate the research data.

Retrieval solutions that can help researchers to find research data globally are a significant challenge. Search engines such as Google are very useful tools for navigating unstructured datasets such as websites or PDF documents. It is far more difficult to search, combine and filter structured information in databases. In the world of data science, this is a major challenge, in which UiO is a global leader. Professor Arild Waaler heads, for instance, one of the major Centres for Research-based Innovation, SIRIUS, where the key focus is on searches. An important perspective is that there is no single user-friendly interface for all searches/research questions. Customization is used to some degree, so domain-specific solutions are needed.

There is also a considerable body of regulations on data management, and the protection of personal privacy and copyright are particularly challenging. Projects covered by the Health Research Act, the Personal Data Act, the Personal Health Data Filing System Act or the Biotechnology Act must, for instance, have prior approval from the Norwegian Data Protection Authority or Regional Committees for Medical and Health Research Ethics (REC). Copyright issues are related to who owns research data, the content of existing databases, and new databases that are wholly or partly based on existing databases. The use of health registers entails a number of legal issues, for example.

Given the complexity of data management, the variety of players and the huge volumes of data involved, an effective data management system will require both the institution and the individual researcher/research group to have a clear strategy for managing data. This must be based on clear guidelines, while the implementation of this policy will require competence building, support services and a good infrastructure with an intuitive user interface.

3. National guidelines based on international obligations

In September 2014, the Research Council of Norway presented its policy on open access to research data (5). This can be seen in a longer development perspective, and is summarized on the Research Council's website:

- *In 2007, the OECD adopted the 'Principles and Guidelines for Access to Research Data from Public Funding'. Norway has undertaken to follow up these guidelines.*
- *In the two most recent parliamentary reports on research, the Government has emphasized that it wants to facilitate greater access to publicly-funded research data. The Ministry of Education and Research asked the Research Council of Norway to establish a policy for open access to publicly-funded research data.*
- *In 2012, the European Commission recommended that the member states develop guidelines for open access to research data.*
- *As part of Horizon 2020, guidelines were drawn up on open access that cover both publications and data (10).*
- *In 2013, the European Commission established the Research Data Alliance (RDA) in collaboration with NSF in the USA and the Australian National Data Service (ANDS). (6)*

The objectives of the Research Council's guidelines are as follows:

- *Quality improvement in the research through greater opportunities to build on earlier works and collate data in new ways*

- *Transparency in the research process and greater opportunities for verifying scientific results*
- *More cooperation and less duplication of research work*
- *More innovation in the business sector and public sector*
- *More effective management and better utilization of public funding (5)*

It is also worth noting that the Research Council wants to be a driving force for the preservation and sharing of research data. This partly entails:

- *Implementing procedures in the application process to ensure that relevant applications include plans for data management*
- *Implementing procedures in the project follow-up to ensure that the plans for data management are adhered to by the projects receiving funding*
- *Continuing the practice of contracts requiring research data to be archived in a responsible manner for a minimum of 10 years*

Guiding principles govern the concept of open access to scientific research data. The OECD report: 'Principles and Guidelines for access to research data from public funding' serves as a guide, and stipulates that 'open access to research data from public funding should be easy, timely, user friendly and preferably internet-based' (7).

The Research Council's policy also follows the 'open as standard' principle in relation to access to research data. While Horizon 2020's definition of open access includes a requirement for the access to be free, the Research Council believes that the user should cover the actual costs associated with the retrieval of research data. This is closer to the OECD definition of open access, which states that access should be provided at the lowest possible cost.

There are several challenges associated with making some datasets openly accessible. The Research Council's policy specifies some valid reasons for restricting access:

> *Security concerns: When accessing data can threaten personal or national security, the datasets* ***must not*** *be made openly accessible.*
>
> *Sensitive personal data: When open access to the data conflicts with the applicable statutory framework regarding the protection of personal privacy, the datasets* ***must not*** *be made openly accessible.*
>
> *Other legal factors: When open access to the data conflicts with other legal provisions, the datasets* ***must not*** *be made openly accessible.*
>
> *Commercial factors: Data that have commercial value and are generated in projects in which a company is a contractual partner with the Research Council* ***may*** *be exempted from the general principle of open access. In these cases, it is recommended that the data are made available after a certain period of time, preferably after three or five years.*

