TALLINNA ÜLIKOOL

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

UNIVERSITÀ DEGLI STUDI DI PARMA

Education and Culture DG

ERASMUS MUNDUS

Live Håndlykken Kvale

Data Sharing in the Life Sciences

A Study of Researchers at

The Norwegian University of Life Sciences

Master thesis
International Master in Digital Library Learning
2012

ABSTRACT

Digital research data collected in the sciences has the potential to be reused and shared openly. Several arguments for such sharing have come forward both from funders and researchers during the last decade. This study investigates the attitudes towards such reuse along with current traditions for sharing, reuse and the storage of research data in the universities, particularly the Life Sciences in Norway.

A comprehensive survey has been conducted among researchers at the Norwegian University of Life Sciences (UMB) collecting data on various aspects of the 'researcher – research' data relationship. The two main focuses are practical issues regarding storage, sharing and reuse of research data and perspectives on the future of data sharing, issues regarding publishing channels and the usage of online research collaboratories are also covered. The research aims to create an understanding of how researchers handle the data they collect, how they retrieve research data for reuse from other sources and how they imagine the future potential of data sharing. Based on the understanding of these processes, recommendations for the implementation of data repositories are made.

The collected data has been analyzed according to years of experience, research field and previous experience with data sharing as well as compared with data from previous studies in the field, in order to understand which factors influence the researchers' opinion or experience. These factors are again discussed and analyzed before conclusions are drawn.

The thesis concludes that the processes of data sharing are far from optimized as researchers today primarily retrieve data from colleagues. In order to create repositories for data sharing and storage, the researchers must improve their understanding of metadata to ensure that no information about the research data is lost. For the researchers to be willing to share their data certain criteria must be fulfilled, in addition, the fears among the researchers for misuse must be taken into account. Many of the attitudes towards the sharing of research data depend on where the researchers stand in their careers; experience with data sharing is to a larger extent dependent on the specific research discipline.

KEY WORDS

Research data, Open data sharing, E-science, E-research, E-science librarianship, Norwegian University of Life sciences,

"The amount of data that scientists produce continues to increase every year. People are needed to handle, preserve, describe, and organize that data, and, because many of these tasks are similar to what librarians have done with publications for centuries, it makes sense that librarians would have a role in the emerging task of managing scientific data."

(Szigeti & Wheeler, 2011)

ACKNOWLEDGMENT

I want to give special thanks Lisbeth Eriksen, Liv Ellingsen, Andreas Åkerstrøm and other librarians and staff at UMB. It has been motivating to discuss and explore the topic of the thesis with people who have long experience in academic libraries. Furthermore your help with the questionnaire and the sample population was essential for me to be able to complete this study.

Thanks to my supervisor Ragnar Nordlie for all help and feedback on my ideas from the stage of selecting a topic to completing the thesis.

Friends, family, and fellow students have also given me great support, by reading, commenting and criticizing, being patient with my obsession with research data or simply making me think of something else, when that has been necessary.

ABBREVIATIONS USED IN THIS THESIS

Association of Research Libraries	ARL
Centre for Integrative Genetics	Cigene
European Space Agency	ESA
Department of Animal and Aquacultural Sciences	IHA
Department of Chemistry, Biotechnology and Food Science	IKBM
Department of Ecology and Natural Resource Management	INA
Department of Economics and Resource Management	IØR
Department of Mathematical Sciences and Technology	IMT
Department of Landscape Architecture and Spatial Planning	ILP
Department of Plant and Environmental Sciences	IPM
Department of International Environment and Development Studies	Noragric
Norwegian University of Life Sciences	UMB
Permanent Access to the Records of Science in Europe	PARSE
The University of Oslo	UIO

TABLE OF CONTENTS

1. Introduction

In scientific communities the debate on how to deal with growing amounts of research data, often referred to as the data deluge, has received more and more attention over the last 5 years. Questions arising in this debate point both at what possibilities the combination and proper visualization of data can give, and towards the preservation issues. There are also growing expectations regarding what impact e-science could have on the research libraries. “E-science has the potential to be transformational within research libraries by impacting their operational functions, and possibly even their mission”. This was the introductory statement made in a report for developing e-science in research libraries in the US from 2007 (Lougee mfl., 2007). Others argue that collections of research data will become the new special collections (Borgman, 2008). Still no matter what collections of research data will be called, they are already having an impact on digital collections in research libraries around the world likely to expand to all research institutions handling scientific data.

Principles of open science are already a part of the daily life of an academic library with institutional repositories and open access. Borgman traces the notion of Open Science back to Saint Augustine and the fourth and fifth century and the base to be the idea that scholarly information is a “public good” (Borgman, 2007, s. 35). She further splits the idea of public goods in to two elements; that “they can be shared without lessening their value”, and the economic term “non-rival”. According to Borgman, the emergence of e-research is a return to old values of research output being a public good in the context of the technologies. Open research data is a continuation of initiatives towards open access to research publications, and open science forms the umbrella term for both of them.

1.1 Background

This thesis investigates the potential of research data, from the researcher’s perspective. As part of this their attitudes towards open science are also touched upon. There is however little research done investigating the relationship between researchers, scholarly output and technology. Further I hope to add something new to the research field by presenting both data and theories about how researchers reuse, store and retrieve digital research data. This is done through investigation of the obstacles and possibilities researchers expect to encounter with the increased sharing of research data, along with an investigation of how the research data is stored, used, and reused today. These combined give a thorough description of digital research data from the perspective of the researcher.

The analysis is however slightly influenced by the LIS perspective as some special attention is given to the metadata issue and the data curator role. By applying literature from the Social Sciences, the theories are placed in a context of research on researchers. This is one of the issues that are recommended for further investigation as not enough material from knowledge sociology, which investigates the motivation of the scientist, has been found. In addition to theories on how researchers work and how various factors influence their attitudes and experiences with the sharing of research data, recommendations for implementation of data repositories are made. Through these recommendations this thesis will hopefully be useful as a source of background material on the research data life cycle, with suggestions on how to improve the reuse and storage of data.

1.2 PROBLEM STATEMENT

In her book "Scholarship in the digital age" (2007) Christine Borgman quotes computer scientist, Jim Gray, who said, "May all your problems be technical" (Opportunities & Council, 2004, s. 95) indicating how the real challenges are understanding "what to build, for whom, for what purpose, and how their usage of technologies will evolve over time" (Borgman, 2007, s. 3). Technical problems tend to have one or two solutions that make them work well or at least work, whereas the problems Bergman presents do not always have clear solutions.

Data repositories are being constructed for researchers to preserve and access their own data and data from other researchers. The purpose of these repositories are however also more complex as there are many visions of how grids of research data can be used in new ways in the future, by combining them in new ways to explore other aspects of the data than what it initially was collected for. Further validation and the possibility of using research data in education, are presented as other purposes for stored research data. Data repositories are therefore built not only to serve the researcher and their storage and retrieval of research data, but also for the data to be used and combined in new ways. What to build can be investigated by studying the needs and knowledge the researchers have of data repositories, their experiences from similar projects and literature in the field. How the usage will evolve over time is however more difficult to predict. Several experts and expert groups have, as the literature review shows, presented their visions for how scholarly research will be conducted, presented or preformed in the future. A thorough investigation of how the researchers work today and which issues these researchers are dealing with will however put the focus on actual challenges, rather than assumptions about the future. Digital research data is by many of the researchers at UMB used on an everyday basis, by detecting which challenges they expect to meet in the future a more proper answer might be given to all the issues Borgman mentions. Understanding how researchers deal with their data, along with understanding the full lifecycle of research data, are two of the keys to constructing functional technical systems. Such a study will help to place the research data in a context from collection, through usage in scholarly publishing and for the future through long term preservation, as a source for the researchers from which to extract new knowledge.

1.3 Research question

This study investigates the status in the field of data sharing through collecting opinions from researchers both on how they work, share and publish and furthermore on how they perceive the obstacles or possibilities with open sharing of research data, by aiming to answer the following question:

What are the attitudes towards and experiences with sharing and reuse of scientific/research data among researchers in the life sciences in Norway?

1.4 Aims and objectives

The aim of this study is to better understand how researchers work with the data they collect, how they reuse data from others and how they would like to share their data in order to make recommendations for the optimization of data sharing and reuse. By better understanding the current situation, along with fears and threats the researchers see in data sharing, recommendations and theories on how to improve the situation can be made. Data sharing can be seen as an application of the open science principles, and a continuation of open access initiatives were the research libraries are prime movers. The research library is a core support function at a university and an exhaustive study of how researchers use data will help the libraries to meet the new challenges of data preservation. In order to look at the topic from all angles, a broad study is done and the most important challenges that stand out are further discussed and analyzed.

1.5 METHODOLOGY

This study is based on a survey where a questionnaire is used for the data collection among researchers and PhD students at the Norwegian University of life sciences (UMB). This choice was made in order to collect a broad sample of opinions from researchers in different departments, and with different years of experience. Trends and patterns in experience and opinions were then collected and analyzed. The questionnaire used for the data collection is based on previous questionnaires, and by this allows a comparison between different universities. In addition to the survey, a literature review detects challenges pointed out in previous studies to support the findings from the collected data.

1.6 SCOPE OF THE RESEARCH

The study was conducted among the researchers at UMB, a small but specialized university right outside of Oslo. There are a total of 650 people holding academic positions at the university, and they have all been invited to answer the online questionnaire. 23% of the researchers from the different departments responded and the attitudes, opinions, experience and knowledge expressed by these 147 researchers, is the source of the data presented.

1.7 LIMITATIONS

Any study needs to have limits in terms of scope, time and data collection. This thesis is the result of work mainly done within the strict limitations of one academic term. More time could have opened up the possibility both to investigate further the issues that came forward as important and given the researchers more time for responding to the questionnaire. Both initiatives could have helped to produce broader data that could be used more for generalization and not just for indication of potential issues. 23% of the researchers replied to the questionnaire and the study is solely based on these answers given by about one fourth of the total number of researchers at the university. What attitude and experiences the remaining 77% have towards data sharing remains unknown. Also the literature review had to be completed at some time even if there is a constant publishing of new and relevant material in the field so the material used is again limited to what was found at the time. There is definitely more literature out there that would have been relevant to include.

The limitation to UMB and the researchers there is also important to keep in mind; it is not possible to say if other universities would have produced different results. The comparison with previous studies and between disciplines helps to emphasize this limitation.

1.8 Significance of the study

This study will hopefully be important in two ways.

Firstly it is hoped to identify needs and challenges the researchers encounter when dealing with digital research data, so that structures for sharing and with this, the usage of metadata, can be improved. The study will add to a landscape of other case studies and best practices and hopefully add something more by giving full attention to the researchers' perspective.

Secondly, the sharing of research data, data repositories and data curation are still fairly new challenges for the LIS community, where the word e-science librarianship is occurring more and more frequently. In the context of Norway however the development of cyberinfrastructure is taking place without involvement from the LIS community. Hopefully this study can be used in the emerging debate of e-science librarianship in Norway.

1.9 Thesis outline

The thesis is divided into six chapters:

Introduction
The first chapter introduces the research domain, describes the aims and objectives of the study along with the research question. Further methodology, scope, limitations and significance of the research is presented.

Literature review
The second chapter begins with definitions of relevant terms and concepts, and is followed by a comprehensive review of literature relevant to this study and previous research in the field. After a conclusion of what the literature shows is drawn, ideas of what this thesis will add to the field are presented.

Methodology

In the third chapter the selected research method is presented and argued for along with a presentation of the sampling strategy and data collection techniques.

Results

In the fourth chapter the findings from the data collection are discussed according to subject and comparisons. They are then presented and analyzed through text and figures.

Discussion

In the fifth chapter the findings presented in the fourth chapter are analyzed and discussed according to relevant issues and factors that have come forward through the presentation of the results.

Conclusion

In the sixth and last chapter, a conclusion based on the research question is presented. Further recommendations based on the findings are made along with direction for further research in the field

In addition to the six chapters an alphabetical list of references and appendixes are found at the end.

2. LITERATURE REVIEW

The literature review begins with definitions of the relevant concepts, before an investigation of literature in the field is done. The literature is grouped into categories according to aims and issues it discusses. Due to the fact that it has been challenging to find literature dealing explicitly with the issue of data sharing, the literature focuses somewhat more on the role of the libraries in the emerge of e-science. Finally a conclusion based on the literature is made and a presentation of what this study will add to the existing literature is done.

2.1 DEFINITIONS

Before going into the subject, a definition of relevant concepts and terms should be in place to clarify and explain what meaning is added to the essential concepts and terms.

2.1.1 DIGITAL LIBRARY

A comprehensive definition for digital libraries was developed by a multidisciplinary group in 1996. The definition is still found to be valid, and forms the idea of what a digital library is:

> 1. *Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describes various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consists of links or relationships to other data or metadata, whether internal of external to the digital library.*
>
> *2. Digital libraries are constructed – collected and organized – by [and for] a community of users, and their functional capabilities support the information needs and uses of that community. They are component of communities in which individuals and groups interact with each other, using data, information, and knowledge resources and systems. In this sense they are extensions, enhancement, and integration of a variety of information institutions and physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community. These information institutions include, among others, libraries, museums, archives, and schools, but digital libraries also extend and serve other community settings, including classrooms, offices, laboratories, homes, and public spaces. (Borgman, 2007, s. 17)*

These two definitions visualize the variety of material and purposes a digital library can serve, and leave open the possibility to include a collection of research data, properly described, selected and organized with future usage in mind, to be one kind of digital library.

2.1.2 Research data

For research data the following OECD definition is used; "factual records (numerical scores, textual records, images and sound) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated"(Gurria, 2007). A definition that includes much material collected with the purpose of research.

2.1.3 Data curation

Another term that appears is data curation. Here I rely on the definition used by G. Sayeed Choudhury "maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the life-cycle of scholarly and scientific materials" (Choudhury, 2009, s. 96). According to (Szigeti & Wheeler, 2011) data curation is often considered a subfield of e-Science. It is however a subject close to traditional librarianship and is also starting to be included as both a graduate and a post graduate course in some LIS curriculums in the US (Walters, 2009).

2.1.4 E-science and Cyberinfrastructure

The UK National e-Science Centre defines e-science as *"the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists."* («National e-Science Centre definition of e-Science» 2012). It is also referred to as the European version of the American term cyberinfrastructure (Jankowski, 2009, s. 5; Luce, 2009, s. 4). Cyberinfrastructure is then again defined as *"the comprehensive infrastructure required to capitalize on advances in information technology; cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and* *experimental facilities, and an interoperable suite of software and middleware services and tools*"(Friedlander, 2009, s. 78).

The two terms, e-science and cyberinfrastructure as mentioned above, are often referred to as European and American terms for the same concept. There are however some differences as cyberinfrastructure is a term that more strongly emphasizes the infrastructure aspect while e-science refers to the whole concept of the new way of doing science. Amy Friedlander explains the differences; *"Thus E-science and cyberinfrastructure, while closely allied, actually represent slightly different emphases. E-science connotes a research program; while cyberinfrastructure recognizes the role that the engineered and institutional infrastructure plays in instantiating and fostering computationally-intensive research and, importantly, devotes resources to advancing capabilities of the infrastructure."* (Friedlander, 2009, s. 79). In this thesis the two terms are used as related concepts according to Friedlander, rather than the synonymous usage, as this seems more appropriate according to the newest literature where e-science and cyberinfrastructure are used as distinct but related terms.

Kowalczyk & Shankar (2011) describes e-science merely as a consequence of the 2.0 development on the web: *"Web 2.0 is also being utilized in other ways, primarily by the user community. Contemporary science is influenced by competing impulses that permeate the daily work practices of scientists and the institutions in which they work. One such impulse is the need to share research results, data, and tools to promote greater impact of research, leverage scarce resources, conduct longitudinal studies, apply new analytical tools to existing datasets, and verify and refine results of other researchers. Data sharing also has the potential to foster multidisciplinary researc*h". They further describe the researchers' initiatives to do open research while respecting the intellectual property domains as a response to the fact that little research data is freely accessible. Freely as according to the principles of open software, principles developed and put on paper in the late 80's by internet pioneers such as Richard Stallman (Stallman & Free Software Foundation (Cambridge, Mass.), 2002) and have had a large influence on how web based collaboration is done today.

2.1.5 E-RESEARCH

The term e-research is another one appearing in a related context first used in 2002 for all types of net based research in the humanities, from 2004 researchers in the humanities used the term more *"as a critique of the notion of e-science rather than as a generic word for internet research of virtual methods"* (Jankowski, 2009, s. 56) this critique addresses three different aspects of e-science; "*the* *underlying philosophy of science, assumptions about diffusions of e-science across academia, and expectations about the technological infrastructure needed"* (Jankowski, 2009, s. 56).

The usage of e-science versus e-research also seems to be somewhat discipline related, e-research being embraced by the humanities and social sciences while e-science is commonly used in the hard sciences where the infrastructure aspect is more relevant (Jankowski, 2009, s. 61). In the LIS field, the term e-science tends to be used in a broader sense by including networked science in general (Kowalczyk & Shankar, 2011; Szigeti & Wheeler, 2011) and as such is more similar to the term e-research. There is however a continuous development and sometimes disagreement in matters of term definitions. Dreyer et. al. (Dreyer, Bulatovic, Tschida, & Razum, 2007) argue that "e-science should be understood in a broader sense" here referring to the need to have a term that covers more than grid computing and large infrastructures, but rather the new way of doing science.

While some go as far as to call it a new research paradigm («The fourth paradigm», 2010) others refer to it as a new research methodology (Hey & Hey, 2006; Lesk, 2008) while others again speak about a fourth pillar in science: *"High-performance computing (HPC) has played a central role in establishing the importance of simulation and modeling as the third pillar of science (theory and experiment being the first two), and the growing importance of data is creating the fourth pillar"* (NSF, 2012). What kind of changes e-science makes in the sciences is not a question this thesis will answer, it is however evident that new technologies and the possibility to access, store and preform larger parts of the research process in an infrastructure developed for research and science, will have an impact on how researchers work and possibly affect and change several aspects of how science is done. As Michael Lesk (2008) says "*As we get larger and larger data resources, and more and more intelligent data mining software, it becomes easier to make discoveries by going through existing data, rather than collecting new data."* The present change in how research is conducted, and the increasing extent based on data mining, is creating a need to establish good routines for data curation and a reward system for researchers who produce quality data that is shared with the community. "*We need to recognize the increased role of data, and the importance of rewarding people who save it and know how to analyze it"* (Lesk, 2008). As an example of the need to focus on data curation, (2008) uses the example of NASA losing its original copy of the video recording of the first Moon landing, and temporarily losing even the audio recording, eventually finding it in a box labeled "bad tape". Most of us do not hold material of such unique value; it does however illustrate how little attention is being given to data preservation and what kind of valuable data can get lost from malpractice in this field.

2.1.6 Open science

As presented in the introduction Borgman (2007) puts the origin of the notion “open science” back to the fourth and fifth century. She states that “As an economic framework, open science is based on the premise that scholarly information is a “public good.” Public goods have two defining elements. One that they can be shared without lessening their value; the economic term is “non-rival.” The second characteristic of public good is that it is difficult and costly to hold exclusively while putting it to use; the economic term is “non-excludable.”” (Borgman, 2007, s. 35–36). Borgman further places e-research in the context of open science by stating that “e-research did not emerge from a political vacuum. In many respects, it is a return to centuries-old principles of open science in response to challenges wrought by changes in technology and social policy, especially with regards to intellectual property” (Borgman, 2007, s. 44). With this statement, Borgman highlights the new possibilities given by technology in the old tradition of the sharing of scholarly output.

2.1.7 Open research data

Open research data is a term that is used in this thesis, despite the fact that no definition of the concept has been found in the literature. It is however found to be a necessary term in order to distinguish between the open access initiatives that are primarily working with access to research publications and initiatives that are making research data openly available. The term, open data, is found to be too general as in LIS community it is often used then talking about access to metadata and library catalogues, and in the broader community it is used when talking about providing access to all data produced with government founding. Open research data is then used as a term applying the principles of open science defined in 2.2.6 with research data defined in 2.2.2. Research data produced with government funding is viewed as a common good that should be preserved for the future and made openly available. Openly available can imply access with no restrictions or access to a defined user community. In order to facilitate proper preservation, issues such as selection criteria and metadata become relevant.

2.1.8 Grid computing

One last term that currently appears in the literature of the more technical kind, is grid, the grid or grid computing. According to Borgman (2007, s. 17) the word grid can be understood as “a power grid, which provides pervasive access to electricity" and the usage of grid in the context of grid computing as "one of several technical architectures that will underpin digital scholarship". The term is not central in this thesis, it is however still given space among the definitions in order to provide a context to the term cyberinfrastructure.

2.2 Literature review

E-science is a topical issue both in LIS and in science in general and this is reflected in the literature. Several articles, papers and even an open access journal on the topic, journal of e-science librarianship, were published during the first two months of 2012. Due to the usage of different terms and the involvement of different scientific fields the collection of relevant literature has been a long process, where coincidences have sometimes led to the finding of new and relevant material. Occasionally the material found has been focused on the scientific field in which data is being made available, rather than the openness itself. In others again the technical focus on the grid infrastructure has made the material less relevant, and focused more on covering a tangible field where the computer technology background is key to understanding certain concepts. Reading has however been the only way to sort out relevant literature, and often again led to the discovery of new literature. The starting point for discovering this field has however been the chapter "Data sharing in the Sciences" (Kowalczyk & Shankar, 2011) in the Annual review of Information Science and Technology. Exploration of the bibliography has led to a better understanding of the field as well as to further literature. Another bibliography that has been used is "Science and technology resources on the internet" (Szigeti & Wheeler, 2011). In addition, searches in databases have been performed. These search results have tended to be "noisy", both due to the sometimes inconsistent use of terms, and as a result of the term e-science being used and mentioned in a broad context. "Data sharing" in the LIS field is more commonly used in the sharing of metadata records than on the involvement in e-science. Open research data does not appear as a concept in the literature, still it is a term that most specifically describes the field of interest and is a core in the concept of e-science librarianship. An example of inconsistent terminology is how these three articles in the journal of e-science librarianship use the following key words:

Article 1: eScience, DataONE, data-intensive science, cyberinfrastructure
Article 2: Data Management, Data Curation, eResearch, Competencies, eScience Librarianship
Article 3: research data, library services, tiers of service, NSF data management, University of Massachusetts Amherst, education, consultation, infrastructure

These three articles all being of interest to this literature review have no coherence in the terms used as keywords. Furthermore, the keywords reflect the terminology used in the articles, so it would therefore have been difficult to conduct one search to find all the three articles in a more general journal in the LIS field. However searches in journals and databases have led to findings, where the relevant references again have been explored. Other relevant literature has been discovered through networking and communication with others in the LIS field interested in the future of e-science.

The literature has in general three directions, or a combination of these. One is the vision about how the sciences and publication channels will develop and how e-science can stimulate new sciences. This direction is related to the vision presented earlier, such as the ones from PARSE.insight and Microsoft research. The other direction focuses more on technical challenges, often metadata issues, and towards issues of grid computing. The third focus is the role that information professionals or data curators can play to help with these challenges and which qualifications are needed in such a profession. Much of the literature combines the three as they are closely connected.