Other factors: When open access to data will have major financial or practical implications for those who have generated/collected the data, the datasets ***may*** *be exempted from the general principle of open access if a satisfactory argument is made for this (5).*

4. Current situation at UiO

UiO is a large university with a broad scope, and it is no surprise that there are many different types of research data, with correspondingly different needs for infrastructure, services and support. The working group has conducted a survey of the current situation at all units, which shows that there is considerable variation even internally within the units. However, there are also many researchers who have the same kind of research data and similar needs across the units. There are therefore opportunities for the development of generic solutions. The survey shows that many of UiO’s researchers do not have a well-considered approach to archiving and sharing research data. Despite the fact there is no widespread culture of sharing research data openly, researchers at UiO are generally positive to sharing. At the same time, the researchers require good solutions that are adapted for the researchers and that protect their interests. Such solutions also require support in the form of a transparent internet resource and not least good guidance. That means that one of our main challenges will be cultural; although it is easy to argue for the overarching concept of sharing, effective implementation depends on the research community collectively feeling that they are well served by the arrangements introduced.

The survey shows that there are many examples of good practice at UiO. There are several research groups, and even entire disciplines, that have good systems for data storage and sharing, such as in languages, environment and astronomy. Not surprisingly, the units with a carefully devised IT strategy and well-developed IT support generally have far greater awareness and practical solutions than other units.

We also have examples of unique generic systems developed by the units that attract a great deal of attention. USIT, for example, has established Services for Sensitive Data (TSD), a platform for the safe handling of all types of research data with statutory or self-imposed requirements for information security. Common to such data is strict authentication and authorization, aimed at controlling who accesses what data and how the data are processed and used. The strict access regime encompasses not only users but also operators. Developing TSD has been a challenge because the security aspect is often incompatible with users’ wishes and established operating procedures. Nevertheless, we now have a TSD platform that is flexible in the sense that new open archives can be included, and that the solution enables the secure exchange of information between systems within TSD and outside.

UiO also has an agreement with the Norwegian Social Science Data Service (NSD), which entails NSD making a prior assessment of projects that intend to process personal data in order to ensure that they are carried out in line with legislation. NSD also does a follow-up at the end of the project to ensure that the projects are concluded in line with what has been reported and statutory requirements. This agreement has some limitations, however, in relation to archiving and sharing research data, and we therefore recommend a clarification and, if necessary, a revision of the agreement. One example of a relevant issue for many UiO researchers is the storage of research data in an audio or video format. In this connection, NSD notes that 'If audio or video recordings are made (which can identify individuals) in connection with the project, these must also be deleted/destroyed or censored if the data needs to be anonymous'.[1] Many researchers at UiO use audio and video data as the primary source, or as a reference to give meaning to other types of research data (e.g. sensor data). Here the current practice is that most of the recordings must be deleted after the project is concluded, even although the material may be a valuable source for future research projects. This arrangement is not optimal, and there is generally a significant need for training in the relevant regulations, support services for clarifying and interpreting regulations and best practice solutions for regulatory compliance. There is also a need in this context to clarify the responsibilities and powers of the various national agencies, for example, in connection with the protection of personal privacy.

In general, the situation at UiO (and nationally) is not satisfactory in light of the goals set. A small number of the storage systems used at UiO for storing the University's own research data can, however, be used for open or access-controlled sharing. The majority of the storage systems at UiO are internal systems in which research data can only be shared within the research group internally at the University. Researchers wishing to share data with other collaborative partners tend to use e-mail or external cloud services such as DropBox. This type of sharing solution does not meet the guidelines of the Research Council of Norway or the EU's pilot for open access to research data (5,8) (and nor is it intended to do so).

[1] *E-mail K. U. Segadal NSD, 9 March 2015*

It is important to note that this report covers the storage of research data. UiO's archive system, ePhorte, is excellent for its intended use, but is not a solution for archiving research data with a view to global accessibility.

Based on the responses from the survey, there is currently no general solution for complying with the Research Council and the EU's policy for the vast majority of users at UiO. The challenges are many. The next chapter outlines some measures that will advance our work. The emphasis here is on solutions for the vast majority of users, who do not currently have effective national or international solutions.