One last point that is made in the literature is the need to reward those who create good data and have good routines for saving and sharing their data (Lynch, 2007). This can be connected to the vision of a different publication system, where a wider range of publications would need to get the reward and accreditation. Heuer on the other hand is more critical regarding using money as a motivator for the sharing of research data stating that "Financial stimulations will only come when you show that you are developing a strategy, never think first about the money, first ideas then money." (Le Diberder, Heuer, Mele, &Diaconu, 2009) The balance between using money as a direct motivator or reward versus investing money in projects based on sustainable ideas for data preservation, is important to keep in mind for the founders, not only when speaking about the preservation of research data.

2.3 Methods used in the literature

The material collected can be split into different categories based on its aim and function. Some literature gives political guidelines, roadmaps and makes declarations aiming to point out the direction and describe paths and visions for the future. Examples of this include the OECD guidelines, the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities and the Science Data Infrastructure roadmap by PARSE.insight. What these have in common is that they are often written by committees or groups of people trying to predict and make strategies on how to meet the needs of the future. This literature is political and indicates how policy makers want research to develop. It also offers a vision on how they expect science to be conducted in the future. Further, there is the literature that states opinions about e-science based on experience from the field of science, LIS or both, sometimes based on studies, other times based more on experience. Other literature again includes studies or descriptions of various e-sciences and related projects forming a more practical approach to the topic, often giving best practice advice. Both of these are found in journal articles, books and book chapters, presentations, blogs etc. As there is a growing amount of material out there, portals collecting material are also appearing. This review intends to give a thorough overview of what has been written and done in the field of e-science, limited to the two available languages, English and Norwegian. According to Kowalczyk & Shankar, much of the literature is based on mixed method case studies of data. These focus mainly on “data access and use in cyberinfrastructure”. Kowalczyk & Shankar are further requesting “more comparative studies across disciplines and infrastructures, as well as case histories of successful and unsuccessful data sharing” (Kowalczyk & Shankar, 2011, s. 33). This is mainly supported by the literature reviewed here. There is however, a growing amount of success stories and best practices, examples include articles written by Sayeed Choudhury at the John Hopkins University (Choudhury, 2008, 2009, s. 93), documentation of eSciDoc in Germany (Dreyer mfl., 2007) and DataONE (Allard, 2012). In addition to mixed method cases studies, surveys using questionnaires collecting both qualitative data (Creamer, Morales, Crespo, Kafel, & Martin, 2012) and quantitative data (Thaesis & van der Hoeven, 2010) are used in investigations of the topic. The comparative studies across disciplines and infrastructures that Kowalczyk & Shankar are requesting are still not present.

Borgman (2007, s. 10) requests research exploring “the continuum from primary to secondary sources” and “the entire life cycle of data, from data generation trough the preservation of the scholarly products that set the data in context.” In other words; research that goes into the research process and investigates and describes recurring issues.

In Norway, what has been found is the survey by Grøttum (Grøttum, 2010) using a quantitative questionnaire that was mentioned earlier. Further, Amuth Gastinger (Gastinger, 2012a) has done a study of the presence of e-science in University policies, with a focus on library involvement. A large part of the material on the topic is still based on a combination of experience and visions, rather than scientific studies. Some such as the PARSE.insight group combines the two, by performing a large scale study to create arguments supporting a large scale vision for the future of e-science.

2.4 Political documents -

Governmental guidelines, roadmaps, declarations and initiatives

Governments, being the primary funders of research in European countries are adding sharing and availability to their vocabulary when speaking about research output. While outreach and education stand as the main arguments for the funders, the possibility to combine, prolong, test new ideas and make new discoveries stands as the main argument from the researchers advocating data sharing (CERN, Le Diberder, Heuer, Mele, &Diaconu, 2009). In much of this literature the two sides of the argument for sharing are combined and presented as the strategy for future research. There are however also researchers arguing against sharing and trust, prestige, integrity and copyright are some of the key words used in their arguments. In the political literature these arguments are often presented as challenges or potential obstacles to the development of e-science.

One of the earliest documents in the political guideline category in Europe is probably "The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities", drafted in 2003, to "promote the internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, founding agencies, libraries, archives and museums need to consider" (Luce, 2009, s. 8). Further, the roadmap and vision for 2020 promoted in the report by Microsoft Research (Microsoft Research, 2005) written by researchers and computer scientists from Europe and the United States initiates with the statements: "A scientific revolution is just beginning. It has the potential to create an era of science-based innovation that could completely eclipse the last half century of technology-based innovation; and with it, a new wave of global, social, technological and economic growth." In their report the focus is primarily on the potential that lies in using advanced computer technology in particular within the natural sciences. The impact of the report is found in later literature where the vision presented for the future of science is shared and often referred to.

In 2007 the "OECD principles and guidelines for access to research data from public funding" (Gurria, 2007) were published forming an important guideline in the field of e-science. "A recommendation is a legal instrument of the OECD that is not legally binding but through a long standing practice of the member countries, is considered to have a great moral force" (Gurria, 2007). These OECD guidelines and definitions are used and referred to in much of the literature in the field. The main arguments for accessibility to research data presented in the OECD guidelines are as follows:

> *The good stewardship of the public investment in factual information;*
>
> *The creation of strong value chains of innovation;*
>
> *The enhancement of value from international co-operation, more specifically, improved access to, and sharing of data:*
>
> *Reinforces open scientific inquiry;*
>
> *Encourages diversity of analysis and opinion;*
>
> *Promotes new research;*
>
> *Makes possible the testing of new and alternative hypothesis and methods of analysis;*
>
> *Supports studies on data collection methods and measurement;*
>
> *Facilitates the education of new researchers;*
>
> *Enables the exploration of topics not envisioned by the initial investigators;*
>
> *Permits the creation of new data sets when data form multiple sources are combined.*
>
> *(Gurria, 2007, s. 10)*

Further the OECD guidelines state that *"sharing and open access to publicly funded research data not only helps to maximize the research potential of new digital technologies and networks, but provides greater returns form the public investment in research".* Following this comes their recommendations regarding the 13 principles; openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability and sustainability. Different aspects are described and recommendations are currently being made.

2.4.1 UNITED STATES

In the US the National Science Foundation (NSF) with their Office of cyberinfrastructure are important advocators and founders of the e-science development. Their latest publication in the field is the "Cyberinfrastructure for 21stCentury - Science and Engineering Advanced Computing Infrastructure - Vision and Strategic Plan" Here they state that their new NSF vision "supports both computational and data-intensive research coming from simulations, scientific instruments, 'cloud' computing, and sensors" (NSF, 2012).They illustrate how things are inter-connected in order to function in the following figure, where the cyberinfrastructure is placed within, and forming only a piece of the total context of e-science.

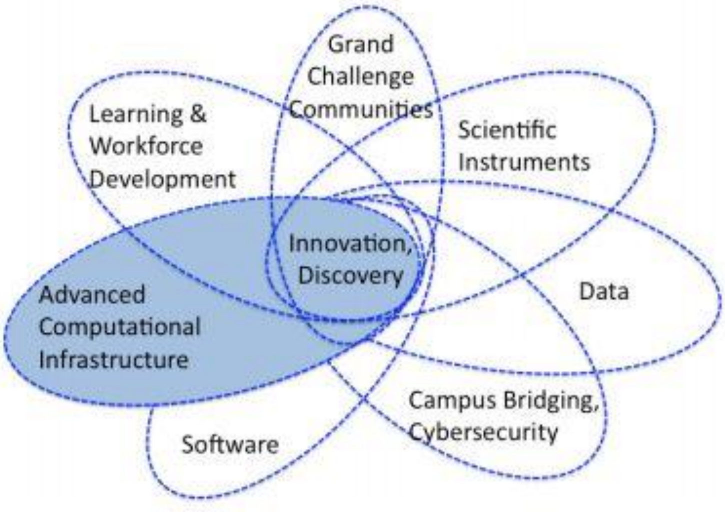

Grand Challenge Communities
Learning & Workforce Development
Scientific Instruments
Innovation, Discovery
Advanced Computational Infrastructure
Data
Campus Bridging, Cybersecurity
Software

FIGURE 2.1. CYBERINFRASTRUCTURE FRAMEWORK FOR THE 21ST CENTURY (NSF, 2012, S. 6)

2.4.2 Europe

In the European context the largest study on data sharing infrastructures was conducted by PARSE.insight from 2008, the results were presented in the report, “Insight into digital preservation of research output in Europe” (2010).They found that some of the reasons why data sharing was less attractive than OA was the lack of clarity about which data should be preserved and which not, and likewise the lack of a uniform European policy on digital preservation. Furthermore the fact that the profession of data curator does not exist and the responsibility for the long term preservation of data is unclear. The general conclusion was however that among different stakeholders there was agreement both about the reasons for preservation and the threats regarding digital preservation. The report does however focus more on long time preservation than making the data available. In addition the report contains recommendations, best practices and a roadmap. These results are taken further in the report from the European commission in 2010, “Riding the wave”. In this report, scenarios, challenges, a wish list and at last a vision for 2030 is presented. All of these offer clear guidelines on where the initiatives for creating a cyberinfrastructure and promoting e-science in Europe is heading.

2.4.3 Norway

In Norway a roadmap (Research Council of Norway, 2012a, Research Council of Norway, 2012b) was presented by the Research Council at the conference, “INFRAday 2012”. It was an updated version of the report “tools for research” (Forskningsrådet, 2008), and largely influenced by “Riding the Wave” (European Union, 2010). It is however in a style similar to the Strategy report by the European Strategy Forum on Research Infrastructures (2010). This report aims to make visible large and important research infrastructures crucial to help reach the research and political goals and to further be normative for long term planning of investments in research infrastructure. It will guide private and public funders of research infrastructure, by promoting projects that are planned, but still lack funding and be part of a tool used by the research council to make decisions on how coming announcements of funding for research infrastructure should be collected. It will also make visible the Norwegian participation in the international research infrastructures and show the balance and reaction between such participation and national investments. As such this roadmap is using existing projects as examples, rather than giving independent directions, and is by this quite different from “Riding the Wave”. It does however show where and how the Norwegian research council invests, or would like to invest by presenting various projects. The main focus for all the projects is investing in expensive equipment and establishing databases to be used in research; the longtime data preservation aspect is however absent. The funding is given for projects with limited time, each of them often generating large amounts of data without any long term preservation plan, such as the “bit stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years” listed in the wish list from “Riding the Wave”. The Norwegian road map seems somehow to be more focused on separate projects and research infrastructure in general, and does not intend to meet the list of goals from the “Riding the wave” report. Furthermore the involvement of libraries is limited to two language related projects in the field of Humanities where the National Library and the University library in Bergen are involved.

2.5 Examples of initiatives for e-science and the sharing of research data

In addition to this there are also a number of initiatives on data storage and data sharing with different levels of documentation. The most complete structure is the Australian National Data Service (ANDS). Their aim is “to build the Australian Research Data Commons: a cohesive collection of research resources from all research institutions, to make better use of Australia's research outputs” («Australian National Data Service»). ANDS does not store data, but assists in the building of institutional data repositories and then connects these together. In addition to being a portal, ANDS is aiming to create an Australian data commons, where commons means shared ground with common resources owned by the community rather than individuals. They further work for the promotion of a culture of data citation by the usage of doi and rewards as in the model illustrated below.

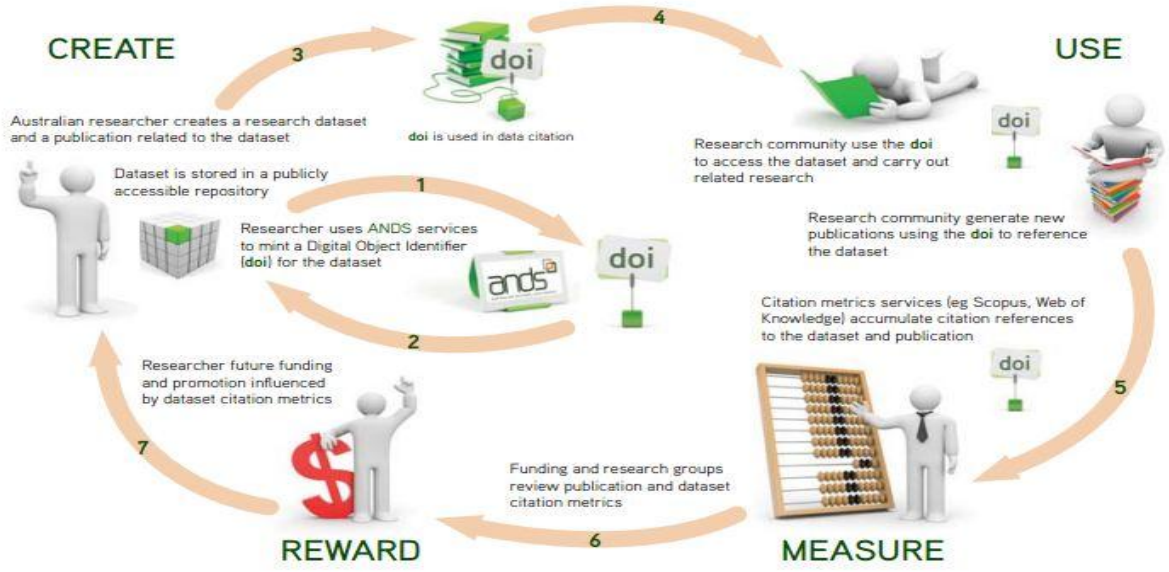

CREATE
3
doi
4
USE
Australian researcher creates a research dataset and a publication related to the dataset
doi is used in data citation
Research community use the doi to access the dataset and carry out related research
doi
Dataset is stored in a publicly accessible repository
1
Researcher uses ANDS services to mint a Digital Object Identifier (doi) for the dataset
ands
doi
Research community generate new publications using the doi to reference the dataset
2
Citation metrics services (eg Scopus, Web of Knowledge) accumulate citation references to the dataset and publication
doi
Researcher future funding and promotion influenced by dataset citation metrics
5
7
Funding and research groups review publication and dataset citation metrics
6
REWARD
MEASURE

FIGURE 2.2. BUILDING A CULTURE OF DATA CITATION («ANDS»)

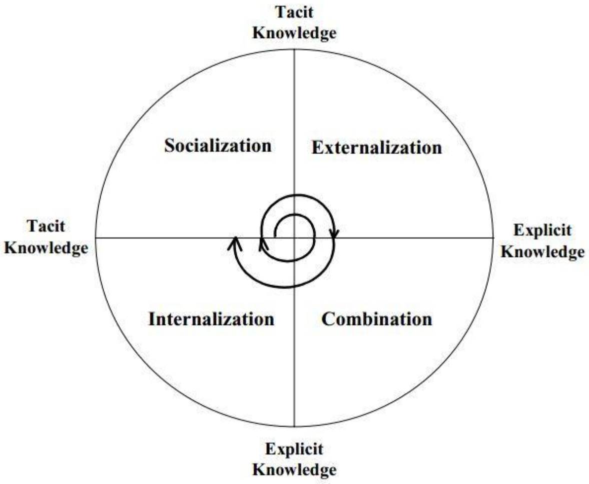
ANDS is so far the most complete national initiative for providing access to research data. Launched in 2008, it already has some years of experience in this much unploughed field, however they are still expanding and developing, so that data citation can become an integrated part of Australian research. In the US an increasing number of storage systems are being built, but these are primarily connected to the research field or the specific institution. The Johns Hopkins University, as the pioneer, started their first e-science project, Sloan digital sky survey, in 1992 and began leading in the development of citizen science with their Galaxy Zoo («Galaxy Zoo: Hubble», 2010). Other initiatives such as the e-science portal for New England libraries («e-Science Portal», 2010) function as a support for librarians working with e-science, by collecting relevant material for e-science librarianship and related areas. Also the Association for Research Libraries in the US is present in the e-science debate, by making an agenda (W. Lougee mfl., 2007),conducting studies of their member institutions (Soehner, Steeves, & Ward, 2010) and by responding to policies made by the National Science Foundation («Unpacking the NSF Requirement», 2010)

2.6 ARTICLES AND PRESENTATIONS FROM NORWAY

At Oslo University, Norway (UIO) a short study related to the OECD guidelines was conducted in 2009 (Grøttum, 2010). It showed that the researchers at UIO where generally open to the idea of sharing their data, but were worried about the amount of work that would be needed in order to connect the necessary metadata to their data. This survey did not include questions about the existing reuse of data and how data is located today. Also, it did not include in-depth questions about fears and threats related to open sharing of content and how the researchers would deal with metadata issues. No additional studies on research data sharing in Norway have been found. As a result of this there is little to no research literature in the field from Norway. What has been located from the academic library perspective is limited to a presentation by Randi E. Taxt, from the University of Bergen (Taxt, 2011) and a presentation by Almuth Gastinger from the University of Trondheim at the annual Norwegian Library meeting (2012). Both of them emphasize the importance of involving the academic libraries in the development of a cyberinfrastructure and e-science. This is in order to keep up with the development happening in the sciences and to offer expertise on significant topics in the field, such as the building and maintenance of metadata, ontologies and thesauruses for refining and proper storage facilities. This illustrates the need to conduct further studies in order to explore the potential of e-science in a Norwegian context, and create an open debate on e-science librarianship. On the annual Munin conference in Tromsø, 2011, Cameron Neylon spoke to Norwegian librarians and researchers about the future of e-science. Several of the attendees found his speech inspiring, but there is however no current debate on the issue. Furthermore the participation on the INFRA day, 2012, showed that researchers are dealing with cyberinfrastructure mostly without including the academic libraries in their work.

2.7 THE ROOTS OF THE LITERATURE ON E-SCIENCE LIBRARIANSHIP

The literature goes back more or less to the year 2000 with the articles by (Moore mfl., 2000a, 2000b)*Collection-Based Persistent Digital Archives 1 and 2*. In the book, *The Data Deluge* (Marcum & George, 2009), Joan K. Lippincott points out that the libraries did not pick up the terms e-science and cyberinfrastructure at once. “In the early years of the twenty-first century, many of those in the information technology, at large research universities, understood the notion of E-science and its implications for network infrastructure, development and distributed tools and massive storage, but only five academic librarians had heard of the terminology of e-science or cyberinfrastructure.” (Lippincott, 2009, s. 73). The literature on e-science and research data is building on principles from archiving, but also using knowledge management theories developed in the 90's to support the development of e-science. An example of this is the article, "Information ecology: Open system environments for data, memories, and knowing" (Baker & Bowker, 2007), where lines are drawn from the data management of e-science back to theories of information ecology by Davenport (1997) The model developed for illustrating the life cycle of knowledge by Nonaka (1994) is evident as the basis used for a model illustrating the connection between knowledge and data.



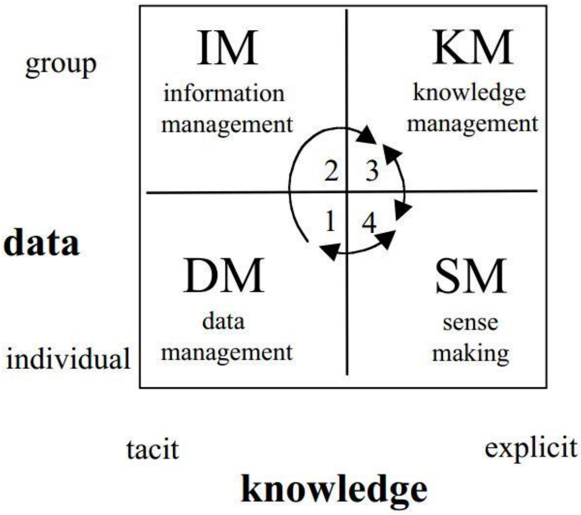


FIGURE 2.3. KNOWLEDGE CREATION BY NONAKA & TAKEUCHI 1995

FIGURE 2.4. DATA-KNOWLEDGE GRID BY BAKER & BOWKER 2007

As such e-science librarianship is a way of dealing with all levels of information, from data to knowledge, as valuable resources, both for further research, but also possibly as an economic value. The data-knowledge grid by Baker & Bowker (2007) combines what is often referred to as the knowledge life cycle (Borgman, 2007, s. 10) with the knowledge life cycle produced by Nonaka & Takeuchi. The data life cycle is a term used to describe the process from data generation trough to the preservation of the scholarly products that set that data in context. Borgman further requests more research on research data, both from primary to secondary sources and through the entire life cycle.

2.8 VISIONS FOR THE FUTURE OF E-RESEARCH

"Research data are the new special collections" states Christine Borgman (Borgman, 2011),an important advocator for e-science librarianship, arguing that the collection of research data is the responsibility of the research libraries. A researcher presenting his ideas of what the future of e-science will bring, is Cameron Neylon, in his presentation “The need to publish more and read less” at the annual Munin conference in Tromsø (Neylon, 2011b) There he argued that “In the modern world of research and innovation the traditional paper, the traditional research output on which we have based the scholarly enterprise for the last two or three hundred years, is not enough”. He sees the traditional article as outdated and advocates for publications of a more dynamic kind that fit the channel they are published in. He argues that the research process should be open, and research data published when collected. Then when referred to in presentations or articles, tweets or whatever other channels, the researcher finds suitable for communicating their findings. One of the main aims he presents for doing this is linking materials in a way that the publication becomes a “side effect of what you do anyway”. The core material to link from in such a chain is the research data. In his blog,’ Science in the Open’, Neylon deals both with issues of the more concrete kind such as data repositories (Neylon, 2011a) as well as the more overall debates for Open science, such as the blog post, “They. Just. Don’t. Get. It…” (Neylon, 2012). This connection illustrates how the advocacy for e-science and data sharing can be seen as a prolonging of the open access and open science advocacy, as suggested by the Microsoft research group “Those who focus on open access, far from being radical, are not being nearly radical enough.” This connection between the research data and the research publication in combination with ambitious visions for how these will be interconnected in the future makes it sometimes hard to distinguish what the subject is. Open research data in a virtual context invites the creation of stronger bonds between data and publication. At the same time the publications themselves are, as Neylon points out, changing both form and formats, as they now come as podcasts and videos in addition to various text formats. Much of the visions for the future of e-science do not speak explicitly of research data, but instead generalize about scholarly output, where the research data represents the earliest stage.

Clifford Lynch said that “E-science represents a significant change, or extension, to the conduct and practice of science; this article speculates about how the character of the scientific article is likely to change to support these changes in scholarly work.” (Lynch, 2007) and goes by this into the line of visionaries who are expecting big changes in the traditions of scholarly publication with the rise of e-science.

The book, The Data Deluge, from 2009 edited by Deanna (Marcum & George, 2009) presents both challenges and perspectives on how research libraries can deal with e-science. In the chapter “E-science and research libraries: an agenda for action” Wendy Pradt Lougee cites computer scientist, Hal Berghel’s vision of the future from 2001:

> *"By 2100, our current view of electronic publications as copyright-able artifacts will be viewed as a historical allegiance to a pre-participatory, non-interactive essentially dull and lifeless era of publishing - an era in which one thought of digital libraries…. as a collection of linked "things" rather than articulated processes and procedures. The current digital publication will become a relic, an obscure by-product from the horse and buggy age of digital networks" (W. P. Lougee, 2009)*

The vision Lougee presents is much similar to the one Neylon argues for, where the research data is one of "the linked things". According to Lougee, this change is already starting to take place, as new genres of interactive publishing are developed, by including "contributions or commentary from a distributed community of users". This change in academic publishing that is described to happen now and take place in the future, will affect the libraries as they facilitate storage and the possibilities to cite and link research data.

In some way they all have something in common with the old visions about the web, from the beginning of the 20th century, such as Vannevar Buch`s visions of two way links and Paul Otlet`s attempt to create a system of linking that would tell the reader whether the material linked to, agreed or disagreed with the document in question (Wright, 2007). By having an optimistic view on how technology can be used, it eases the workload of the researcher and connects material in new ways and helps make new discoveries.

2.9 METADATA ISSUES

Rolf Heuer, the director of CERN, emphasizes the importance of metadata in the statement "You need to preserve the data but also the knowledge. When talking about data preservation it is always data and metadata, that is the important point" (CERN, Le Diberder, Heuer, Mele, & Diaconu, 2009). The different value of data with or without good metadata is an important issue, both in terms of functionality in reuse but also in terms of trust. Heuer further states that 50% of the researchers in their field (High Energy Physics) could do better science by reusing data, but are worried that they are using something that is not documented enough". That way quality metadata is facilitating data integrity "the assurance that the data are whole or complete, consistent, and correct" is supported by Kowalczyk& Shankar (2011) which further expresses the fact that "without provenance data, objects cannot be considered preserved". Metadata issues can therefore be regarded as highly important in data preservation and e-science.

The future potential for rediscovery if good metadata is applied, recognized by a group from Microsoft research, writing about Scientific data management in the coming decade states the following, expressing both the vision and the aims to reach this:

> *"In an ideal world there would be powerful tools that make it easy to capture, organize, analyze, visualize, and publish data. These tools would do data mining and machine learning on the data, and would make it easy to script workflows that analyze the data. Good metadata for the inputs is essential to make these tools automatic. Preserving and augmenting this metadata as part of the processing (data lineage) will be a key benefit of the next- generation tools" (Gray et. al., 2005).*

This idea that the tools for good data mining are still under development, and that metadata must be made for functions yet not known, is supported by the group working with eSciDoc at the Max Planck Society: *"this agnostic approach to data and metadata has led to the development of an implementation of abstract content models, to enable us to deal with different and yet unknown types of data and structures"* (Dreyer mfl., 2007). This makes the challenge of creating high quality metadata schemas to use in data preservation a complex task. Not only shall metadata be created for purposes not yet thought of, but it must also be created in a way that little or no data loss can be expected in the processes of migration of the data between platforms.

There is no doubt that the creation of metadata can be an expensive process in the LIS community. The tradition is to manually create metadata or copy it from records that were once typed in manually. In the context of research data, this is not a functional method, and the need to create algorithms that automatically create most and possibly all the metadata will be necessary in such a context. "*Rather than manually typing metadata, the metadata could be extracted from existing files generated from the context upon creation."* (Harms mfl., 2011) This requires again as Richard E. Luce points out, a need to have a plan for the metadata before the research data is collected:

> *"In order to create tools that effectively search, categorize and give context to this data, highly enriched metadata is required to describe provenance, contents, formats, usage conditions, rights and so forth. Specifying these requirements up front, as well as creating taxonomies to support cross-disciplinary research, requires the skills and knowledge of information scientists in the broad sense" (Luce, 2009, s. 6).*

This point is supported by others, who also emphasize the need to develop a plan for data preservation before the data is collected as an implemented part of any research proposal (CERN mfl., 2009). If early planning of metadata collection is done and strategies to collect as much metadata as possible by automatic processes are implemented, the burden for the contributing scientist would be less. "T*he goals of automated context creation are to generate more accurate and consistent data, to create sufficient context for precise data discovery, and to ease the burden of creating metadata form the contributing scientist"* (Kowalczyk & Shankar, 2011). This would however, as Heuer points out, require that the strategies for metadata collection are included from an early stage, preferably in the research proposal. Even if the importance of metadata is repeatedly pointed out, no study investigating researchers' traditions for adding metadata to their research data has been discovered. Still (Kowalczyk & Shankar, 2011) cite that *"lack of metadata is one of the major barriers to data sharing because it affects both machine-readable processes and human searching"*. Once a discovery is made the issue of trust or credibility is largely depending on the documentation of the research process, this documentation should be found in the metadata following the research data. As the literature illustrates, the challenges regarding data preservation that relate to metadata are many: planning, automation, cost and trust being the concrete ones that are faced today while anticipation of the future challenges such as rediscovery in new systems with functionalities of the future is another. Migration between platforms not yet developed is another. (Duval, Hodgins, Sutton, & Weibel, 2002) state that *"there is usually a direct relationship between the cost of metadata creation and the benefit to the user: Describing each item is more expensive than describing collections or groups of items but clearly more useful for data discovery".* This statement does not seem however to be based on a study, but rather on general experience.

2.10 THE SKILLS OF A DATA CURATOR AND AN E-SCIENCE LIBRARY PROFESSIONAL

The literature shows several requests for a data curator or data archivist that has professional skills in the long time preservation of research data. *"There is a new job in our field that should be installed and it is what we call the data archivist position, it is a person that takes care of the preservation of data and makes sure that the system is running"* (CERN mfl., 2009).The report by PARSE.insight states that "*the profession of data curator is missing"* (Thaesis & van der Hoeven, 2010), also the report by the National Science board (Simberloff mfl., 2005) requests such a new professional to help with the curation of constantly growing amounts of research data.

Szigeti & Wheeler (2011) ask the question “*What role do librarians have to play in the growing field of e-science?”* and partly answers the question by stating that *“people are needed to handle, preserve, describe, and organize data, and, because many of these tasks are similar to what librarians have done with publications for centuries, it makes sense that librarians would have a role in the emerging task of managing scientific data”.* The role of the data curator seems though to be composed of more than a librarian’s qualifications as the archivists and the computer scientist are also needed in the data curator role. There is however “much potential for librarians to become more integrated in the science workflow. This includes working closely with scientists at all stages of the data lifecycle, as well as participating in the data literacy education of the next generation of scientists by helping coach science undergraduates and graduates on best practice relating to data” (Allard, 2012).In order to fulfill the potential schools of LIS, the US has started to put data management courses on their curriculum for both graduate and undergraduate students. Data management must be taught to academic librarians so that they can teach data literacy and preservation skills to the students. At the same time data management must be taught to LIS students so they have skills to manage the challenges of data curation in their profession. There are examples of data management being taught at Massachusetts Medical school (Piorun mfl., 2012)and the University of Illinois’ Urbana-Champaign («Scientific Data Management at UIUC») These are two examples. But the work of ARL to involve the libraries in research data management is definitely getting results. In their report from 2007, Lougee et. al. list the current library capability to support e-science in a list expressing what skills librarians should have and build on the work with e-science support.

Current library capability to support e-science

A case can be made that research libraries already have existing capacity and experience that they can bring to bear to support e-science. By virtue of their experience in the service and data management, and, for main, their mission, they are capable of advising and developing infrastructure to support the needs of scientists working in a cyberinfrastructure-enabled environment. For example, research libraries have:

Expert understanding of the policies and principles related to open exchange of scholarly information, as well as the roles that can be played by institutional repositories in assuring that exchange, and a demonstrated ability to offer and support both institutional repositories and domain repositories.

Experience with developing and supporting integration and interoperability tools for information distribution and discovery, both within and across disciplines is required.

> *Experience is also needed in developing and supporting both business and technical strategies for long-term archiving.*
>
> *Also, the understanding of archival and life-cycle aspects of scientific information, including the importance of assuring access and usability over long term." (W. Lougee et. al., 2007)*

Others have tried to fit these skills and the definition of LIS professionals with data curators. In "the Data Deluge", James L. Mullins defines library science as *"the systematic study and analysis of the sources, development, collection, organization, dissemination, evaluation, use and management of information in all its forms, including the channels (formal and informal) and technology used in its communication.*" This definition is from Joan M Reitz' Online Dictionary for Library and information Science (ODLIS). Mullins uses the definition to argue that the principles and theories of library science can be applied "first by analyzing and understanding the data sources created by the domain scientists, then by creating an organizational structure and retrieval point through the application of metadata and taxonomies/ontologies, which will facilitate the dissemination of the data, and finally, by establishing standards for the preservation and archiving of the data". Clearly many of the domains the data curator needs to deal with are covered by the LIS professional, and it might be that the profession will develop further in the direction of digital curation also in the academic libraries, and that the material dealt with will no longer will be limited to publication but also include material curated earlier in the research process.

2.11 Additional literature

As this thesis intends to study the behavior and attitudes of the researchers, literature from the social sciences has also been used. According to Borgman, Robert Mertons norms establishes a "neat and concise way to explain how science operates" (Borgman, 2007, s. 37). Borgman further argues that the norms he constructed can "still be used as templates to construct the tools and services of e-research" (Borgman, 2007, s. 37). In his essay "Behavior patterns of scientists", Robert K. Mertons aims to explain the behavior and competition between the scientists. Rivalry and pressure to publish quickly are some of the issues he mentions. He argues however that these issues are more common in the "hot subjects (Robert K. Merton, c1979, s. 331). His work is however, as Borgman argues, of importance to the field of e-science as it thoroughly describes the patterns and norms of the sciences before the appearance of advanced technology, and is relevant to this thesis as he formed norms and descriptions of how scientists worked together. Also literature from the field of Social studies of Science and technology (STS) is of interest when exploring e-science librarianship and the book by "E-research: transformation in scholarly practice" (Jankowski, 2009) that is used and referred to several times in this literature review comes from the STS field. Manuel Castells describes in his book, "The rise of the network society" (Castells, 2010), how technology has had and will continue to influence the society and his theories on how the network functions can also be applied to the sciences and the emergence of e-science. When Castells describes "the development of horizontal networks of interactive communication that connect the local and global environments, as a multichannel system of digital communication that integrates all forms of media" (Castells, 2010, s. 23), he does not speak specifically of scientific works, but still we find that the changes he is describing are also happening here and can be applied to the vision, that researchers such as Cameron Neylon hold of the future of scholarly publications.

2.12 Conclusion

The literature shows several approaches to e-science, e-science librarianship and research data management. There are, however, many issues still to be discovered, already from the definitions it is clear how the matter is still being explored and established, this forming of the subject makes it of current interest to do further investigation. As described, much of the literature in the field is visionary, or based on case studies or best practices and new publications are coming out month by month, making it hard to end a literature review. While much has been written on e-science since 2000, the literature on e-science librarianship has been growing much in the last two years, indicating that e-science is becoming a more visible and integrated part of the activity in academic libraries. These libraries have been working with open access publishing for some time, and e-science librarianship seems to be the way to continue making the academic library the core of any learning institution. As Clifford Lynch described it:

Investments in infrastructures for research data are being made; it is however the library that has to invite themselves and their competencies in the field to join the party and offer their expertise on issues such as metadata and knowledge retrieval.

"Notable and important examples include the movements towards open access to scientific literature; movements towards open access to underlying scientific data; demands (particularly in the face of some recent high-profile cases of scientific fraud and misconduct) for greater accountability and auditability of science through structures and practices that facilitate the verification, reproducibility and re-analysis of scientific results; and efforts to improve the collective societal return on investment in scientific research through a recognition of the lasting value of much scientific data and the way that the investment it represents can be amplified by disclosure, curation and facilitation of reuse." (Lynch, 2007).

2.13 WHAT THIS STUDY WILL ADD TO THE FIELD OF EXISTING LITERATURE

In Norway the possibilities of e-science have yet to be thoroughly explored. The focus is much on the technical infrastructure and less on long term preservation along with access. In the development of grid stores held by external partners, the academic libraries are not present and the link between facilitating access to research data and the ideals of open science are not in focus. As the literature shows this is quite different in the US where research libraries and ALA are making the e-science librarianship visible and data curation is starting to be included in the LIS curriculums. This study aims to support those who try to break ground for e-science librarianship in Norway by identifying and describing attitudes and needs among the scientists who are gradually forced into a world where much of both collaboration and data collection is done electronically. The academic library has through time been a support function for the researchers. E-science is forcing the libraries to include data so that links between research output at different stages can be made. Data, information and knowledge is often presented in a hierarchical way (Dixon, 2011), there are however disagreements about whether data or raw data should be included as a responsibility of the library. Unlike Dixon, and with support from e-science librarianship, I argue that the libraries should aim to cover the whole pyramid.

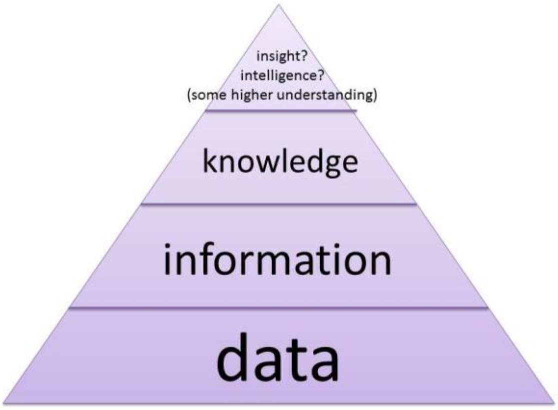


FIGURE 2.5. THE KNOWLDEGE PYRAMIDE

This view has still not gained acceptance in the Norwegian library community where many resources have been invested in the establishing of open access publication channels, but so far the library involvement does not extend the published material. As the literature points out librarians needs to work more closely with the researchers in order to assist in the publication and preservation of research data. Norway is not likely to be an exception, but a study of a Norwegian University will help with not only which issues Norwegian researchers are struggling with, but also help to put the e-science involvement of academic libraries on the map here.

3. METHODOLOGY

In this chapter the research approach and selected method are presented and argued for. Further the sampling strategy, data collection method and data analysis method are presented. Finally ethical considerations and trustworthiness of the study are discussed.

3.1 RESEARCH METHOD AND APPROACH

This research aims to collect data that describes and explains researchers' attitudes towards sharing research data and at the same time uncover which routines the researchers already have for storing and sharing research data. As far as the research method is concerned, the survey was selected and a mixed method questionnaire was used for data collection. This research does not aim to falsify or support a hypothesis, but rather to conduct a broad study in order to explore the attitudes among the researchers and the experiences they have with data sharing. The descriptive survey is therefore found suitable. Pickard (2007) describes the purpose of a descriptive survey; "to describe a situation and/or look for trends and patterns within the sample group that can be generalized to the defined population of the study." According to Babbie "surveys are excellent vehicles for measuring attitudes and orientations in a large population" (2007, s. 244). According to Pickard, the data collected in descriptive studies "does not lend itself to more sophisticated statistical analysis, indeed this is not its purpose."(Pickard, 2007, s. 96). As later described, the data analysis is not based on pure quantitative methods, but rather on a qualitative approach to mixed data. Pickard further presents the descriptive study as being suitable to describe the current situation in a given environment. As this research aims to investigate both the attitudes and the more practical side of how researchers deal with the data they collect, the survey is found to be the most appropriate method. It is difficult to give a proper definition of the survey, as a range of studies are labeled with this method, it can however be described by some typical features:

> The use of a fixed, quantitative design;
>
> The collection of a small amount of data in standardized form from a relatively large number of individuals;
>
> The selection of representative samples of individuals from known populations" (Robson, 2002, s. 230)

Pickard explains the survey as "the research method used to structure the collection and analysis of standardized information from a defined population using a representative sample of that population" (Pickard, 2007, s. 95). The survey used in this study is descriptive as it tries to "explain a situation and/or look for trends and patterns within the sample group that can be generalized to the defined population of the study" (Pickard, 2007, s. 96). The process of conducting a survey is as described by Pickard as a fairly linear process from the identification of a general topic area, through the establishment of aims and objectives, identification and selection of a sample population, selection of a data collection method, design and data collection to data analysis, described by Pickard to be the most challenging part of making a descriptive survey, before the final presentation of the results.

3.2 Sampling strategy

The Norwegian University of Life sciences at Ås (UMB) has been selected for the survey as it is a university focused mainly on the natural sciences, and therefore a university which is likely to produce much data that fits to the OECD description of "sharable data". The group of researchers behind the Microsoft research report(2005) claim that "It is in the natural sciences where the 2020 Group argue the greatest impact of computer science will be felt." It is therefore also likely that computer science based possibilities such as the sharing of research data along with an increased production of research data often referred to as the data deluge, will have a more instant and evident impact on the Natural Sciences than the Social Sciences.

UMB is however a university combining both the hard and the soft sciences. Having institutions with both social science approaches like the Department of International Environment and Development Studies (NORAGRIC) and at the same time several departments that focus on the applied sciences like the Department of Mathematical Sciences and Technology (IMT), there are a total of 8 departments at the University. In addition there are several research centers connected to the university such as the Centre for Integrative Genetics (CIGENE) and an Aquaculture Protein Centre. UMB should therefore provide a sample population that shows the differences between different disciplines and at the same time data from disciplines that are likely to be strongly affected by the e-science development.

In Norway the Universities of Bergen, Trondheim, Oslo and Tromsø represent the larger Universities. Over the last 10 years, several college universities have however gained the status of university and UMB is one of these. UMB became a university in 2005, but has institutional roots dating back to 1859 when it was formed as The Agricultural College University of Norway. The school has currently about 3800 students. The total number of staff is 1120, but only 650 of these hold scientific positions. These 650 researchers and PhD students are the sample population for this study. During the next few years the university will grow as a result of a merger with the Veterinary Sciences school, adding another 1000 or so to the current number of students. In combination with the merger several organizational changes are also taking place, it is therefore an institution in change, growth and development.

3.3 Data collection method

A mixed method questionnaire has been selected as the appropriate method for data collection. According to Ringdahl (2007, s. 181) much attention must be given to constructing questions that aim to measure attitudes in order to avoid leading and multidimensional questions and at the same time measure the strength of opinion. Robson refers to the challenge of designing unambiguous questions, an issue of "internal validity" (Robson, 2002, s. 231). The questionnaire used is largely based on the survey done by PARSE.insight at CERN, Humbolt University and ESA in 2009. As a result of this reuse, the problem of internal validity is to some extent avoided, as the questions have already been tested. The results from the PARSE.insight study formed the background for the Insight report (Thaesis & van der Hoeven, 2010). Some questions are however taken from another questionnaire done at the University of Oslo in 2009 by Per Grøttum (2010) investigating the researchers attitudes in regards to the OECD guidelines for research data (Gurria, 2007). This survey was conducted in Norwegian, but a translation of the questions was done, creating the risk that some meaning was lost or changed in translation. In addition to the questions from the two previous surveys some additional open ended questions were added for the respondents to express, in own words, some of their attitudes and concerns regarding data sharing.

"The quantitative element of a questionnaire will serve purposes such as looking for a potential generative mechanism by seeking promising patterns between measurable variables, whilst at the same time acknowledging the difficulties in finding measurable indicators of complex social phenomena" (Greener, 2011, s. 40). By combining the quantitative and the qualitative approach, attitudes and opinions for the researchers are collected in measurable form and analyzed with a more qualitative approach aiming to understand and explore other sides of the issue.

The questionnaire is designed to discover how the researchers use, find, share and store data and the attitudes they hold towards data sharing. Furthermore the different questions can be grouped into different categories, together providing a broad understanding of how the researcher sees and deals with digital research data. Together these categories aim to give a complete picture of the researhcers and their relation to research data as illustrated below.

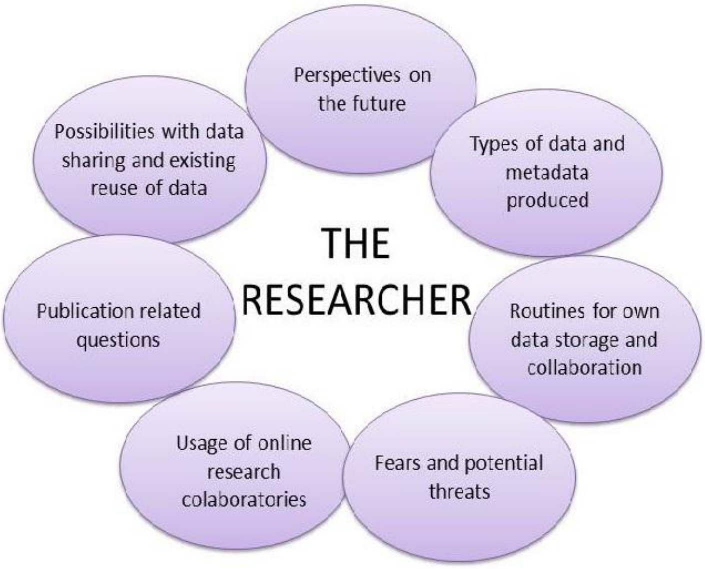

THE RESEARCHER
Perspectives on the future
Types of data and metadata produced
Routines for own data storage and collaboration
Fears and potential threats
Usage of online research colaboratories
Publication related questions
Possibilities with data sharing and existing reuse of data

FIGURE 3.1. THE DIFFERENT ASSPECTS COVERED BY THE QUESTIONAIRE.

The questionnaire contained 34 questions and was estimated to take about 15 minutes to answer. After first introducing the research to a group of representatives from the different the departments at the university, the questionnaire was sent out to researchers and PhD students at UMB, a total number of 650 people. 23% of the sample population answered the questionnaire during the one month it was active. According to Babbie and Pickard, improving the response rate is one of the main challenges when conducting surveys, and so is deciding on which response rate gives sufficient material for drawing conclusions. Babbie refers to a review of published social research literature suggesting that a response rate of 50% is considered adequate for analysis and reporting (Babbie, 2007, s. 262). According to Greener, the important thing is not the percentage of the respondents but rather “to think about what kind of a case they need to make in order to make that claim” (2011, s. 51).The 23% respondent rate in this study gave enough respondents to combine questions in order to detect patterns in the way the respondents reply. It does not however provide data strong enough to make generalizations; the findings will therefore not aim to generalize but rather use a qualitative approach to indicate trends and patterns in the group of respondents.

3.4 Data analysis

The answers gathered were analyzed in two different ways in order to both understand the answers themselves, and see how the responses from UMB are compared to previous responses to the same questions. Together these two analyses should give a comprehensive picture both of what is the more general opinion towards the sharing of research data among researchers and what is particular to the Norwegian University of life sciences.

Firstly the questionnaire was analyzed according to different variables such as experiences in research and institutions the researchers were connected to. These variables were selected based on a preliminary combination of questions done early in the process of analysis where indications were found that these two factors had an impact on how the respondents replied. Some questions were again combined with other questions, in order to identify if other variables had an impact on the responses. One example was to see if the reasons for sharing of research data given by those who already make their data openly available were different from the average opinion. By doing such comparisons it will uncover which variables had an impact on the respondent's opinion. The identified variables were grouped, placed in relation to and discussed according to the facts they are found to impact. As the facts were analyzed according to relevant variables key findings were identified. In this way both the individual story of each question was explored and discussed, before the key factors of the findings were extracted and discussed in the bigger picture, with the research question in mind. Ian Greener describes the challenge of analyzing survey data as "how to both examine each question`s responses and to tell a coherent story about what the results as a whole seem to add up to" (Greener, 2011, s. 49). In this thesis space is given to both so that the process of extracting the essential findings presented in the discussion can be traced back to the data presented in the results.

Furthermore the results from the PARSE.insight study and the study by Per Grøttum were used as a benchmark when analyzing the results. The PARSE.insight study gathered material from three different institutions in different European countries. They were the most prestigious in their field, while UMB is a small and fairly new university. Other differences might be explained with the years between the studies. The PARSE.insight study was conducted between 2008 and 2009 and this difference in time might affect the results. For the questions taken from the Grøttum study, a comparison was made with the responses there to detect possible differences in the responses at UIO and UMB.

Findings from the different analysis will then be discussed in order to answer the question of what the attitudes are towards sharing and reuse of scientific/research data among researchers and what experiences the researcher at UMB have had with sharing and reuse of research data.