5. Storing, archiving, sharing and curating data – proposed outline solution

The current situation is characterized by a wide variation both between and within UiO's units. There are many challenges, and the solution we outline can be summarized in three main points:

i) Establish clear guidelines for storing and sharing research data at UiO (see Chapter 6 and appendix).

ii) Support competence building and establish support services for all relevant target groups, and thereby develop a collective understanding of systems that can enable the individual researcher to put in place sound strategies for data management.

iii) Establish a solution for storing different types of research data that is easily accessible from several client platforms and which is also flexible, secure, transparent and long-term. Naturally it is important to ensure a smooth transition to archiving in national and international archives. The user perspective must be in focus and effective interfaces must be developed towards the end users.

If we are successful, we will be able to both share our research data with others and in principle utilize others' research data effectively. In order to fully exploit the data explosion, we must also develop tools for successful retrieval of research data. Points ii) and iii) are discussed in more detail below, while Chapter 6 outlines more specific proposals for measures at UiO. Challenges that require national coordination are described in Chapter 7.

Competence building and support functions

General competence building in the establishment and organization of efficient support functions is required to enable UiO to implement a well-planned policy for research data in practice. Many researchers lack knowledge of and competence in archiving and sharing their research data, finding relevant research data in archives and reusing others' research data correctly in their own research. Courses and training must be offered. Support functions must also be established in an increasingly data-intensive workday. Therefore, we need workshops, formal training programmes, courses in data citation and systematic data collection, curriculums that deal with the challenges of data-driven research etc. Chapter 6 presents proposals for specific measures, and puts a strong emphasis on the strategic development of IT support services that will increasingly be directly involved in education and research. It is imperative that these are based on the needs of teachers and researchers.

eInfrastructure and the interface between user and system

We must define and delimit the problem area before we can outline a more detailed technical solution aimed at satisfying as many as possible of our researchers' perceived needs for eInfrastructure and interface solutions. We have chosen to define four areas of responsibility for research data infrastructure and services. This division allows us to specify measures and to outline a realistic solution at institutional level, even though it does not provide a stringent definition of what research data are included in each area of responsibility. The division into four areas of responsibility is illustrated in Figure 2. It largely corresponds to the data cycle in the research process (see Figure 1).

The research process (project) begins with a research question based on existing knowledge and previous observations (research data), and the researcher designs a process that will lead to new knowledge. The process involves the collection of research data – either new or existing – and further manipulation and analysis of these data. The process often generates new research data. The kind of infrastructure necessary when processing research data is individual and dynamic. The individual researcher or the research project itself bears responsibility for this. The figure shows this as the lowest responsibility level (level 1). The UiO report on IT in research (*IT i forskning*) (3) discusses eInfrastructure generally and includes infrastructure, services and the support functions employed in this part of the data cycle.