3.5 Ethical considerations

Ethical considerations must be made in all research, in social research ethical considerations are often related to anonymity and privacy issues. Data identifying an informant cannot be shared without permission from the person. In order to protect the privacy of the respondents on the questionnaire, the email addresses used for inviting researchers to answer the questionnaire have been made anonymous by the university. Further anonymity functions in the software were applied to make sure that it was not possible to trace the respondents from their email addresses. There is always a possibility that a combination of several questions can reveal the identity of the respondents. When quotations from the respondents are used they are therefore only identified by department, not years of experience.

3.6 Trustworthiness of the enquiry

In order to define the trustworthiness of a study certain criteria must be taken into account, according to Pickards "trustworthiness model", credibility, transferability, dependability and conformability are the different criteria that must be considered.

The four criteria are respected by taking the following measures into account:

Comparisons with results from other studies strengthen the credibility as well as the transferability by identifying which attitudes and traditions tend to be; institution specific. Further a rich description of the selected institution and population of respondents strengthens the transferability. The dependability is ensured by a detailed description of each step in the research process and how the research method has been appropriately followed. In addition to this information about the research process, the results from the survey are described and shared in a blog at sharingandaccess.wordpress.com. The blog is again shared with the informants as this ensures transparency in the research process. To ensure conformability much of the results are presented and visualized in figures. This ensures openness in the process and by this also conformity

4. Results from the questionnaire

Of the 650 researchers and PhD students at the Norwegian University of Life Sciences (UMB) selected as a sample population for the questionnaire, a total of 147 respondents, or 23%, replied during the month that the questionnaire was out. Prior to the questionnaire a presentation on the subject, introducing the study and the upcoming questionnaire, was held for a group of researchers at UMB including representatives from each department, and questions regarding the study were answered. The representatives from the different departments were then asked to encourage their fellow researchers to take time to answer the questionnaire. In addition to the first invitation, a reminder was sent out after 2 weeks, and again after 2 weeks a note thanking people for responding, announcing that the questionnaire would only be active for five more days was sent. During the first week 70 people responded, after the second notice another 70 responded and the day the thank you note was sent out, another 7 people replied to the questionnaire. Several out of office notes came as replies to the invitations, however only 5 participants were out of the office during the entire period. Two of these claimed to have read their email, leaving only 3 that possibly were not reached at all.

The questions where grouped into the following broad categories:

- Types of data and metadata produced
- Routines for own data storage and collaboration
- Usage of data collected by others
- Fears and potential threats
- Possibilities with data sharing
- Perspectives on the future
- Publication related questions
- Usage of online research collaboratories

Based on this categorization two main focuses are given to:

1. Types of data and metadata produced, routines for own data storage and collaboration, usage of data collected by others
2. Perspectives on the future and fears and potential threats and possibilities with data sharing.

The results regarding publication related questions and usage of online research collaboratories are presented in a following chapter called, “Other Results”.

The division and selection is based on the research question aiming to uncover the attitudes towards and experiences with the sharing and reuse of scientific/research data among researchers. 1 is focusing on practical issues, experiences with sharing and metadata while 2 is focusing on the attitudes towards sharing and reuse. In order to put these answers in context, space is also given to results regarding publication and usage of online research collaboratories. Further comparisons with results from the PARSE.insight study and the study done at UIO, in 2009, will help to put the results from this survey into a context of previous studies.

4.1 Variables

In the first question the respondents were asked to indicate which department they are connected to. The list of departments was taken from the UMB English web page and listed a total of 15 research centers and departments. There were few or no respondents from the different centers, but the departments where all represented with between 10 and 32 responses. The sum of respondents connected to each department is higher than the total number of respondents as some researchers are connected to more than one department. This gives a perecentage of 106,6% when summing the responses from the different departments. In the analysis of the results, this double connection will lead to some responses being measured twice when comparison between departments is done, in particular the respondents connected to the centers that have a double connection. The research centers are however excluded in comparison with disciplines or departments as they have a low number of staff. By doing this, counting the same respondents twice, is also avoided. It is however worth mentioning that the research center, Cigene, with 5 respondents, often has responses that differ more significantly from the average than any of the departments. Due to the low number of staff, and overlap in staff with different departments, they are however not listed in the comparison between the departments. When the data is combined with information from the university on number of staff connected to each department we see that the respondences in each department goes from 10,6% in IMT to 55% in ILP. Most of the departments have however, a respondence rate close to the average of 23%.

Table 1. . Question 1: Which department/center are you connected to? And number of staff from the University

Department/Center	Short name	Nr of staff	Respondents	Percentage based on nr of staff on dep.	Percentage of total
Dept. of Animal and Aquacultural Sciences	IHA	91	23	25,3%	15,6%
Dept. of Chemistry, Biotechnology and Food Science	IKBM	98	14	14,3%	9,5%
Dept. of Ecology and Natural Resource Management	INA	87	19	21,9%	12,9%
Dept. of Economics and Resource Management	IØR	52	15	28,9%	10,2%
Dept. of Mathematical Sciences and Technology	IMT	113	12	10,6%	8,2%
Dept. of Landscape Architecture and Spatial Planning	ILP	58	32	55,2%	21,8%
Dept. of Plant and Environmental Sciences	IPM	107	21	19,6%	14,3%
Dept. of International Environment and Development Studies	Noragric	43	10	23,3%	6,8%
Aquaculture Protein Centre	APC	1	1	100%	0,7%
Animal Production Experimental Centre	SHF				
Centre for Plant Research in Controlled Climate	SKP				
Centre for Continuing Education	SEVU				
Centre for Land Tenure Studies			2		1,4%
Centre for Integrative Genetics	Cigene		5		3,4%
Norwegian Centre for Bioenergy Research					
Other, please specify (2 Nofima, 1 LMD+INA)			3		2%
Sum		650	157	23%	106,8%

The majority of the respondents have been involved in research for more than 20 years, second came the new researchers with less than five years of experience, most of them being the PhD candidates that were also invited to answer the questionnaire. The total number of PhD students invited where 163. Of these, 38 or 23% replied, giving a representative number of PhD students among the respondents. Data is not available on how many years the other researchers at UMB have been doing research. The data on the PhD students does however indicate that it is likely to be a more or less balanced distribution of respondents according to years of experience.

Table 2. Question 2: How many years have you been involved in research?

Yearsinvolved in research	Respndents	Percentage
< 5 years	39	26,5%
5-10 years	22	15%
10-20 years	36	24,5%
20+ years	49	33,3%
Totalt		100%

4.2 PRACTICAL ISSUES REGARDING STORAGE, SHARING AND REUSE OF RESEARCH DATA

In this section the results from the questions regarding metadata, routines for storage and usage of data gathered by others as supplementary data, are presented and discussed. This helps to understand the data cycle and how it can be improved to facilitate sharing not only among a network of fellow researchers but in a larger community no longer depending on personal relations, but on a common interest in research. Not only have the researchers been asked about how they would be willing to share their data, they were also asked about how they deal with metadata. This included questions about whether they used formalized metadata standards, if they apply administrative and or technical metadata, or if they apply no metadata to their research data at all. Metadata is the key for creating context and storing knowledge about the research data in a larger system with the perspective of long time storage (CERN mfl., 2009). It is therefore interesting to see how the researchers regard and deal with these issues. The issues of metadata are put into context by investigating what kind of research data the researchers produce, and how they store their data.

4.2.1 TYPES OF DATA PRODUCED

In the OSEC standard, relevant research data is defined as "factual records (numeric scores, textual records, images, sounds) used as the primary source for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings" (Gurria, 2007). One third of the respondents stated that they produce this kind of data very often, or always in their research. In the institutes, IHA and INA, 77 to 80% of the respondents replied that they produce this kind of data very often or always, while in IØR only 27% of the respondents replied very often or always. This illustrates that there is a difference between the disciplines in how often they produce the kind of research data covered by the OECD definition, however an average of 2/3 illustrates that most of the researchers at UMB produced research data covered by the OECD definition often or always. It would therefore be relevant for UMB to study the OECD guidelines and how these could be implemented at the university.

When asked to further specify which formats the data is produced in, standard office documents were most common with 82% of the respondents producing research data in this format. A further 55% produced images in various formats (JPG, JPG2000, TIF, PNG, GIF, etc.), 49% produced PDFs and 42% produced device specific raw data. Other formats used were scientific and statistical formats such as SPSS, FITS, GIS etc., at 35% and plain text in various encodings, 33%. There are variations between the different departments on which formats the researchers use in the departments and IKBM, INA and ILP, raw data and/or scientific and statistical data is used by more than 50% of the respondents. In IMT software applications source codes are used by about 50% of the respondents, but as an average of all departments only 19% use these formats. This tells us that some formats such as office documents, images and PDF's are used by most of the researchers, so is device specific raw data of different kinds. Other formats are generally more discipline specific.

4.2.2 Routines for own data storage and collaboration

The most common arena for the sharing of research data is the research group and colleagues in research collaboration; about half of the respondents make their research data available in this way. Additionally 28% do not share their digital research data but would like to do so in the future; many of these respondents are new to research with less than 5 years of experience. In this group 50% of the respondents replied that they would like to share their research data in the future. 24% stated that access to the data of their current research is temporarily restricted and 23% of the respondents that their data is available for everyone. As the figures show, the attitudes towards the sharing of research data differ according to years of experience. The numbers below are given in percentages.

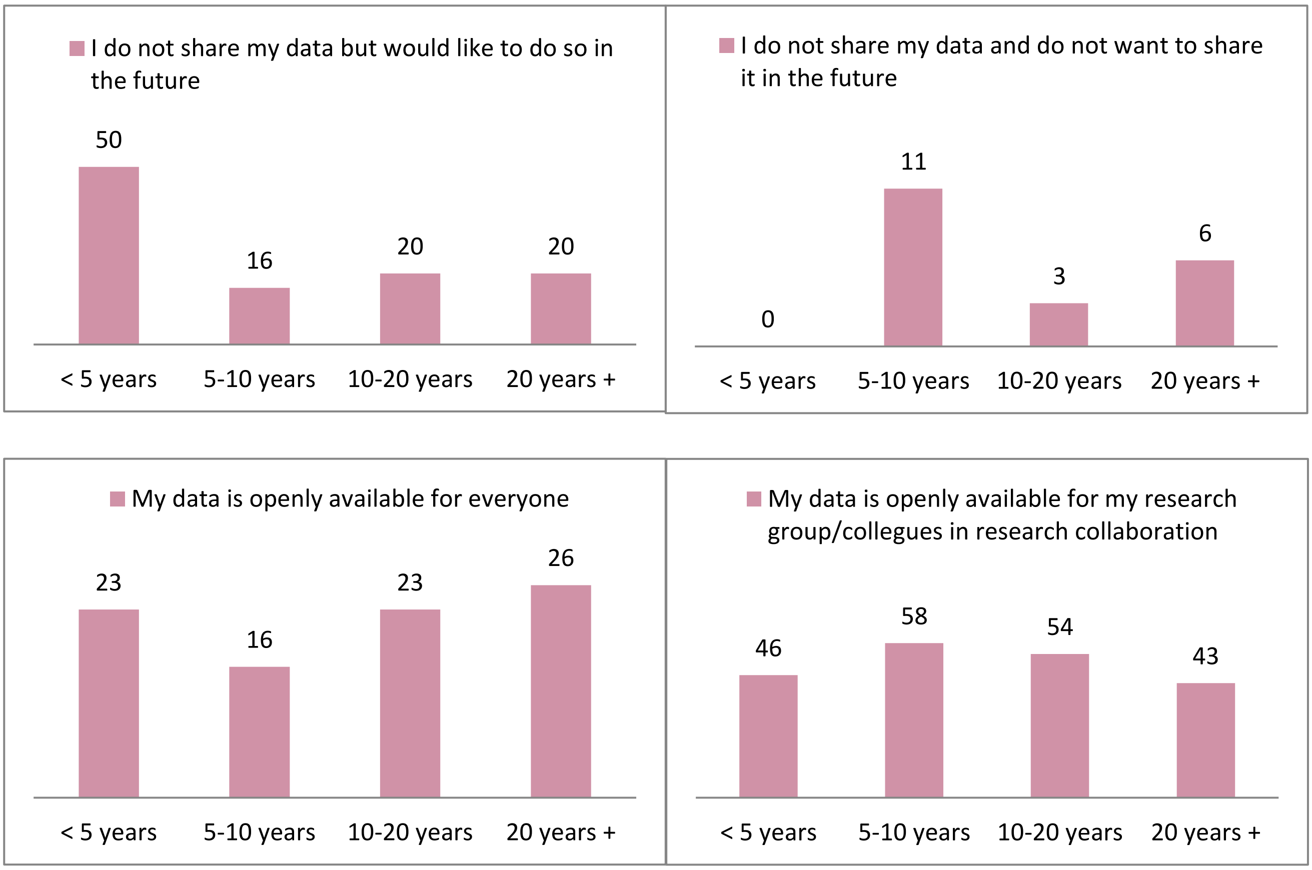


FIGURE 4.1: QUESTION 12 - WHICH OF THE FOLLOWING APPLIES TO THE DIGITAL RESEARCH DATA OF YOUR CURRENT RESEARCH?

Figure 4.1 indicates that there are clear differences in the attitudes and routines for data sharing among researchers, depending on where they stand in their careers, or how many years of experience they have as researchers. It is however not possible to give a firm explanation of why such a pattern exists. One explanation might be that ambitions change depending on where the researchers stand in their careers. The PhD students, being less experienced, are more open towards exploring new possibilities, while the researchers in the next phase become more focused on establishing a name and reputation for themselves, and as a result of this, the fear of what might happen if someone else accesses your data is stronger, and so the will to provide open access is less. Later on in your career getting established and becoming known becomes less important, and after being a researcher for more than 20 years the researchers no longer fear that others might "steal" from them, but rather hope that what they have done might be of use to others. It is however impossible to know if these are the reasons for these differences without investigating the issue further. These results show however that there are differences in how the researchers responding to this survey deal with their data, and what attitudes they have towards sharing their own data, and further that these differences depend on how many years of experience they have.

4.2.3 FAMILIARITY WITH METADATA AND STANDARDS FOR DIGITAL PRESERVATION

When it comes to familiarity with standards or guidelines used in digital preservation; one researcher from each institute seems to know something about standards while the remaining 90% of the researchers are not familiar with preservation standards. Of the 3 % who selected "other", one person from Cigene has listed a long list of standards (CellML, MIRIAM, SED_ML, KISAO, RICORDO) the rest of the respondents who used this option seem not to have understood the question, replying "scientific journals and databases" and "UMB has storage attached to Cristin[1]" or they simply do not seem to know. Responses that clearly indicate that the question has not been understood and that the percentage for none of the above, in reality, is even higher.

According to figure 4.2 it is evident that teaching researchers about metadata standards and the advantages of using such standards for optimized data exchange is necessary. This would help to prevent misinterpretation of data, as a result of data loss or a lack of metadata. The competence of researchers in using metadata is extremely low. There are however some persons with key qualification on this topic at the university.

[1] Cristin; Current Research Information System in Norway, "The Research Information System CRIStin is a tool for researchers and research institutions in Norway. It is aimed to record and promote the publication data, projects, units and competency profiles. The system is also used for reporting of publication points."

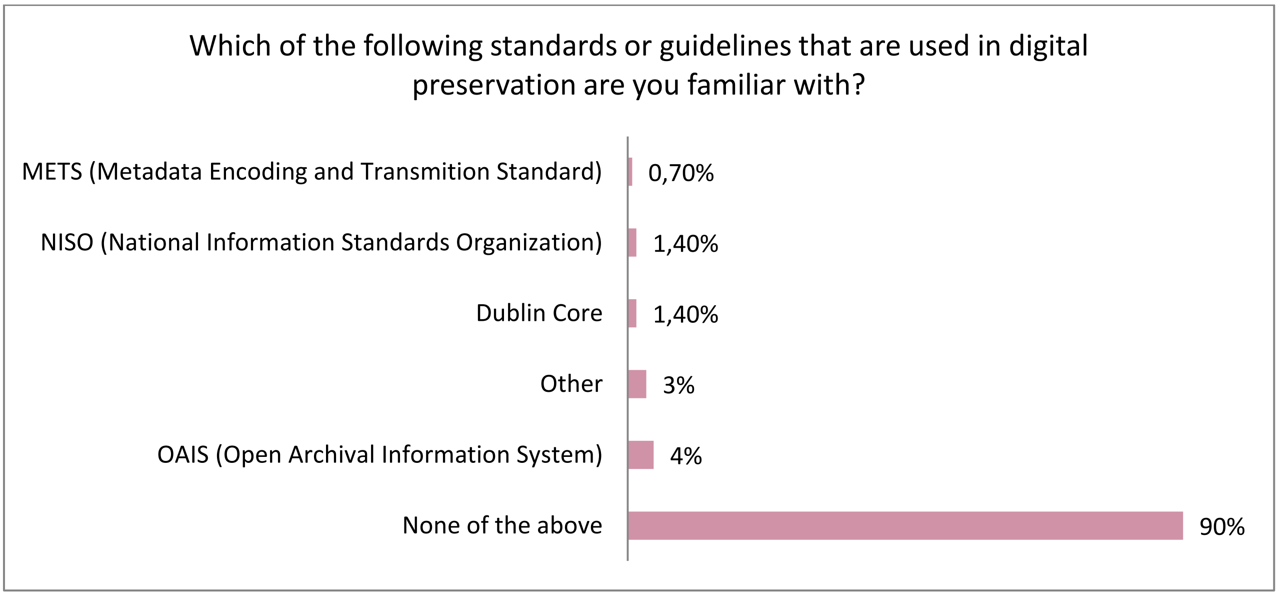


FIGURE 4.2: QUESTION 9

- WHICH OF THE FOLLOWING STANDARDS OR GUIDELINES THAT ARE USED IN DIGITAL PRESERVATION ARE YOU FAMILIAR WITH?

The researchers were asked in an open question whether existing standards are important in order to share their research data. Almost half of the respondents did not think so. None of them used the possibility to explain why. Others again, did not know and most of the respondents skipped this question. There are however also respondents who expressed and explained the importance of having standards for sharing research data in their field, which tells us that those who find standards to be important are also those who have more experience with data sharing.

> *"Yes, I primarily work with computer simulation, and open standards facilitates the reuse, refinement and critical examination of published models and results" (Cigene)*

> *"Some of the data I publish will not be accepted if they do not fulfill existing standards for what is required to be included. Therefore, yes. Existing standards makes the publication process more trustworthy" (IKBM)*

These comments were made by researchers form Cigene and IKBM who already make their research data openly available to everyone. And based on how the respondents replied, 'for and against', combined with other questions identifying the respondents background and experience, it seems that researchers with more experience on data sharing have a greater understanding of the importance of quality metadata and standards.

Another question asked about their routines for assigning metadata to their digital research data is reported in figure 4.3. Half of the respondents replied that they do not assign any additional information (metadata) to their digital research data. 35% replied that they assign administrative information and 30% assign technical information. Between the two there is an overlap of 16% assigning both administrative and technical information. In the additional comments some of them show a lack of experience with the matter such as the comment, "Do not yet have enough experience to answer this question". The highest results for assigning metadata are the researchers with 5 to 10 years of experience where 43% of the respondents assign administrative metadata, 38% assign technical information and only 38% do not assign any metadata to their digital research data, as illustrated in figure 4.4. The other respondents show only smaller differences from the average when sorted by years of experience. The department that the researchers are connected to, has however, a larger influence on their routines for assigning metadata, as illustrated in figure 4.5. In the departments, IKBM and ILP, 70% to 73% of the respondents assign additional metadata, while in INA and in Noragric, 37% to 40% of the respondents assign such information to their digital research data. This has possibly to do with the types of data used in the different disciplines. Researchers at IKBM are more likely to primarily use numeric data sets, while Noragric, the social science department collect data more in forms of interviews and questionnaires. In the latter some additional information about context can normally be extracted from the raw data, for numeric datasets it is more crucial to store additional information about where they come from and when and why. Further, traditions for sharing within some disciplines are likely to have an effect on these large differences. When combining the question with another question in trying to detect the researchers' traditions for the sharing of research data it becomes evident that those who make their data openly available to everyone or to their research disciplines are better at assigning metadata, here only about 25% do not assign administrative or technical metadata.

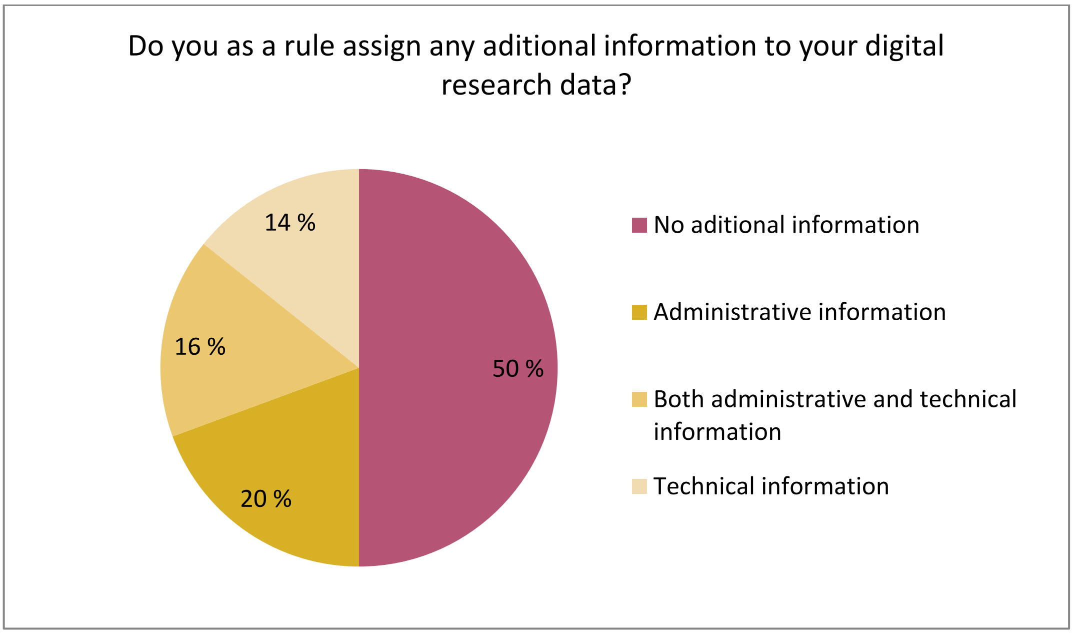


FIGURE 4.3: QUESTION 10

- DO YOU AS A RULE ASSIGN ANY ADDITIONAL INFORMATION (METADATA) TO YOUR DIGITAL RESEARCH DATA?