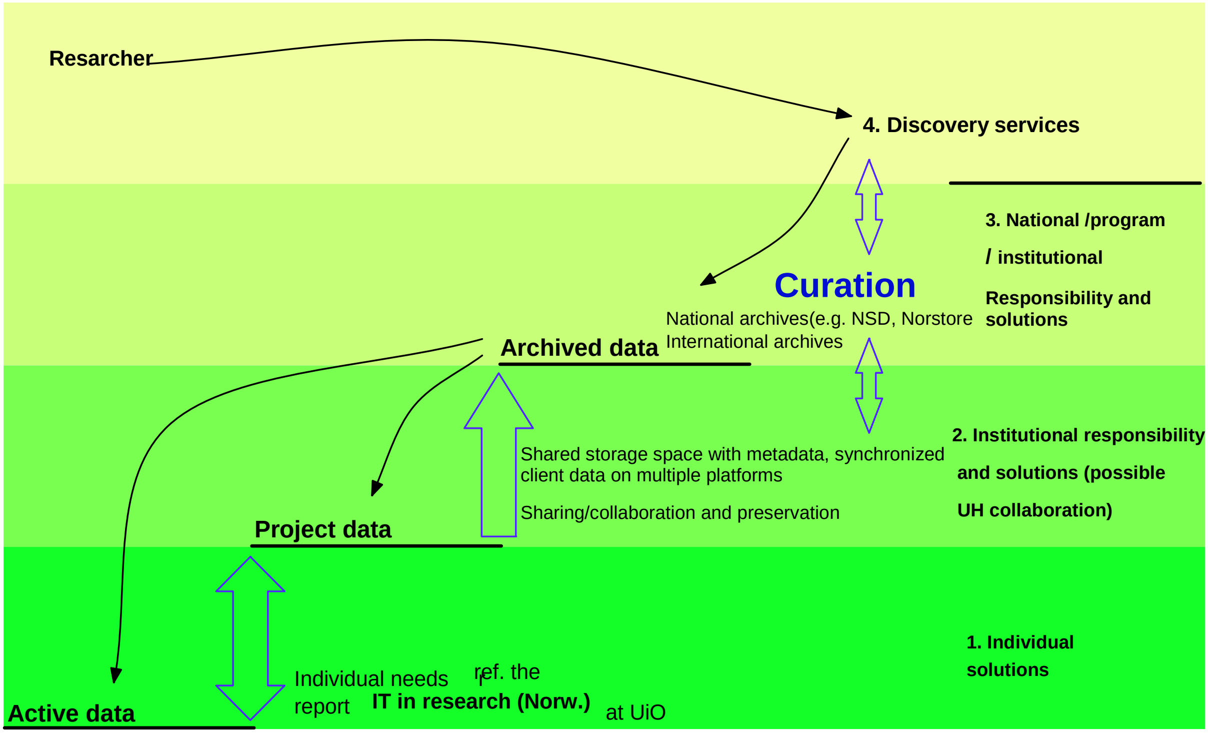

Figure 2. Responsibility for solutions and dynamics in a fourfold division of 'the management of research data'.

The next two areas of responsibility include research data that are not merely accessible to the individual researcher but that are shared and have a degree of openness. How open and freely accessible the shared research data will be depends on several factors such as legal, copyright or commercial requirements. The difference between research data at levels 2 and 3 lies in assumed or established research value, and is often correlated with the research data's degree of 'maturity'.

Level 2 represents a central solution for storing and sharing research data that is secure and accessible on several client platforms (user systems). Various implementation solutions exist today but current solutions pose a number of significant challenges. User interface is a keyword. Without a good user interface, the task is insurmountable for the individual researcher/research group. Moreover, an essential functional requirement to the solution is that it secures metadata and has appropriate access levels. The working group believes that it is necessary to establish a good infrastructure for level 2 at UiO. This level lies between the individual solutions linked to the active use of research data (level 1) and established or international systems for the archiving of final research data (level 3). At the same time, a new infrastructure of this kind has the potential to give researchers the necessary incentives to register metadata and thereby promote the sharing of research data.

Level 3 is associated with analyzed research data of value to the research community, i.e. research data that is worthy of archiving. Here we find published datasets that are equipped with a DOI/PID, which must be quality assured, enriched with metadata and 'locked' so that no changes can be made. There are both national (Norstore, Norwegian Social Science Data Service (NSD), National Archives of Norway, museum archives) and international programmes/services and institutions that are responsible for the archiving of data and making the data accessible (even though there are also challenges in this respect). Our researchers can choose the most appropriate solutions for archiving (in relation to their research area and any applicable statutory framework). At the same time, these archives must offer an efficient and high-standard user interface. The current systems are far from optimal. This is discussed further in Chapter 7: 'Challenges we cannot solve alone'.

It is noteworthy that the figure shows a strong correlation and dependence between the different areas of responsibility. As an extension of individually owned research data, there is a need to share and use research data jointly (internally in a research project but also with external national or international partners). Moreover, a secure and structured 'temporary storage' option is needed for data where no decision has been made on whether the data should be permanently archived or if there are other circumstances, for example legal requirements or a ban on publication, that preclude full openness or automatic storage. The level 2 solution we propose would cover this. Research data can be stored at level 2 pending quality assurance, structuring and annotation using appropriate metadata. These processes may be time-consuming.

Retrieving and thus reusing stored and archived research data demand that the associated metadata are of high quality. A central service for storing and sharing research data at UiO for research projects without an established infrastructure (level 2) must set out requirements for adequate metadata such that the subsequent archiving (level 3) is easy to perform. We propose built-in incentives, for example free storage (or storage at cost price for large projects) when the essential information (metadata) has been entered.