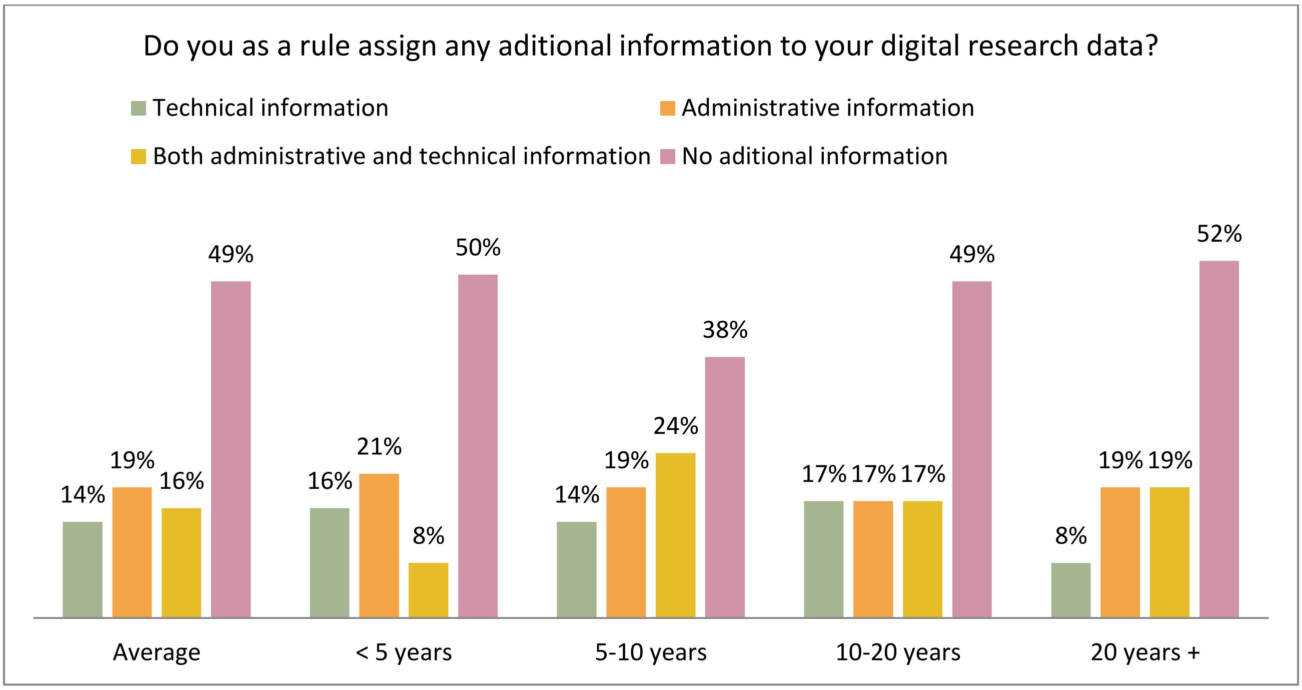


FIGURE 4.4: QUESTION 10 (ACCORDING TO YEARS OF EXPEREINCE)

DO YOU AS A RULE ASSIGN ANY ADDITIONAL INFORMATION (METADATA) TO YOUR DIGITAL RESEARCH DATA?

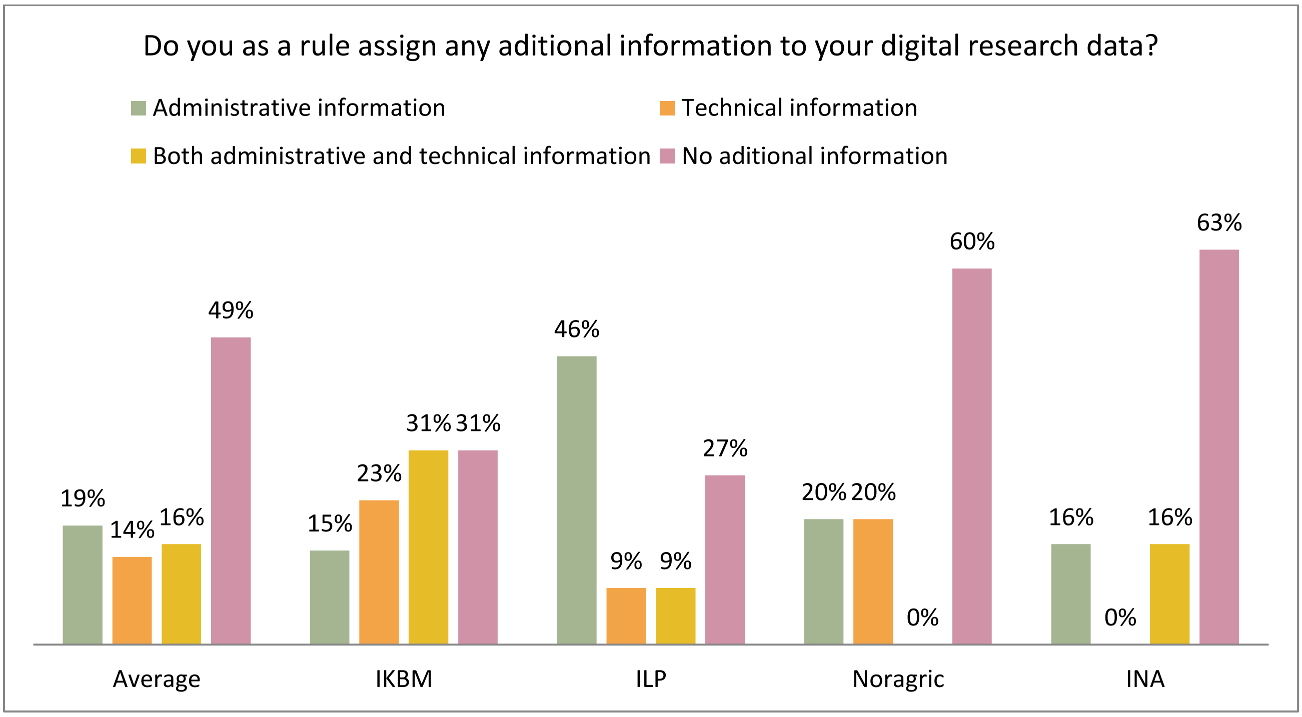


FIGURE 4.5: QUESTION 10 (ACCORDING TO DEPARTMENT)

- DO YOU AS A RULE ASSIGN ANY ADDITIONAL INFORMATION (METADATA) TO YOUR DIGITAL RESEARCH DATA?

As half of the respondents do not assign any additional metadata to their research data, this data will not be possible to interpret correctly in the future without assistance from a researcher who was involved in the data collection. In this way results are lost even if the data is stored. There is however a trend among the researchers with 5 to 10 years of experience to add metadata. This routine can be explained with these researchers being "digital native researchers" as they started with research after the millennium; they are not likely to be familiar with an "analog" way of doing research. As a result of this they also adapt their personal systems for storing and retrieval to the digital context, where metadata is necessary. Naturally the tradition for assigning metadata is stronger among researchers who openly share their data today. A further differentiation is identified between the disciplines that might be explained by different disciplines using and collecting different types of data.

The researchers were also asked about what was required of them when submitting data to external facilities. 40% responded that they needed to comply with a standard format, 22% that they have to transfer copyrights and 16% that they needed to supply additional information such as manuals and software. The researchers were however not asked if they had experience with submitting data to external facilities and most of the respondents showed a lack of familiarity with the issue. The respondents rate for, “I don’t know”, remains the highest one, selected by about 50% of the respondents from the different options. ”Supply additional information such as manuals, software”, is in between, with 51% “not knowing whether this is necessary” when submitting data to an external facility. The high level of uncertainty might reflect the fact that most of the respondents do not have experience with submitting data to external facilities. By combining this question with the respondents who answered ‘yes’, on having to include data when submitting materials to journals, the level of uncertainty decreases. The most evident change is that the need to comply with a standard format rose from 40% to 75%. This clearly shows that the group of researchers with experience in some kind of data exchange, have a better knowledge about metadata issues, while researchers who have not yet had the experience of making their data available to others are unfamiliar with standards and exchanges format that are needed.

4.2.4 Collaboration in research

When asked about their collaboration with other researchers in their current projects, as seen in figure 4.6, it was found that only 14% of the researchers did not collaborate with other researchers at all. 34% of the respondents replied that they collaborate with researchers from other projects in their discipline, 6% that they collaborate with other researchers in other disciplines and 47% replied that they collaborate both with researchers from their own and from other disciplines.

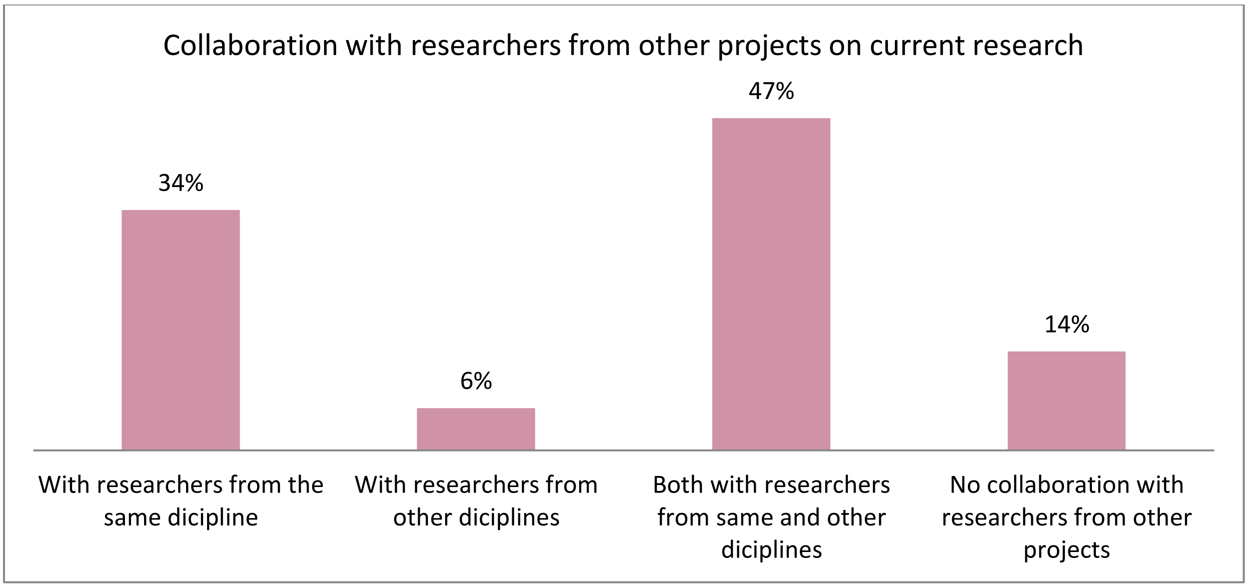


FIGURE 4.6: QUESTION 13

- REGARDING YOUR CURRENT RESEARCH, DO YOU COLLABORATE WITH RESEARCHERS FROM OTHER PROJECTS?

The high level of collaboration tells us something about how the researchers work in the university. There is a high level of not only disciplinary collaboration but also interdisciplinary collaboration. One of the popular arguments for the sharing of research data is that it will promote new collaborations particularly between disciplines that traditionally are not connected. Much interdisciplinary collaboration at UMB might be explained as the University is a small institution where researchers know each other. Even if they are not in the same department, the focus on certain topics such as food, encourages researchers from different sciences to work together in order to cover all angles of the topic.

4.2.5 Routines for and interest in using research data collected by other researchers

The results from the questions about collaboration are also reflected in the reuse of data collected by others. As illustrated in figure 4.7, much collaboration is done both within the same discipline and with researchers from other disciplines. More than half of the respondents currently make use of data collected by other researchers in the same discipline, while about one third use data collected by other researchers in other disciplines. There are however, as the figure illustrates, large differences between the different departments on both the disciplinary and cross disciplinary reuse of data.

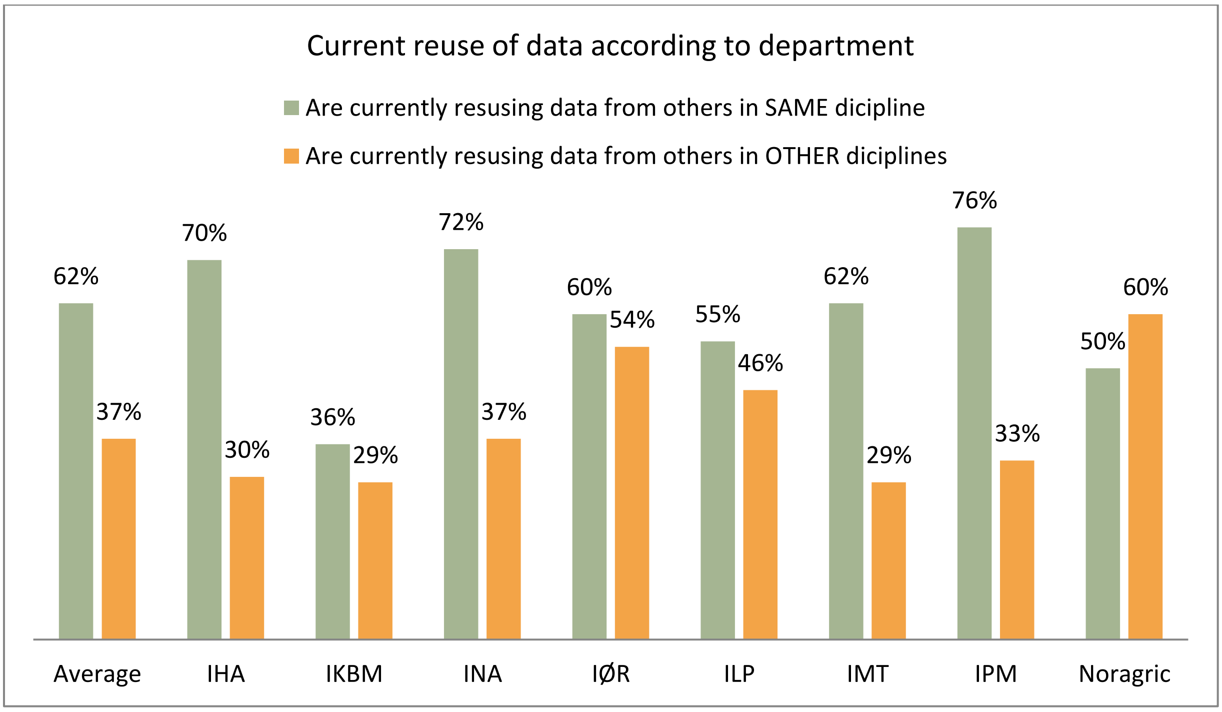


FIGURE 4.7: QUESTION 15 & 16- DO YOU PRESENTLY MAKE USE OF RESEARCH DATA GATHERED BY OTHER RESEARCHERS?

Also according to level of experience in research there are differences; the researchers most active in reusing data from others are those with 10 to 20 years of experience, in fact for the reuse of data from other disciplines the numbers here are, as the figure 8 shows, more than double compared to researchers with 5 to 10 years of experience. It is somewhat surprising to find that this kind of reuse decreases from the new researchers and PhD students to those with 5 to 10 years of experience, both in the same discipline and in other disciplines.

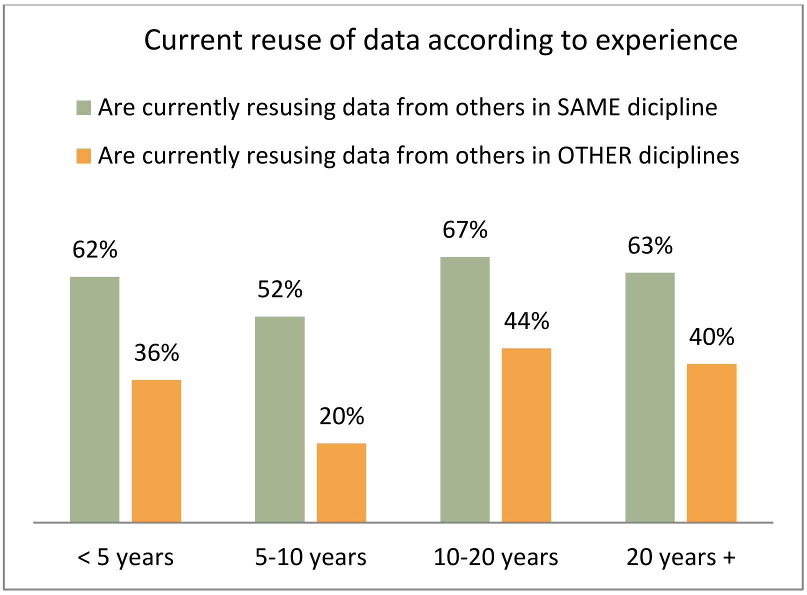


FIGURE 4.8: QUESTION 15 & 16 (ACCORDING TO YEARS OF EXPEREINCE)

- DO YOU PRESENTLY MAKE USE OF RESEARCH DATA GATHERED BY OTHER RESEARCHERS?

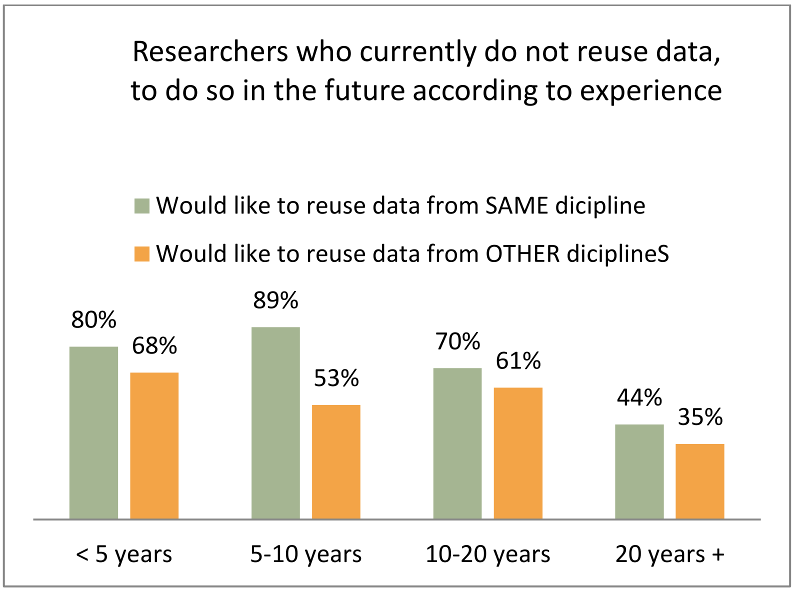


FIGURE 4.9: QUESTION 17 & 18 (ACCORDING TO YEARS OF EXPERIENCE)

- WOULD YOU LIKE TO MAKE USE OF RESEARCH DATA GATHERED BY OTHER RESEARCHERS?

About two thirds of the researchers that do not use data collected by others in their discipline would like to do so. Figure 4.9 shows that the top score here is among the researchers with 5 to 10 years of experience and the researchers with less than 5 years of experience, after 10 years of experience the interest in reusing data from others, among those who currently do not use data from other researchers, decreases. Less people are interested in using data from other disciplines; only about half of those who do not use data from others in other disciplines would like to do this. Here the variation is however large, between the different disciplines as illustrated in figure 4.10. There is also a decrease in the interest of data reuse from other disciplines as the researcher gets more experienced. Among the researchers with less than 5 years experience, two thirds would like to reuse data from other disciplines while among researchers with more than 20 years of experience only one third would be interested in reusing research data collected by researchers in other disciplines. Still the general interest in reusing data from other researchers is high and the fact that it is found to be higher among younger researchers may be explained with the fact that their work routines are less established, and that they to a larger extent are still developing their network both in and across disciplines. This indicates that researchers have the network for accessing and reusing data from others, after 10 to 20 years of experience, as the interest in reusing data among those who currently do not reuse, goes down, while the existing reuse of data from others stays on top.

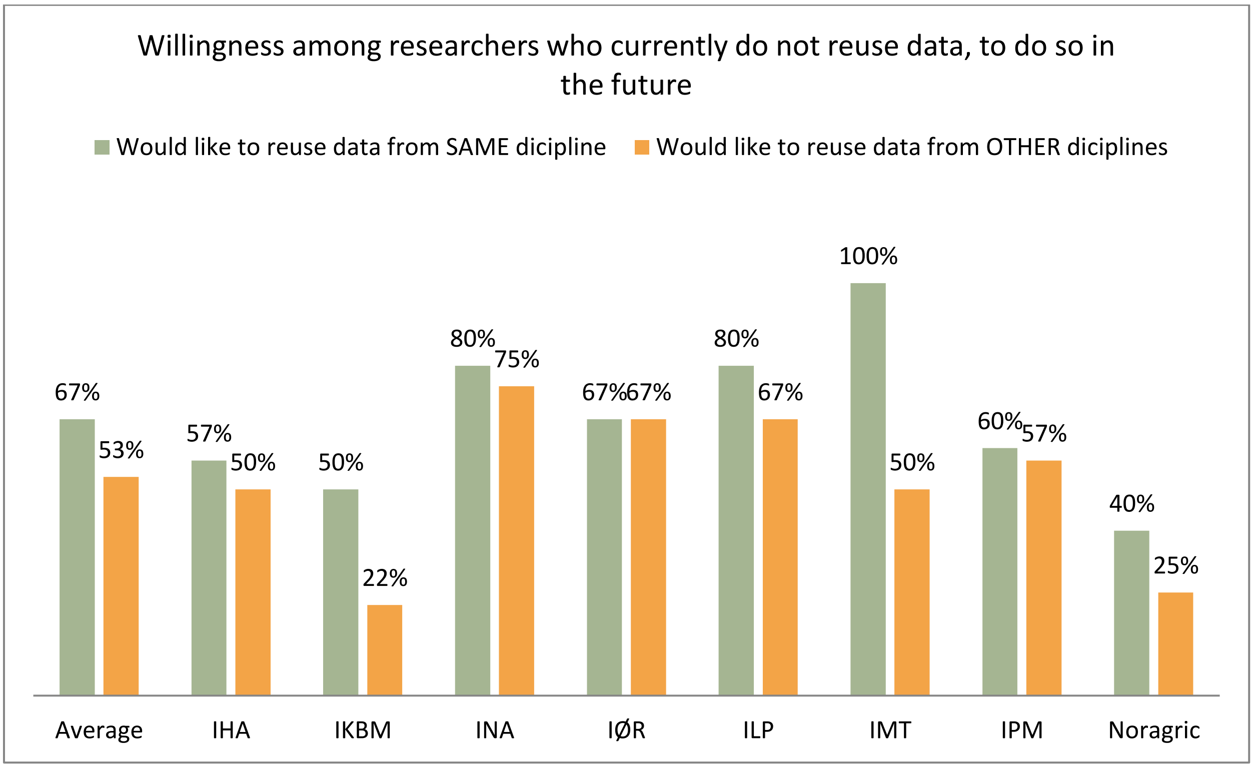


FIGURE 4.10: QUESTION 17 & 18 (ACCORDING TO DEPARTMENT)

- WOULD YOU LIKE TO MAKE USE OF RESEARCH DATA GATHERED BY OTHER RESEARCHERS?

As for locating digital research data, figure 4.11 shows that a majority of the researchers replied that they find and access digital research data via colleagues. Other popular channels are formal literature and institutional databases and search facilities followed by general search engines such as Google etc. The variation between the research disciplines is large, some being close to the media, others with quite different patterns. When it comes to years of experience, the variations are smaller. New researchers tend however to use several channels, while the more experienced ones use only one or two. The usage of social media channels is also highest (13%) among the researchers with less than 5 years experience and is currently not used by more experienced researchers.