It is vital that the institutional solution supports research collaboration and facilitates a sufficiently secure sharing of research data. Today, research data are shared via all kinds of channels from email to Dropbox, which entails risks related to traceability and security. Institutional storage should therefore be as simple to use as cloud solutions, and should be independent of platforms and allow secure authentication and management of access. By the same token, it must also be possible to give access to and share research data with partners outside UiO. The authentication mechanisms and other structures facilitating this already exist in the higher education sector in Norway, and there are also international authentication mechanisms that can be used (e.g. eduGAIN).

So far, we have dealt with systems that assist researchers to make their own research data accessible through archiving. Conversely, the potential and incentive for the researcher are linked to the possibilities inherent in using others' research data. This demands sound retrieval systems for research data that harvest metadata from relevant archives and show where research data are accessible. These are not in place today. Such retrieval systems should be established nationally since a range of small, scattered solutions will quickly become chaotic, complex and of varying quality. We would emphasize the importance of putting such retrieval systems in place, and suggest a specific measure for health data in Chapter 7.

6. Operationalization and the road ahead

Based on the current situation and the applicable statutory framework as well as the opportunities inherent in a functioning, global research society that shares and uses research data across research groups and national borders, we propose the following measures:

i) Establishment of clear guidelines for data management at UiO. Proposed guidelines are given in an appendix to this report.

ii) Establishment of sufficient IT competence and support throughout the organization. Many units need to devise an IT strategy, incorporated if possible into the unit's main strategy, together with a related plan of measures. In this connection, needs for expertise in data archiving and data curation must be met. The same applies to the need for a data strategy. In some cases, the needs of subordinate units can perhaps be covered by a centralized unit at faculty level. However, the working group's survey shows that IT needs nowadays play such a major role in teaching and research in all disciplines that they must be more clearly highlighted in overall budgeting than is currently the case. In clear terms, IT needs must be stressed not only in the annual budget of the unit but also in the individual research project budget. What is required in the project, and how will the related costs be covered?

iii) Establishment of a pilot that will ensure that a programme for competence building and good research support services is created. The survey shows that UiO researchers are generally positive to sharing research data. However, they also pose clear requirements regarding good technical solutions that are tailormade for the researchers and that safeguard their interests. Exploitation of these technical solutions also requires strong research support in the shape of well-structured webpages (see an excellent example of how this can be done at data.bristol.ac.uk), training and high-quality supervision. The support will include guidance to researchers on choice of infrastructure, what tools can be employed, retrieval and citation of research data, what metadata standards are relevant, and also what storage systems best safeguard researchers' wishes and at the same time fulfil the requirements of funding sources. Training must be based on a limited number of short training modules that cover user needs effectively. We suggest that a pilot is set up with the aim of devising an internet resource and a comprehensive cluster of training modules, such as:

- Data management for project leaders.
- Start-up module for master's degree students under the auspices of the University of Oslo Library. Use of existing search engines and citation practices should be included.
- Similar or related module for PhD students according to demand.
- Range of courses (internal and external) for general competence enhancement among potential course holders at different levels.

In light of the work input of our academic employees, it is crucial that these modules are concentrated and user-focused. We must provide a range of modules that motivate rather than frighten. Online training should be considered. The pilot must assess the balance between the general and the generic, and more specialized training must take place at faculties or units. User involvement in the pilot is essential in addition to ongoing user evaluation of the courses that are developed.

The webpage that is created must offer good guidance (templates, examples of good practice etc.) but it must also provide support in respect of legal aspects such as privacy protection and copyright (see below).

iv) Establishment of a level 2 solution for temporary storage of research data with metadata. This must support the researchers and must therefore offer a well-developed user interface. It must offer an option to share research data and co-authorship with colleagues nationally and internationally. Incentives work, so at the very least, research projects should be offered low-cost or even free storage space when submitting project information and metadata. A working group including researchers and representatives of the IT organization must be set up to specify the offer in more detail. The specifications will form the starting point of an institutional initiative, for example through the Advisory group for eInfrastructure proposed in the report on IT in research (3).

v) In line with the recommendations of the EU's High level expert group on scientific data, UiO should also consider developing stronger programmes of study that would ensure that the knowledge society is provided with data scientists – experts who have the ability to exploit the possibilities offered by a future well-functioning global system for sharing research data. The University has an option that can function as a platform for the further development of a dedicated course of study.

In addition, there are a number of challenges we cannot solve alone. These are described in more detail in Chapter 7.

The distribution of responsibility must be clarified. In our view, *one* body should have overarching responsibility while being operationally strong enough to secure progress overall. Other units can and should have responsibility for subsets of the overall problem area. We believe that the future the Advisory group for eInfrastructure should be responsible for coordinating and driving the work going forward, including:

i) Follow-up of policies and guidelines

ii) Implementation of institutional solutions (level 2)

iii) Competence building and good research support services

7. Challenges we cannot solve on or own

The challenges we cannot solve on our own as an institution vary in character and importance. Deciding on a system for the national archives and national solutions for identifiers and metadata is the most important.

- **National archives for research data.** Keywords are responsibility, work sharing, user interface and user focus. What is required for the next generation Norstore? Can NSD cope with the requirements the Research Council of Norway sets for the research community to archive research data with NSD's assistance, or is there a gap between the requirements set and what is offered? A well-ordered, overarching national system for archiving research data that cannot be deposited in international archives is essential.

- **Identifiers and metadata.** Research data must be archived in systems that can be shared with the rest of the world, where international standards for metadata and digital identifiers such as DOI for objects and ORCID for the researcher accompany the datasets. The archives must be curated to enable reuse. Moreover, harvesting of metadata through open APIs should be permitted so that the data can be retrieved by the various retrieval systems established. This will also allow the establishment of systems that generate overviews of the use and reuse of research data. This is necessary in order to establish reward systems for researchers (see below). It is essential that national providers of identifiers are established as quickly as possible so that NSD, Norstore and other archives can use these. Moreover, it is highly desirable that established archives should provide open access to metadata, for example through APIs, and that the metadata adhere to relevant international standards, e.g. CERIF.

The next two challenges are different in character because the statutory framework and the various actors already exist. Nevertheless, we are of the opinion that we need a joint national interpretation/guidance for teachers and researchers. This is a complex matter, and from the perspective of cooperation, work sharing and concentration, it seems unnecessary that the work is duplicated at different institutions.

- **Privacy protection.** Research activity is subject to strict regulation to ensure privacy protection when necessary. There appears to be a need to clarify the responsibility and authority of different bodies. Since the Health Research Act came into force, REC has had responsibility for prior approval of research projects covered by this Act. The individual institution must put in place institutional internal control. For research not covered by the Health Research Act but that comes under the Personal Data Act, a prior assessment by the Data Protection Official for Research at the Norwegian Social Science Data Services (NSD) is required. NSD’s current role and mandate in connection with long-term archiving of research data is perceived today as unclear and outdated (the mandate is primarily based on text-based data). A clarification of NSD’s/the Data Protection Official’s role therefore appears to be necessary. Sensitive personal research data cannot be made openly accessible. Even open anonymized data can lead to sensitive personal conclusions if such data sources are incorrectly assembled. However, the definition of sensitive personal data is unclear and must be expanded to include inappropriate collation of data (REC defines video data as sensitive personal data per se, while NSD/the Data Protection Official assesses this on a project-to-project basis). Visual data/video data pose a challenge to the applicable statutory framework, and a broader understanding of sensitive data is needed. It may be expedient to initiate a dialogue with the Norwegian Data Protection Authority and NEM in order to ensure a joint understanding in line with the technological development.

- **Copyright.** We must clarify how research material subject to copyright such as literary texts, newspaper photographs, TV reports and recordings of concerts can be used, archived and shared. In this connection, several levels of copyright may be applicable: composers have rights to their own work, musicians have rights to their performance, record labels have rights related to the release. Many of these rights apply for a considerable period of time after the persons concerned are dead, and there are different rules in different countries. How shall we deal with this at the different levels of access and sharing in the systems that are developed?

The final two challenges we will focus on are of a different nature. The need for a rewards system discussed at the end is quite challenging although not as pressing, and it may be an advantage to postpone it for some time.

- **Retrieval solutions.** Sharing and reuse are enabled through good retrieval systems for research data, and a strategy is required for developing and making such systems accessible. In general, domain-specific solutions will provide the solution going forward. These retrieval systems may equally well be international as national, but nevertheless it is tempting to propose a national system for searches in structured health data since Norway is in a unique position as regards health registers. UiO already possesses solid competence in this field, and a project in cooperation with OUS/HSØ would strengthen both institutions.