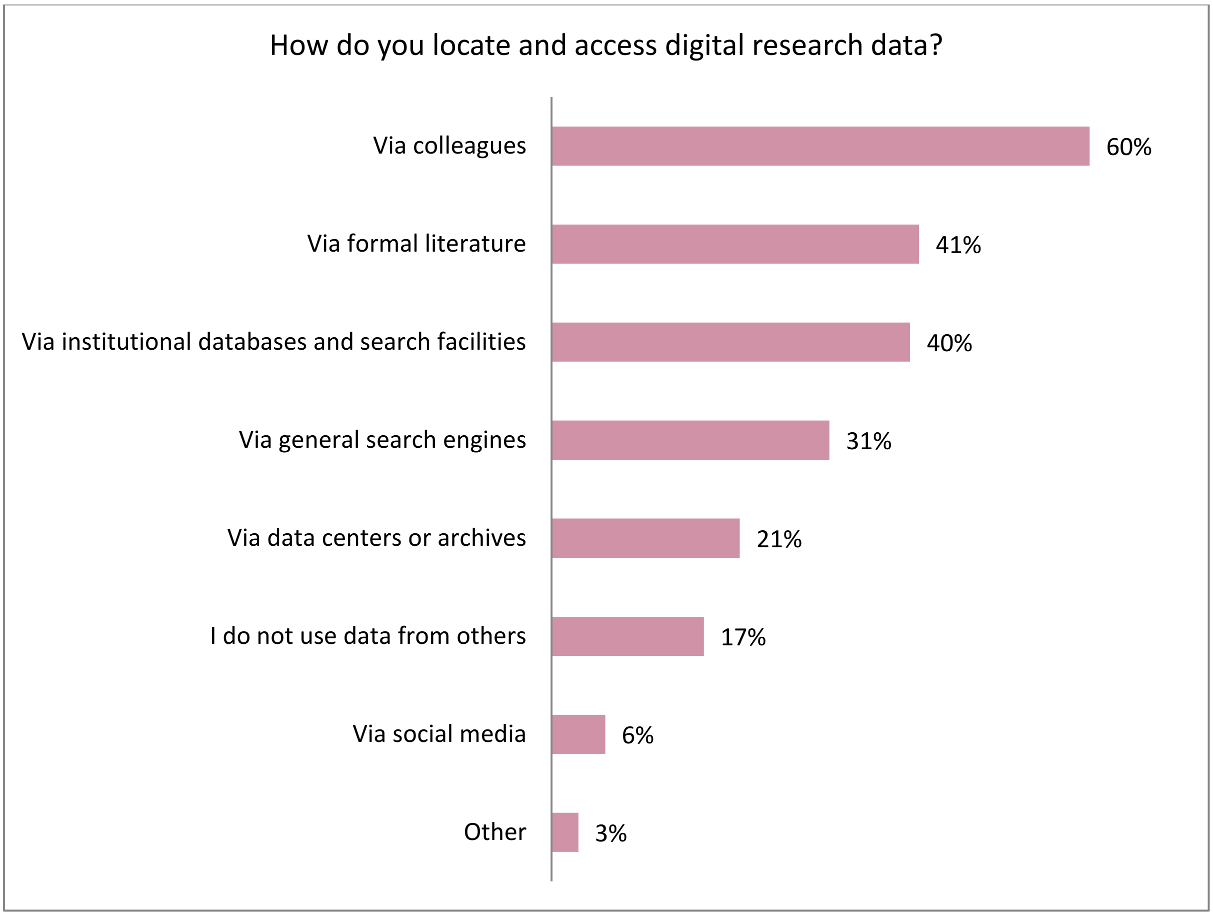


FIGURE 4.11: QUESTION 19

- HOW DO YOU LOCATE AND ACCESS DIGITAL RESEARCH DATA?

37% of the researchers have experienced to the need research data from others that was not available, the number of researchers answering that they did not experience this was slightly lower. However 28% replied that they did not know if they ever needed material from other researchers that was not available. The fairly high number of researchers experiencing the need to research data from others that was not available, illustrates that the routines for preserving and sharing are fare from optimized. The differences between the disciplines appear; indicating how the usage of data from others varies much between the different disciplines. Not getting access to data gathered by others was most common among the researchers with 5 to 10 years of experience, where about half the respondents replied to have experienced not getting data they needed, and was less common among the researchers with the longest experience.

The results regarding routines for storing and reusing research data among the researchers at UMB, shows us that much sharing is already taking place; researchers share their material with fellow researchers in particular with those they collaborate with. This is again supported by how the researchers mainly find research data they need through fellow researchers. These collaborations are however largely based on contact and networks support functions that take time for new researchers to establish, a formalization of the exchange of research data could make more knowledge available for new researchers too.

4.3PERSPECTIVES ON THE FUTURE OF DATA SHARING

Results regarding the attitudes the researchers have towards the sharing of research data, included fears and potential obstacles they see in the future of data sharing. These are presented here. The questions aimed to detect what reasons for preserving research data the researchers found important, and further which risks they foresee in a future of the open sharing of research data. The questions were combined to discover differences according to years of experience or which department the researchers work in, or if the opinions are so general that discipline and experience do not affect how the different issues are seen.

4.3.1 REASONS FOR PRESERVING DIGITAL RESEARCH DATA

Researchers were presented with reasons or arguments for preserving digital data and were asked to rate them on a scale of four options from very important to not important. The option, “It allows for re-analysis of existing data” stands out as the reason found most important followed by “It might stimulate the advancement of science”. The argument that “If research is publicly funded, the results should become public property and therefore properly preserved”, and the possibility that “It may serve validation purposes in the future” also stands out as important to most of the researchers. There are however differences in how they are rated both according to discipline and years of experience. A further combination with the respondents who make their research data openly available today shows that the reason those researchers find to be most important is that “when the research is publicly funded the results should become public property and therefore properly preserved” was regarded as very important by 61% of the respondents in this group, significantly higher than the average of 37%.

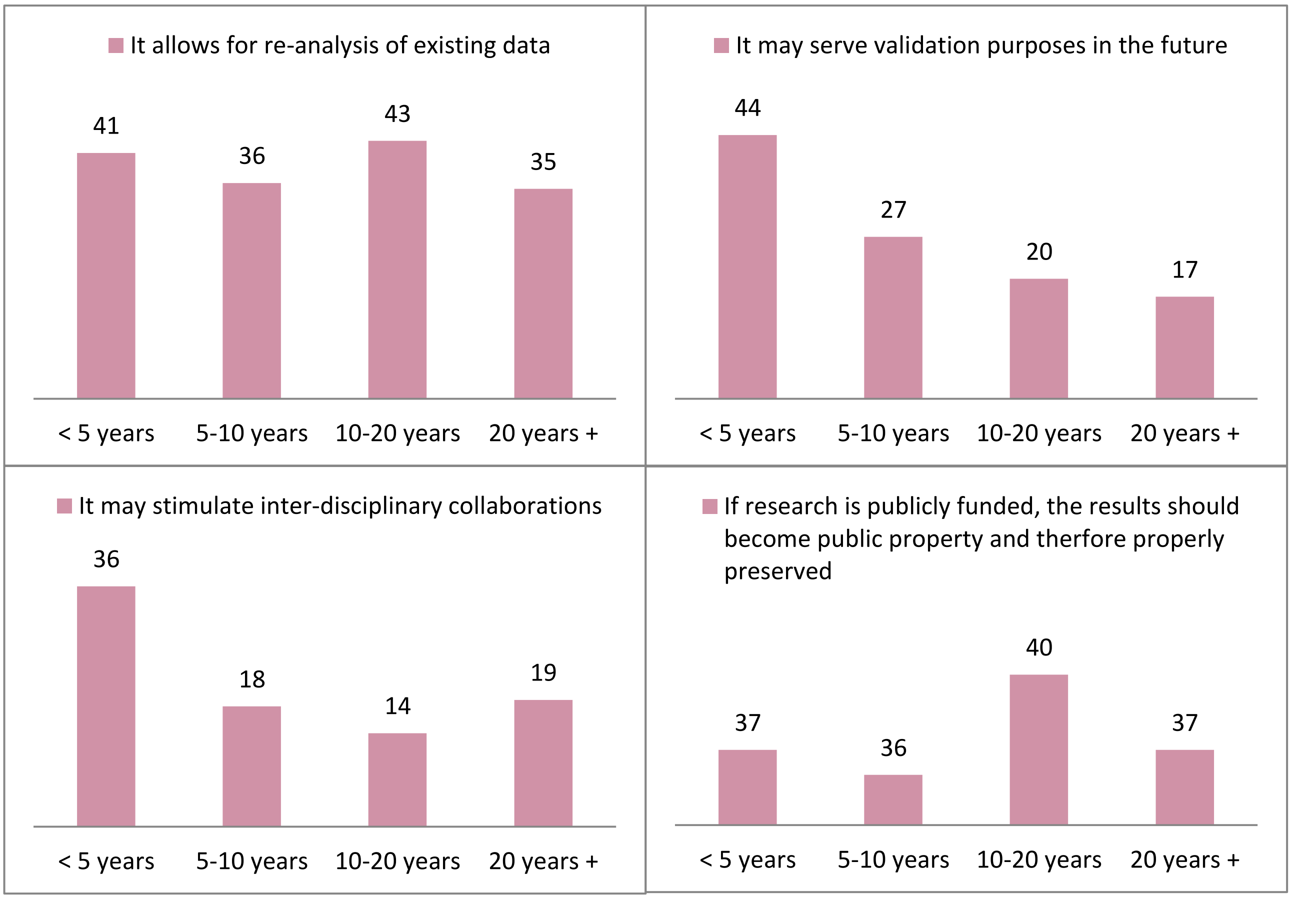


FIGURE 4.12: QUESTION 27 (ACCORDING TO YEARS OF EXPERIENCE)
- REASONS FOR PRESERVING DIGITAL DATA RATED I PORTANT OR VERY IMPORTANT

Figure 4.12 shows how the arguments are rated as very important according to the years of experience in research the researchers have. The idea that the sharing of research data may stimulate inter-disciplinary collaborations stands out as an argument with is much stronger support among the fresh researchers than among the more experienced ones. Similarly the validation purpose for storing data is found to lose importance for the researchers as they gain experience. The argument that public funded research should become public property is stronger among researchers with more experience. This shows that how researchers think about making data openly available depends much on where the researchers are in their careers.

4.3.2 THREATS AND PROBLEMS REGARDING SHARING OF RESEARCH DATA

Of the seven options given regarding different threats towards the preservation of digital research data on a scale of 5, from very important to not important, also containing the option “I don’t know”, the argument found most important was that the “evidence may be lost because the origin and authenticity of the data may be uncertain”. 67% of the respondents regarded this to be a very important or important threat. This was followed by the fear that “users might be unable to understand or use that data e.g. the semantics, format or algorithms involved” was rated important or very important by 54% of the respondents, and the fear that “access and use restrictions may not be respected in the future” by 53%. That “the current custodians of the data, whether an organization or project, may cease to exist at some point in the future” was found to be of importance by half of the respondents, however only 36% feared that “the ones trusted to look after the data may let them down”, making it the issue the researchers worry the least about.

When comparisons with the different departments are made, differences are found; in IKBM and IØR the usage of “I don’t know” ranges from 14% to 38% on the different options in the media. This alternative is used by 6% on “evidence may be lost because the origin and authenticity of the data may be uncertain” to 17% “the ones we trust to look after our holdings may let us down”.

Figure 4.13 illustrate how the different issues are regarded according to years of experience. The first layer is the percentage, rating the issue as very important, then the ratings for important are put above and together they form the percentage of all the respondents giving the issue importance. Again the graphs show that some issues are regarded as more important by younger researchers, but for other issues the importance is regarded as more or less the same. There are however clear variations depending on how many years of experience the researchers have.

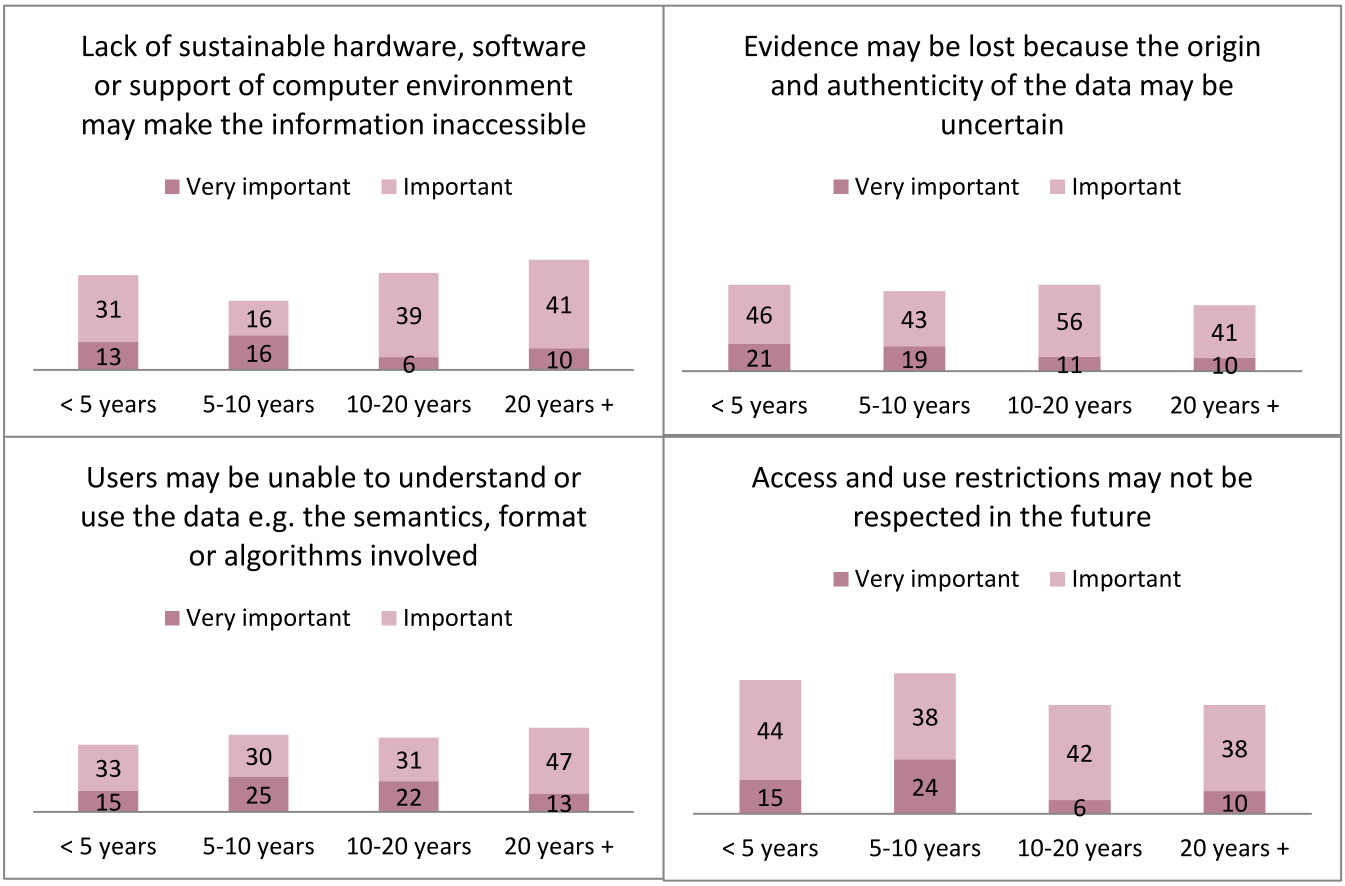


FIGURE 4.13:QUESTION 28 (ACCCORDING TO YEARS OF EXPERIENCE) - THREATS TOWARDS PRESERVATION OF DIGITAL DATA.

What are regarded as the most important threats can also be refined in the problems the researchers foresee in the preservation and sharing of digital research data, as illustrated in figure 4.14. Issues regarding misuse of data showed 48% and legal issues (41%). These are the matters that more of the researchers fear in the future. In IKBM the fear of losing the scientific edge (50%), tops the results, while for the other departments it is either misuse or legal issues that are the problems most of the researchers expect to meet in the future.

The other issues mentioned, such as incompatible data types, lack of financial resources or technical infrastructure do not concern the researchers much, only 18% to23% expect these to be problems in the future. In comparison 18% of the respondents do not foresee any future problems; of these most are, as the figure shows, researchers with more than 10 to 20 years of experience. There are however also other issues where the results are affected by the years of experience, such as “fear of losing the scientific edge” among the group with less than 5 years of experience. “Incompatible data types” worries the group with 5 to 10 years of experience much more than their colleagues.

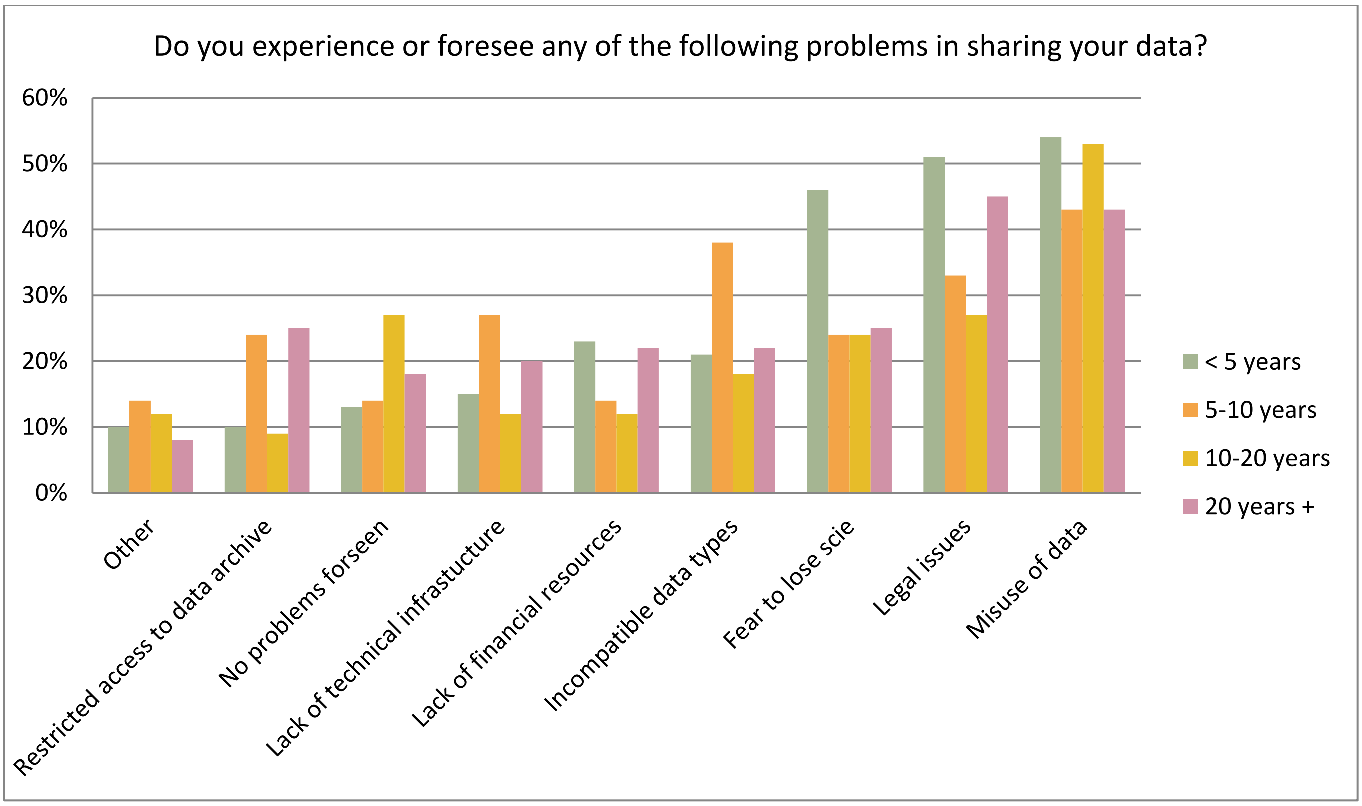


FIGURE 4.14: QUESTION 14 (ACCCORDING TO YEARS OF EXPERIENCE)

INCREASINGLY, AWARENESS IS GROWING THAT DATA SHOULD BE SHARED AS WELL AS PUBLICATIONS.

- DO YOU EXPERIENCE OR FORESEE ANY OF THE FOLLOWING PROBLEMS IN SHARING YOUR DATA?

Other issues mentioned by the respondents are privacy issues for informants, misinterpretation and ethical challenges. Some of the respondents fear that it will be time consuming to give proper documentation for use and that a structure for data sharing will become some kind of unnecessary bureaucracy. The fear that it might lead to other researchers getting credit for their work, or that they simply will not have the time to extract and publish results based on the data within a reasonable time, and then lose the benefit of the first publication, are also expressed.

> “Field data requires a lot of time and financial resources, the person strongly involved in the planning and acquisition of the data should have the benefit of first publication. You don't want to end up as the data producer while others get to do the interesting work. First publication should ideally come early, but several factors can affect this time, having children, having to apply for new funding e.g. in non-governmental institutions where you have to get funds for your use of time...” (IPM)

This shows that the researchers, rather than fearing that others will not have the technical or financial resources to handle their data, fear losing control of the data they collect; that it will be misused or that restrictions for use will not be respected. Also that it will create conflicts between them and their informants and that it will give them less time for research if a mandatory data deposit system is introduced. These fears must be taken into account, both by maximizing automatic extraction of metadata and by establishing clear routines for how to deal with privacy, ethical and juridical issues. Furthermore, a tradition for proper data citation and accreditation must be put in place, along with the benefit of the first publication for the researcher(s) collecting the data.

Also, in a question where the researchers were given the possibility to freely comment on the issues of data sharing the same points where repeated:

> *“I think as soon as it’s published, it could/should become available for general use, but there is a lot of competition for publications so it’s necessary to have exclusive access first.” (IKBM)*

> *“There is a huge difference between the usages of data for commercial purposes vs. research purposes. I would not hesitate to share my data with other researchers, but I would not like them to be used commercially. That could weaken the credibility of my research.” (IØR)*

> *“Many data can by nature not be shared such as interviews. Quantitative data should be exclusive for researcher for a limited period.” (ILP)*

> *“If the data collection has mainly public funding, the data should be free for anybody to use.” (INA)*

> *“But two major concerns: (1) Incentives for collecting data: if all data are to be open and unrestricted, why bother to gather data oneself (2) Making data online/available for all entails extra documentation costs, which further adds to concern (1)” (IØR)*

> *“Much of my primary data are provided by public institutions such as Statistics Norway or the Norwegian Mapping Authority. What I produce will in such cases be secondary and slightly aggregated or streamlined compared to the raw data. Sharing such data is then depending on approval of the original owner. 3 years ago I was part of a project in Malawi where eventually was denied access to data by the president (or rather the national statistical office was denied the right to publish the data. It was politically too sensitive. “(ILP)*

“Often in my case the data are owned by private Breeding companies” (Nofima)

“Who is to pay the costs of making data accessible to others? Other people can use up your ideas…” (Noragric)

“It is difficult to use someone else’s data and misinterpretations/misunderstandings my cause erroneous conclusions to be drawn. Therefore the researcher gathering the data ought to be able to state when data is not applicable to a question...” (ILP)

“Unless it is data of general nature, but if it is specific data generated for a specific study with important findings, one may want to hold on to data until the findings are established and ownership acknowledged.“ (Noragic)

The comments show an interest in the topic, often combined with concerns about the need to regulate sharing of their data. The researchers themselves know best what kind of data they collect, and what issues might occur in the case of open sharing, they are however seldom experienced in making their data re-findable or possible to interpretation by others. Furthermore the comments show the diversity not only with regards to types of data and data formats, but also in terms of where the data was collected. Usage of data from Statistics Norway and the Norwegian Mapping Authority is already openly available, and can be properly cited with the usage of for instance doi`s, as done in the Australian, ANDS. Data from private breeding companies must most likely be dealt with in a different way. Others express that they think data should be shared along with first publication. In addition, the commercial aspect is mentioned as a factor that the researchers would like to be able to put restrictions on. The idea of open research and their data as public property does not include commercial purposes. Concerns regarding privacy and costs connected to making data publicly available are also mentioned.