- **Rewards system.** It is desirable/necessary to reward research and researchers that practice good research data management and give others access to their research data. A considerable body of research data is made accessible when studies are published, but a system rewarding the production of good research data and making it accessible is desirable. In this respect, we note the ongoing establishment of a number of data journals in which it is possible to publish datasets and metadata. This kind of publication will give the authors/data owners credits in the form of publication points.

References:

1. Bell G, Hey T, Szalay A. Beyond the Data Deluge. Science. 2009 Mar 6;323(5919):1297–8.

2. High level expert group on scientific data. Riding the wave - How Europe can gain from the rising tide of scientific data - Final report of the High Leverl Expert Group on Scientific Data A submission to the European Commission [Internet]. European Union; 2010 [sitert 2015 Mar 23]. Tilgjengelig fra:: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

3. Arbeidsgruppe for IT i forskning ved Universitetet i Oslo. IT i forskning ved Universitetet i Oslo - Rapport fra arbeidsgruppe for IT i forskning ved Universitetet i Oslo (UiO). Oslo: Universitetet i Oslo; 2015 Jan 37 s.

4. Hey T, Tansley S, Tolle K, editors. The Fourth Paradigm: Data-Intensive Scientific Discovery. 1 edition. Redmond , Washington: Microsoft Research; 2009. 284 s.

5. Forskningsrådet. Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd. Norges forskningsråd; 2014.

6. Norges forskningsråd. Forskningsdata skal deles - Norges forskningsråd [Internet]. forskningsradet.no. [sitert 2015 May 6]. Tilgjengelig fra: http://www.forskningsradet.no/no/Nyheter/Forskningsdata_skal_deles/1254000298821?lang=no

7. Gurria A. OECD Principles and guidelines for access to research data from public funding [Internet]. OECD; 2007 [sitert 2015 Mar 9]. Tilgjengelig fra: http://www.oecd.org/dataoecd/9/61/38500813.pdf

8. European Commission. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 [Internet]. European Commission; [sitert 2015 Jun 5]. Tilgjengelig fra: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Appendices:
Background material for the document;
Survey at the faculties conducted by the working group:
https://www.usit.uio.no/om/organisasjon/uav/itf/saker/forskningsdata/kartlegging/

In conclusion, we wish to express our thanks to our academic staff and external staff at the Ministry of Education and Research and the Research Council of Norway for useful and sound input to the report.

Appendix 1. Working group’s mandate, members and work

Mandate:

- Survey existing services and practices, and researchers’ needs for storing and sharing data at UiO.
- Develop proposals on principles and guidelines for storing and sharing/making research data accessible, safeguarding the rights of academic staff.
- Recommend solutions that reflect the increasing requirements of funding sources regarding storing and sharing data, and the researchers’ own needs.

Members of the working group ‘Storing and sharing research data’:
Chair Vice Dean (Research) Svein Stølen, MN
Senior Engineer Torben Leifsen, MN
Associate Professor Tor Endestad, SV
Senior Engineer Torgeir Christiansen, UV
Head Engineer Tore Miøen, OD
Head of Department Alexander Refsum Jensenius, HF
Senior Engineer Espen Uleberg, KHM
Research Director Fridtjof Mehlum, NHM
Senior Adviser Katrine Ore, MED

Secretariat:
Head of Department Hans Eide, USIT (Department for Research Computing)
Senior Librarian Live Kvale, UB (Science Library)
Adviser Margaret Fotland, AF (Education and Research Administration Office)

Legal support:
Lawyer Einar Noreik, AF(Education and Research Administration Office)

Appendix 2. Proposed policy and guidelines:

Policy

The University of Oslo wishes to manage research data[2] in accordance with the most stringent international standards, thereby supporting the development of a global research society in which research data are widely shared. This will promote:

- quality improvement in the research through greater opportunities to build on earlier works and collate research data in new ways
- transparency in the research process and greater opportunities for verifying scientific results
- more cooperation and less duplication of research work
- more innovation in the business sector and public sector
- more effective use and better utilization of public funding

UiO will make it as easy as possible for staff and students to follow the applicable statutory framework at all times. This means that UiO must have clear guidelines for data management, good training programmes in competence building, and webpages that provide support, as well as efficient support services for eInfrastructure. The entire organization together with relevant external partners/funding sources must collaborate in order to implement good practice under the framework conditions set by the legislation and funding sources.