A illustrated in figure 4.15 to 4.17 the researchers were also asked to express their support for different statements regarding data sharing. Here the variation between diciplines was more evident than variation according to years of expereince. This indicates that the importance of some of these issues depends on the dicipline. The statement “sharing of research data would strenghten the research in my field” (figure 4.15) is directly affecting the diciplines.This is reflected in the answers, as everyone in the institutions, the INA, IØR, ILP and IM agrees or are undecided about this statement, whereas in the remaining insitiutions there are researchers who belive that the sharing of research data would not have a positive impact in strenghtening the research. Also, whether open access to research data would stenghten the research, tends to be an issue where the support depends on the dicipline, as no one from IKBM or IØR dissagrees with such a statement.

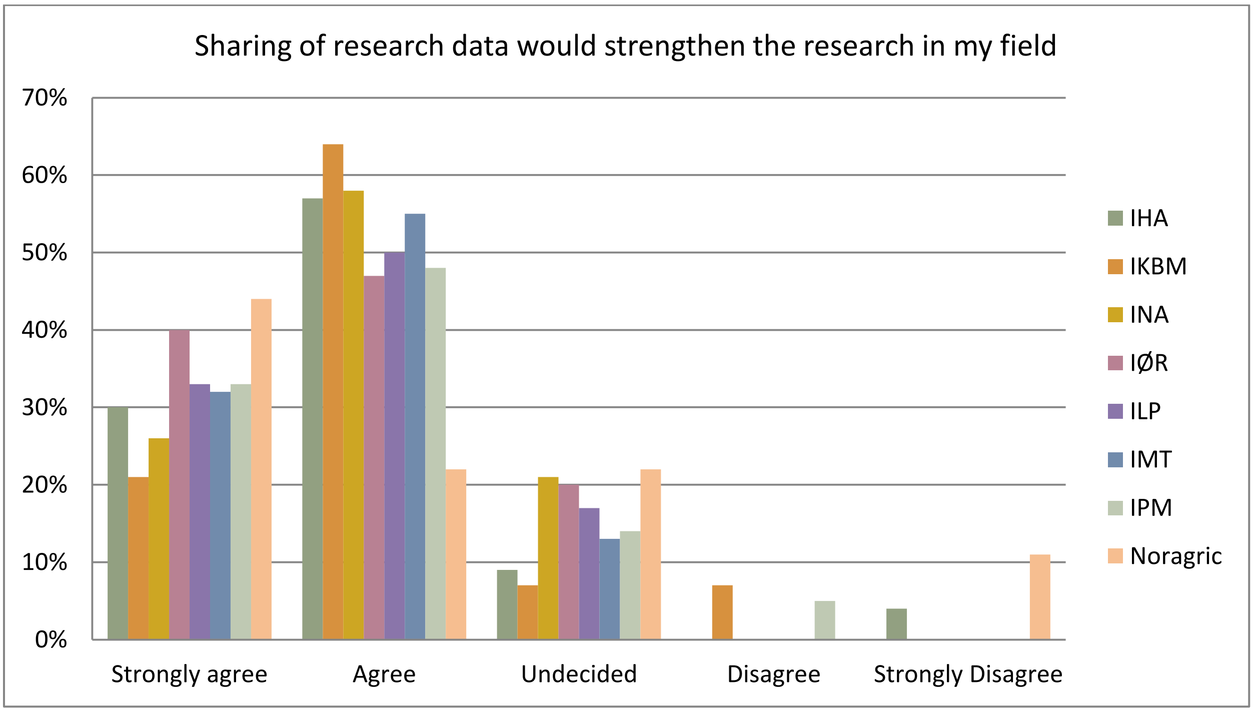


FIGURE 4.15: QUESTION 7.2 (ACCORDING TO DEPARTMENT)

- SHARING OF RESEARCH DATA WOULD STRENGTHEN THE RESEARCH IN MY FIELD.

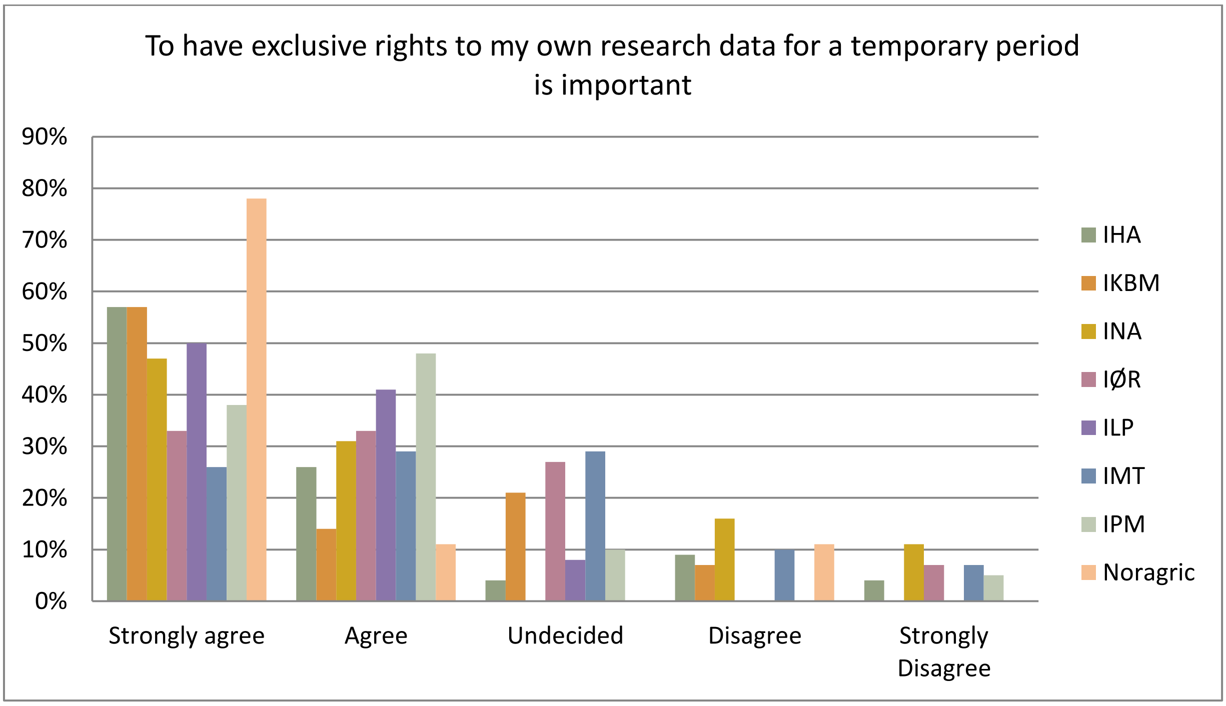


FIGURE 4.16: QUESTION 7.1 ACCORDING TO DEPARTMENT

- TO HAVE EXCLUSIVE ACCESS TO MY OWN RESEARCH DATA FOR A TEMPORARY PERIOD IS IMPORTANT.

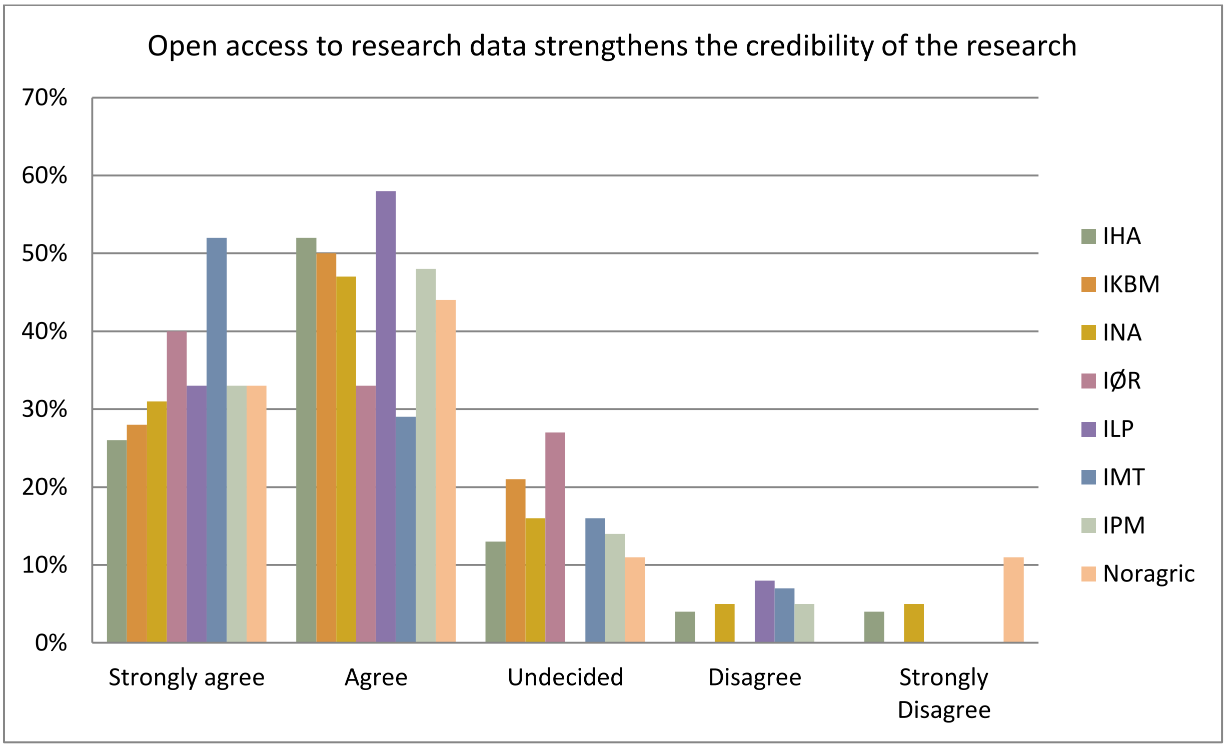


FIGURE 4.17: QUESTION 7.4 (ACCORDING TO DEPARTMENT)

- OPEN ACCESS TO RESEARCH DATA STRENGTHENS THE CREDIBILITY OF THE RESEARCH.

Figure 4.16 tells us that the possibility for the researcher to have exclusive rights to their data for a temporary period is important in all disciplines, whereas the importance of other issues depends more on discipline. Sill most of the researchers in question believe that sharing of research data would strengthen the research in their field and many believe that open access to research data would strengthen the credibility of the research. These are important arguments to keep in mind as it is crucial that the research community believe that sharing and open access would have a positive impact on their profession.

4.3.3 USE OF OPEN RESEARCH DATA AND ACCREDITATION

More than 80% of the respondents believe that it is useful to link digital research data with formal literature. Again there are variations between the disciplines from 93% in IKBM to 74% in INA and ILP. Still the data strongly supports initiatives on linking data to publications, even if the support to some extent depends on the discipline. As illustrated in figure 4.18 the researchers with less than 5 years of experience and the researchers with 10-20 years of experience are the most positive about this while researchers with 5-10 years of experience are more negative and the most experienced researchers are the least supportive of the idea of linking research data to literature.

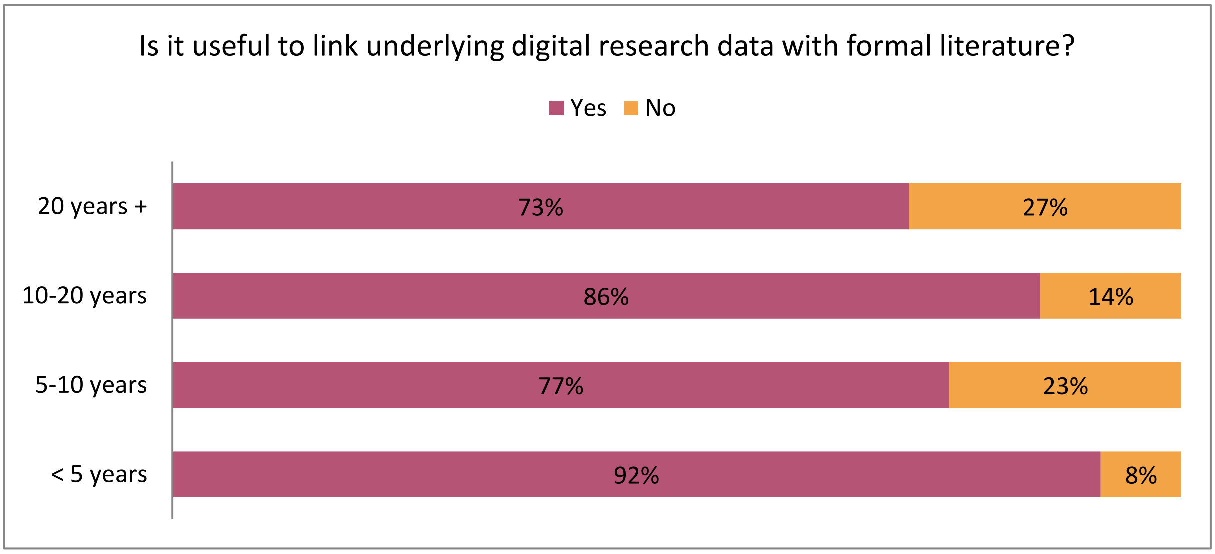


FIGURE 4.18. QUESTION 22 (ACCORDING TO YEARS OF EXPERIENCE)
- IS IT USEFUL TO LINK UNDERLYING DIGITAL RESEARCH DATA WITH FORMAL LITERATURE?

Almost all the researchers (94%) want to be credited when underlying research data produced by them is used by others. This necessitates having a valid system for data linking and citing and possibly even the need to introduce an official accreditation system giving the researchers credits in line with the credits they get for publishing, when data that is made available is reused by other researchers. There is today a high pressure on researchers to publish books and articles, as the universities are "rewarded" depending on this production. If similar rewards were introduced for making data available, or data being cited in other research, this would become a concrete motivator for the researcher. As the results show, the attitude towards making data openly available depends heavily on where the researchers are in their careers.

4.3.4 Perspectives on the future

Figure 4.19 shows that about one third of the respondents would be willing to submit their research data to an open digital archive of the organization in the near future, but options like the publisher, disciplinary data center or a closed archive of the organization are preferred by most researchers. 12% of the respondents would not be willing to submit their digital research data to any external facilities in the near future. Figure 4.20 shows that the main reasons given for this choice are fear of losing copyright, that it will be misinterpreted and that the respondents do not trust the safety of datacenters, journal sites or repositories. The options that were not selected weres "I do not know of any digital archives (repositories or data centers) to which I can submit data", "Submitting digital research data costs money and therefore is not attractive to me" and "I do not want to run the risk of anyone else being able to access and use my digital research data"

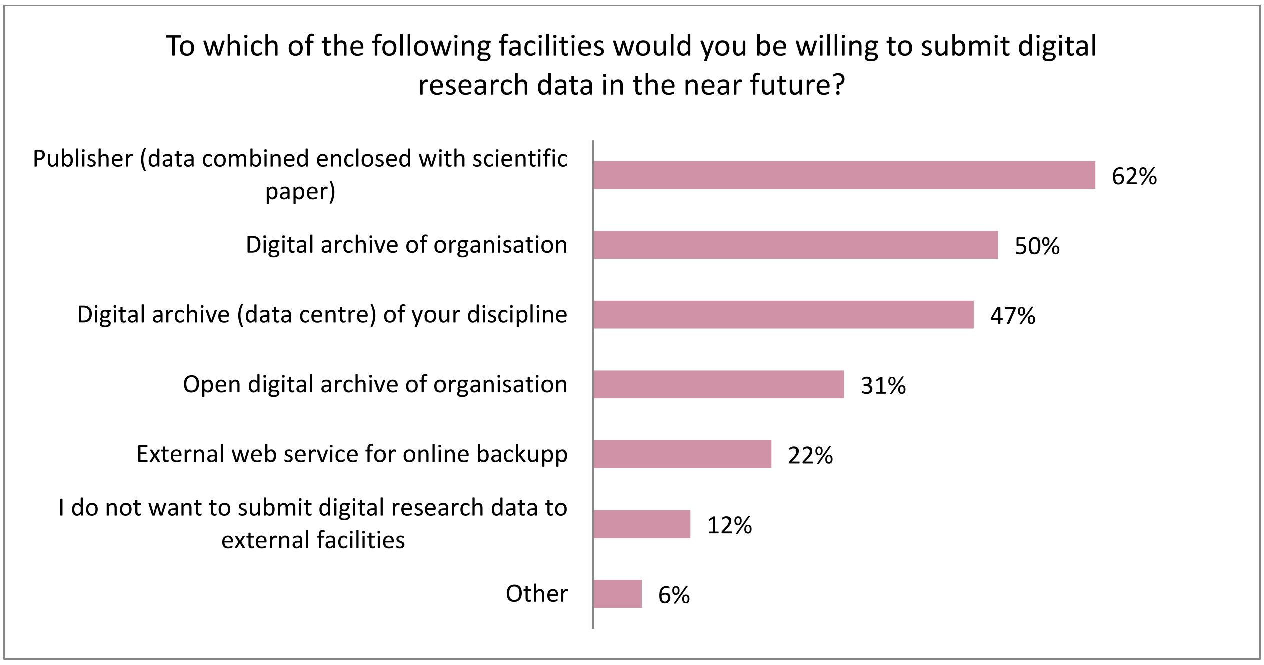


FIGURE 4.19: QUESTION 30

- TO WHICH OF THE FOLLOWING FACILITIES WOULD YOU BE WILLING TO SUBMIT DIGITAL RESEARCH DATA IN THE NEAR FUTURE?

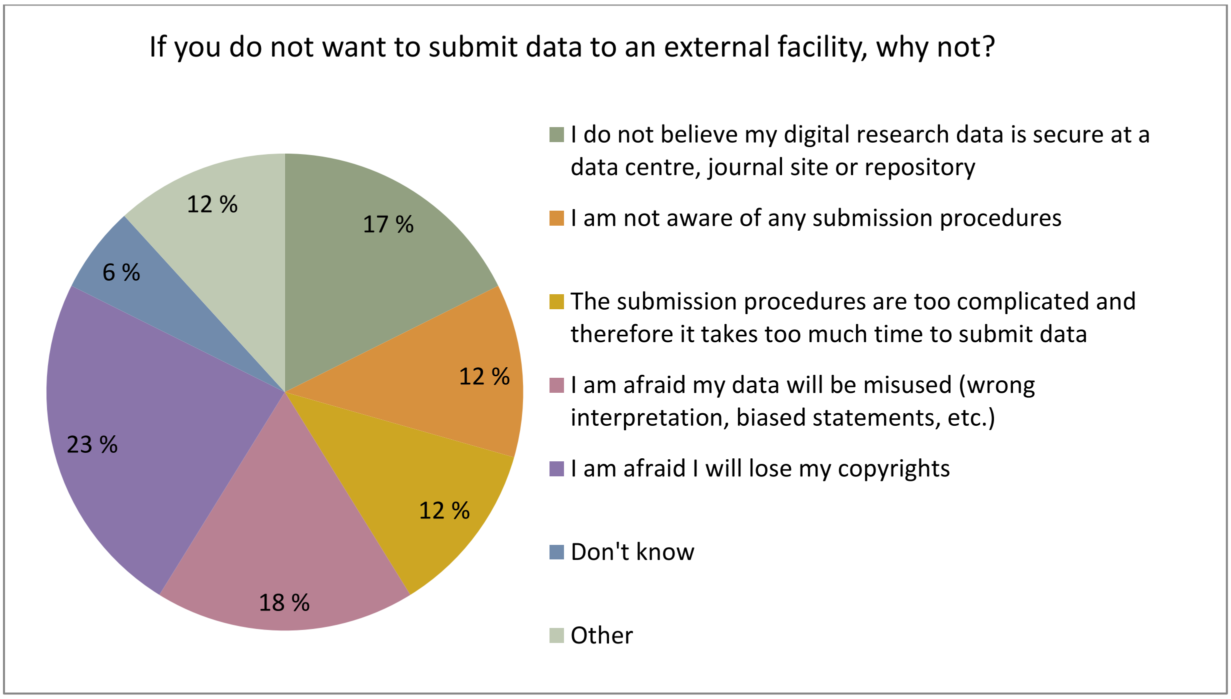


FIGURE 4.20: QUESTION 31

- IF YOU DO NOT WANT TO SUBMIT DATA TO AN EXTERNAL FACILITY, WHY NOT?

The location to which the researchers would be willing to submit data, and the reasons given by those who do not want to submit data, reflects the fear identified in previous questions, that they will lose copyright or that the data will be misinterpreted. The reason why the publishers score high as place were researchers would be willing to submit their research data, might be explained with the fact that journals are associated with prestige and credibility, and in some cases are already requiring background data for an article to be published. However if the academic publishing continues to change and goes in the direction of more being published at earlier stages than the article, as predicted by, for instance, by Cameron Neylon (2011), this change might also become evident in where the researchers would like to store their research data. It is however not clear from the question what type of data storage, facilitated by the publisher, the researchers expect. It might be the possibility of linking figures and the data illustrating their findings to a publication, or it might be the possibility of storing complete data sets in a database. It is however more likely the first alternative, as this is the tendency the publishers are to some extent already offering.

This finding does however present the issues of how to store, what to store and who can offer what. In the publishing world we see that articles are stored in institutional repositories as well as by publishers, though sometimes the versions of the material are not the same. With the storage of data a similar situation might occur, where complete data sets are stored in repositories with visions of future reuse, in combinations, data mining or citing while the publishers store parts of a data set visualized in an attractive way along with a publication. In other words, one does not exclude the other and repository for storage of complete data sets must be constructed with the thought of data exchange and citation in mind.

Almost half of the respondents believe that a national infrastructure for data preservation and access should be created to help guard against some of the threats given earlier; the remaining half contains both people who are not sure or are against the construction of such an infrastructure. Figure 4.21 shows that there are large variations between the different institutes on this question, indicating that such an infrastructure might be more welcome in some disciplines than others.

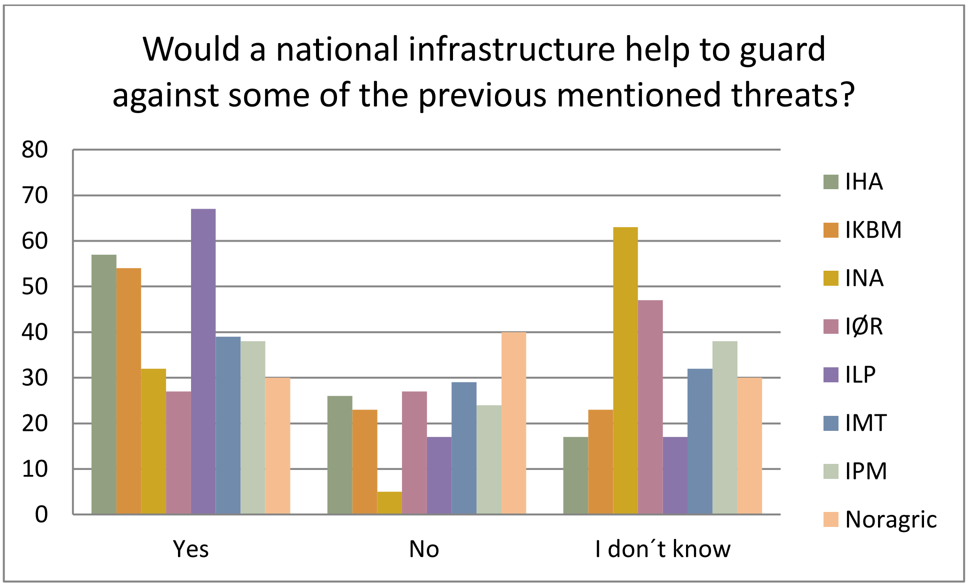


FIGURE 4.21: QUESTION 32 (ACCORDING TO DEPARTMENTS)

- DO YOU THINK THAT A NATIONAL INFRASTRUCTURE FOR DATA PRESERVATION AND ACCESS SHOULD BE BUILT TO HELP GUARD AGAINST SOME OF THESE THREATS?