Research data must be:

a. accurate, complete, genuine and reliable
b. identifiable, retrievable and accessible
c. securely and safely stored, either centrally at the researcher's own institution or in national/international archives in accordance with the stipulated requirements
d. maintained in accordance with legal and research ethics obligations
e. capable of being made accessible to others in line with relevant ethical principles for sharing research data.

[2] Research data means registration/recording/reporting of numerical scores, textual records, images and sounds that are generated by or arise during research projects.

Research data must be stored/archived as long as they are of value to the researcher and the broader research community, and as long as indicated by the funding source, patent provisions, legislation, embargo stipulations and other official requirements. The shortest storage period for research data is three (3) years after publication/release. In most cases, research data will be kept longer than the minimum three-year period. In general, research data should be made accessible at the earliest possible point of time, but after a first right of use period for the research team itself.

If research is supported by a contract/agreement/grant containing specific provisions on ownership, storage and access to research data, the provisions of this agreement/contract will take precedence.

If research data are to be deleted or destroyed, either because the agreed period of storage has expired or on the grounds of legal provisions, this should be carried out in accordance with all legal and ethical principles, as well as the requirements of funding sources and partners, taking into account confidentiality and security.

UiO's guidelines for archiving, making accessible and sharing research data

1. Research data must be stored/archived in a secure manner
    a. The data must be stored in secure archives, either centrally at the researcher's own institution or in national/international archives
2. Research data must be made accessible for further use
    a. Research data must be made accessible to all relevant users under the same conditions, insofar as there are no legal, ethical or other security-related reasons for not doing so (see below)
3. Research data should be made accessible at an early phase
    a. Data on which scientific articles are based should be made accessible as early as possible, and never later than the date of publication
    b. Other data that may be of interest for other research should be made accessible within a reasonable period of time and never later than three years after the project has ended
4. Research data must be supported with standardized metadata
    a. The metadata will enable others to search for and utilize the data
    b. The metadata must adhere to international standards
    c. The metadata must include a description of the data quality

    d. If the data forming the basis of a publication are selected from a larger dataset, the dataset must either be published or described

    e. If observations have been removed from the dataset, the exclusion procedure must be described and justified

5. Research data must be supported with licences for access, reuse and further dissemination

    a. Licences should be internationally recognized

    b. Licences should place as few restrictions as possible on access, reuse and further dissemination of the data

6. Research data should be made freely accessible, but the actual costs of dissemination should be covered

    a. Metadata should be made accessible free of charge and should be published in a way that facilitates automated harvesting and use in searches for research data

7. Research data should be supported with a long-term plan

    a. A plan should be prepared on how to manage data that is deemed to have long-term value

    b. Academic staff should have a well-considered approach to how research data that is assessed as not having long-term value should be managed, or destroyed if appropriate after a certain period of time.

Researchers have a responsibility to manage research data in accordance with the principles and requirements stated above. This means that they must develop and document clear procedures for collection, storage, use, reuse, access and storage or destruction of research data in connection with their own research. This will include the division of responsibility in partnership projects with other institutions. The information must be described in a data management plan. Legal framework conditions and the requirements of funding sources must be safeguarded.

Main principle for open access

UiO’s policy adheres to the ‘open as standard’ principle in respect of access to research data. UiO will therefore help to ensure that research data will be made more freely accessible in principle, and that exemptions are made for data that cannot or should not be made accessible (see below). Access must be provided at the real cost of accessibility. The exemptions include:

Security concerns

- When accessing data can threaten personal or national security, datasets must not be made openly accessible.

Sensitive personal data

- When open access to the data conflicts with the applicable statutory framework regarding the protection of personal privacy, the datasets must not be made openly accessible.

Other legal factors

- When open access to the data conflicts with other legal provisions, the datasets must not be made openly accessible.

Commercial factors

- Data that have commercial value and are generated in projects in which a company is a contractual partner with UiO may be exempted from the general principle of open access. In these cases, it is recommended that the data are made available after a certain period of time, for example after three or five years.

Other factors

- When open access to data will have major financial or practical implications for those who have generated/collected the data, the datasets may be exempted from the general principle of open access if a satisfactory argument is made for this.