Regarding initiatives to raise the level of knowledge about the preservation of digital research data, the youngest researchers were again the most positive ones, still all the initiatives proposed got support from half of the respondents or more, indicating that the researchers are generally positive to learn more about preservation of digital research data.

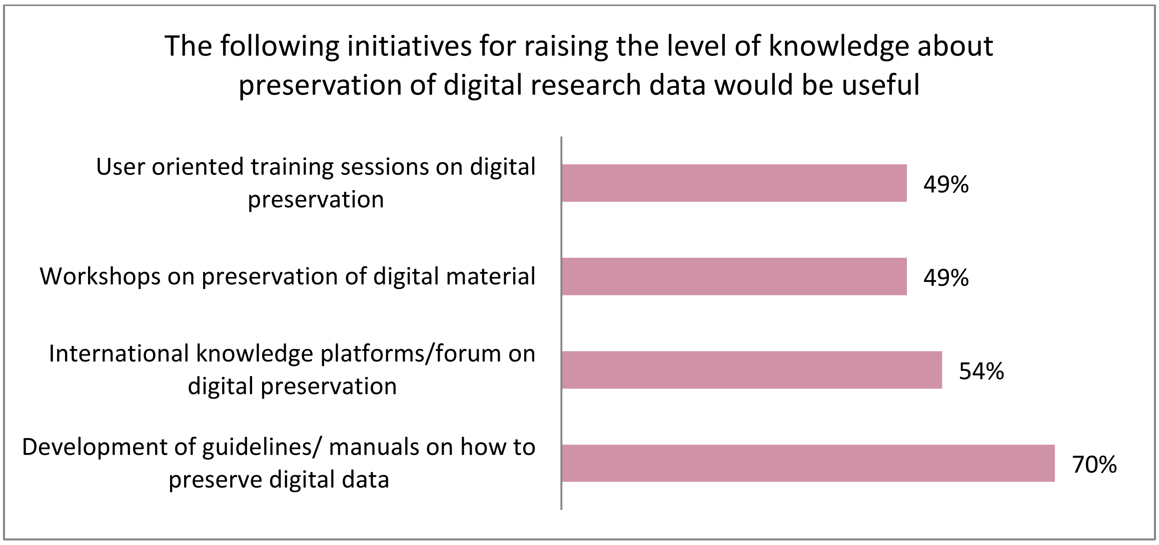


FIGURE 4.22: QUESTION 33

- DO YOU THINK THE FOLLOWING INITIATIVES WOULD BE USEFUL FOR RAISING THE LEVEL OF KNOWLEDGE ABOUT PRESERVATION OF DIGITAL RESEARCH DATA?

Still the numbers presented in figure 4.22 tells us that most of the researchers are positive to learn more about preservation of their digital research data, and that guidelines or manuals for this should be developed, the question arising from this is however who should be responsible for developing these guidelines? The literature shows that in the US, schools of LIS are adding data curation to their curriculum, so far this has not happened here in Norway. It is therefore a relevant question to ask who has the competence to develop guidelines on how to preserve digital data?

4.4 Other issues

In addition to the issues previously discussed the questionnaire also contained some questions regarding academic publications and the use of online research collaboratoratories. These questions are found to be relevant for the topic of this thesis, however they do not fit properly in either of the two main categories used in the analysis of the results. These issues are therefore presented separately as they are believed to provide supplementary information about how the researchers work.

4.4.1 Publication related questions

When asked about channels for publications (figure 4.23) 91% of the researchers answered that they publish in journals available on subscription, whereas 40% used Open access (OA) journals. The second most common form of publishing is conference proceedings, 60%. Books and grey literature (technical reports, white papers, preprints etc.) were used by 26% of the respondents. Again there is a variation between the institutes, however traditional journals are, with only one exception (ILP), the most used publication channel, and also proceedings score high in all the institutes. In Cigene and IKBM, OA journals are more common than proceedings with 100% vs 40% and 69 vs 61%. In INA these two are equal with 42%. Noragric is the only institute where publication of books is more common than conference proceedings. Traditional journals, conference proceedings and OA journals are in this order the three most common publication forms independent of years of experience. There is however a stable increase of the publication of books and grey literature as the researchers gain more experience, and for the researchers with more than 20 years of experience publishing in books is just as common as OA journals (39%).

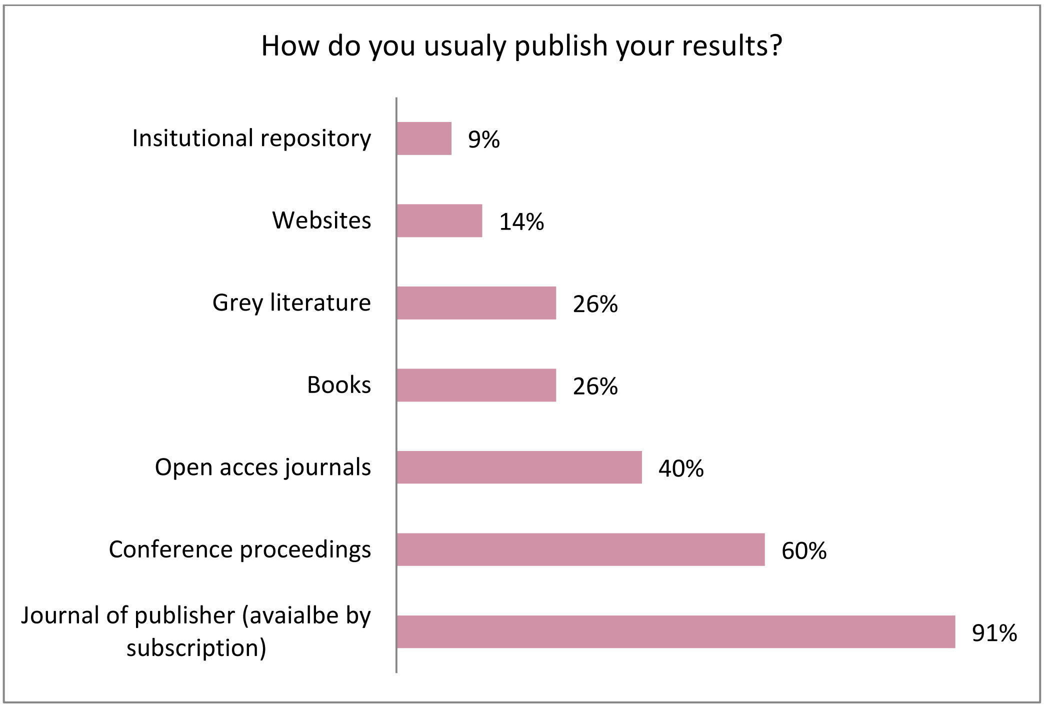


FIGURE 4.23: QUESTION 21 - HOW DO YOU PUBLISH YOUR RESEARCH RESULTS USUALLY?

One fourth of the respondents answers that the journals to which they typically submit work requires them to include relevant digital research data such as tables, figures etc. The tradition of including such material is discipline dependent as 71% of the respondents from IKBM were requested to include such material whereas none of the 19 respondents form INA have experienced such material to be requested by journals. Such requests appear to be slightly more common for less experienced researchers and this then decreases as the researcher gains experience, however the difference here is small compared to the difference between disciplines.

4.4.2 Usage of online research collaboratories

20% of the respondents state that they use online research collaboratories. Again the variation between the institutes is large, from 80% at Cigene to 0 at INA. Age has less of an impact; however the most active users of such collaboratories are the senior researchers with more than 20 years of experience (27%), of researchers with 10-20 years of experience the usage is 11%, and among the younger researchers the usage is somewhere in between. This indicates that within established research communities research collaboratories are more commonly used. Furthermore it tells us that usage of virtual environments is not necessarily the arena for younger researchers as one might think based on previously presented findings showing how younger researchers use a larger number of platforms to retrieve data. To a larger extent working on establishing a network of researchers there is a general assumption that younger people are more active in the virtual environments. Based on the data collected in this survey and presented in figure 4.24 such an assumption is found not to be valid.

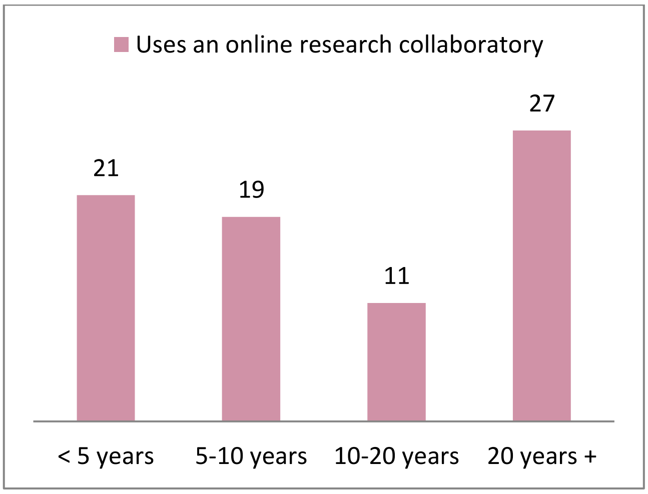


FIGURE 4.24: QUESTION 24 - USAGE OF ONLINE RESEARCH COLLABORATORIES

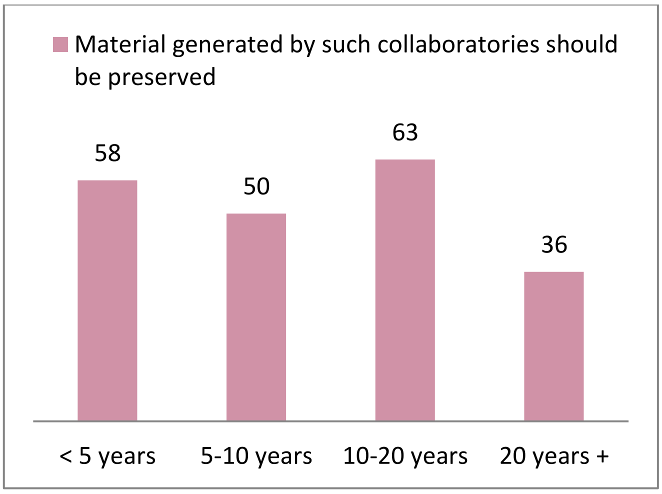


FIGURE 4.25: QUESTION 25 - ATTITUDES TOWARDS PRESERVATION OF MATERIAL GENERATED BY ONLINE RESEARCH COLLABORATORIES

Half of the respondents believe that it is important that materials generated by such collaboratories are preserved. Institutes where such platforms for collaboration are being used generally think preservation of material is more important. However by crossing this question with the previous one, it is found that only 58% of the 29 respondents who use collaboratories believe that information generated by these should be preserved. As illustrated in figure 25, the senior researchers have less interest in preservation of such material (36%) while researchers with 10-20 years of experience believe that is it more important (63%).When compared to the data in figure 24 it seems a paradox that the group that is less familiar with online collaboratories is the one most focused on preserving the material, while the interest in preserving the collected material is lowest in the group with the highest usage. One can only assume that those who are familiar with what is produced do not see the value of preserving it, while those who do not explicitly know what kind of material the question refers to, believe that it might be of value in the future and therefore also important to preserve.

4.5 Comparison with the results from the PARSE.insight study

Most of the questions in the questionnaire were taken from a study conducted by PARSE.insight in 2009. Their study was sent out to researchers at CERN, the Humboldt University in Berlin and the ESA. In addition to questioning the researchers about long time preservation issues, other stakeholders were also asked to express their opinion on the topic. In this thesis we are however only dealing with the questions and results from the part of the study that was conducted among researchers at these three institutions.

Unfortunately the complete dataset collected in the study done by PARSE.insight, is no longer openly available. On their website there is a dead link to such a document and attempts to contact the group leader requesting the data have not given any results. The data used for the comparison is therefore data presented in the reports published by PARSE.insight (*Interim insight report*, 2009; Thaesis & van der Hoeven, 2010). The selection of questions used in the comparison was then based on the data available.

4.5.1 The research data created

In order to detect what kind of material was dealt with the researchers were asked to select from a longer list, which formats of digital research data they produce. The two most used formats where the same in the two studies: standard office documents- 82% at UMB and 94% in the parse study, followed by images used by 56% of the respondents at UMB and 79% of the parse insight respondents. In the parse study networked-based data were also used by 79%, at UMB only 24% replied to using this kind of data rating it as the 7th most used of the 17 options given. Other data formats popular among the researchers at UMB are raw data- 42%, Scientific and statistical data- 35% and plain text- 33%. In the parse study raw data, software applications, databases and source codes were all formats used by almost 50% of the respondents. This tells us that apart from the most commonly used formats there is a large variation in which formats the different researchers use for their raw data output. Furthermore, more of the researchers from the PARSE.insight study use several different formats in their research, as the scores were generally higher.

4.5.2 Existing traditions for data sharing

When it comes to the current sharing of research data the numbers from PARSE.insight are again higher than the numbers collected from UMB, telling us that the tradition of sharing is stronger in the institutions questioned in the PARSE.study. The variations are however under 10%. Unfortunately the PARSE.insight does not share their numbers for the rest of the question, but in UMB, 24% stated that access to their data is temporarily restricted and 28%, that they do not share their data but would like to do so in the future. Without the PARSE.insight numbers for the remaining options, the comparison remains incomplete and only gives an indication of differences without giving the whole picture.

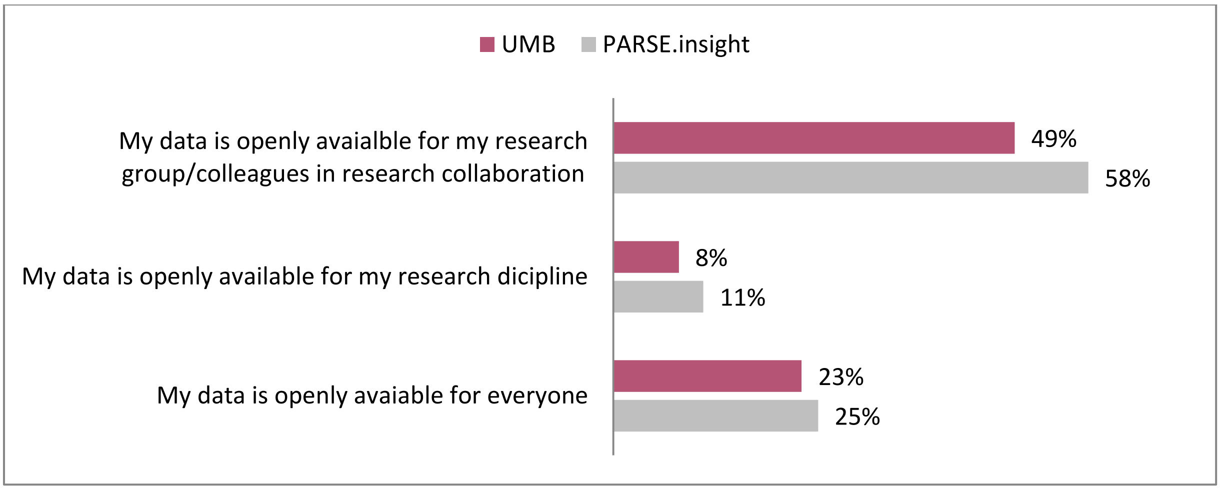


FIGURE 4.26: COMPARRISON ON HOW DATA IS MADE AVAILABLE

Figure 4.26 shows that the traditions for sharing are slightly stronger at the institutions form the PARSE.insight study; the results are however quite close to the results from UMB.

4.5.3 Willingness to reuse data from other researchers

While sharing among the UMB researchers tends to be lower than the results from PARSE.insight, the willingness to reuse data from other researchers is, as figure 4.27 shows, slightly higher at UMB. This, when it comes to its own disciplines, where the figure shows 67% of the responding UMB versus 63% of the researcher replying to the PARSE.insight questionnaire, and from other disciplines where the difference is larger with 53% versus 40%.

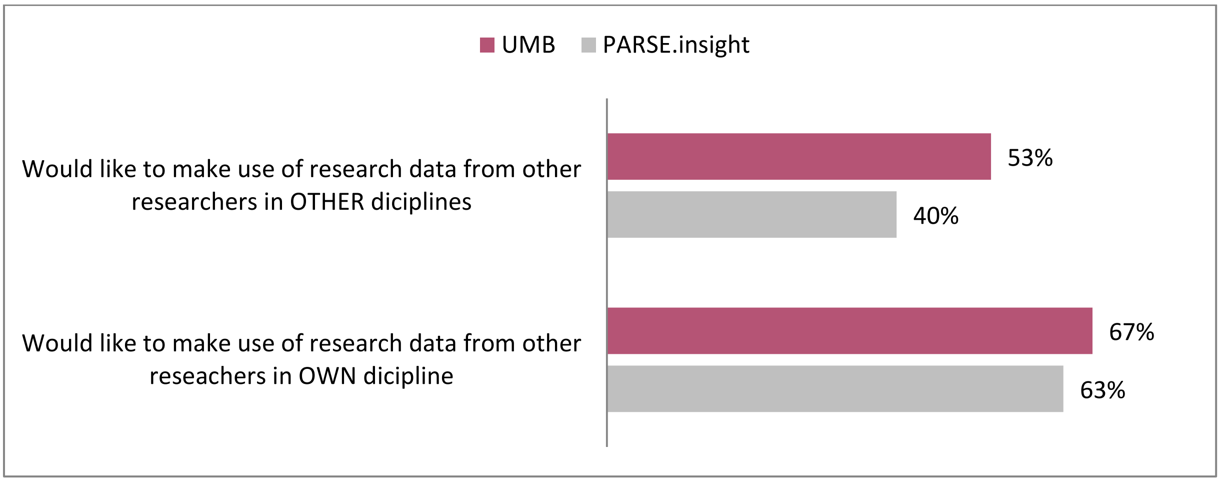


FIGURE 4.27: COMARRISON ON MOTIVATION TO REUSE DATA FROM OTHER RESEARCHERS

The differences points out that the unexplored potential for sharing is higher for the researchers at UMB than for their European fellows. It might be explained by going back to the results presented in a previous chapter where the new researchers at UMB are more positive than the experienced ones, as one might find the researchers from the PARSE.insight study to be more experienced than the researchers from UMB,

4.5.4 Reasons for preserving digital research data

When asked to range the importance of various arguments for the preservation of digital research data, the UMB researchers generally find all the different issues to be less important in the comparison to the researchers from the PARSE.insight study. The differences are illustrated in figure 4.28. As the questions here are identical in the two studies other variables must cause this difference. As mentioned earlier UMB is an average to small university in the Norwegian context, the institutions questioned by PARSE.insight are all leading in their fields in Europe. This could possibly affect how the researchers see the importance of saving the research data they collect, but it would be impossible to be sure of what causes this difference without further studies. The arguments ranging as most important for the researchers at UMB are “It allows for re-analysis of existing data”, “If the research is publicly founded, the results should become public property and therefore properly preserved” and “It will stimulate the advancement of science”. These all ranged as very important reasons by approximately one third of the respondents and as important reasons by about half of the respondents. Also “It may serve for validation purposed in the future” was found to be an important reason by more than half of the respondents. In the PARSE.insight study the same four reasons were found to be the most important ones. All four were rated as very important by about half of the respondents and as important by 35 to 45%. The only argument were the importance was rated about the same was “It potentially has economic value”. This ranged as the least important in both studies. The following figures present the differences in the data from the two studies.

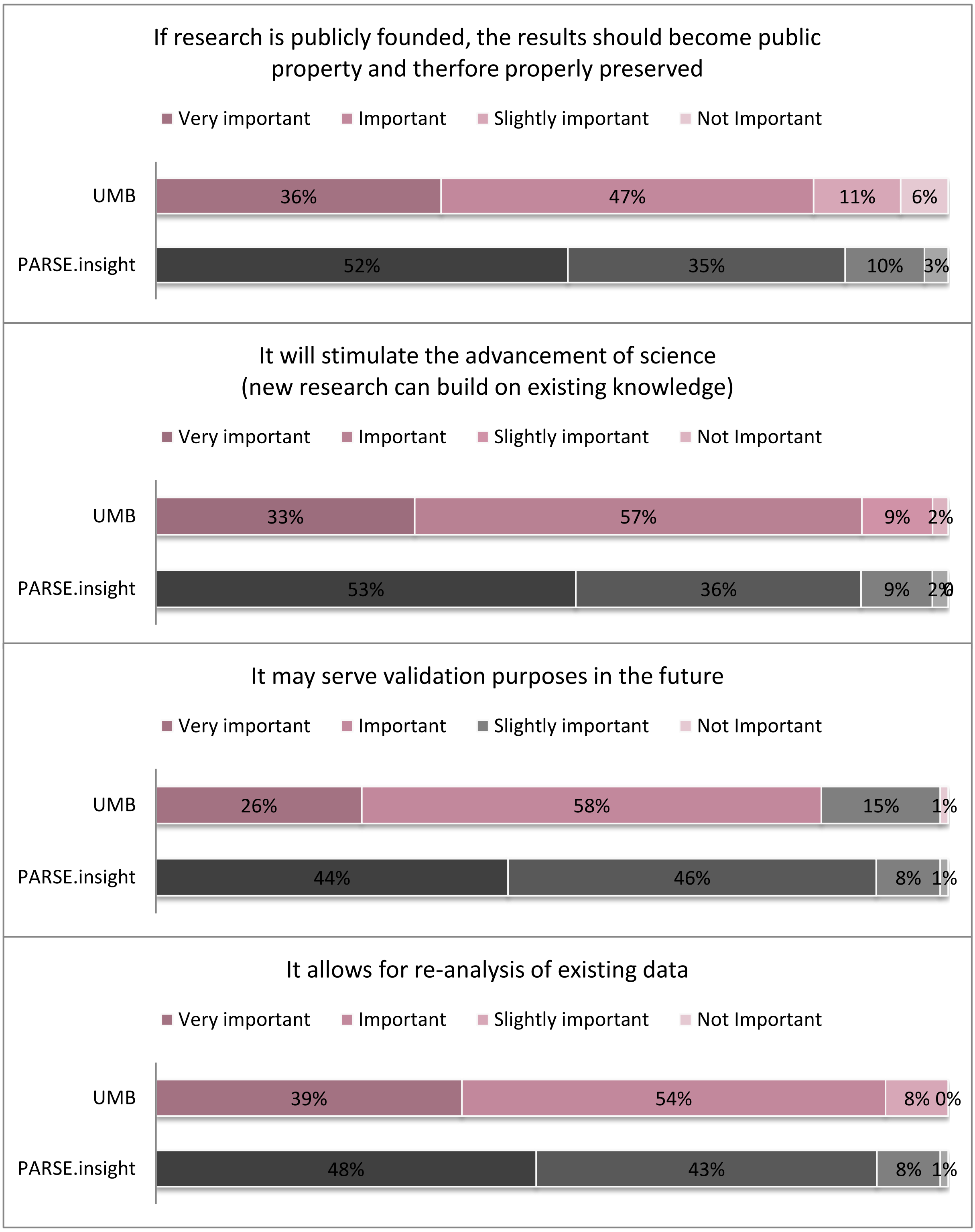


FIGURE 4.28: COMPARISON OF ARGUMENTS FOR PRESERVATION AND ACCESS TO RESEARCH DATA

4.5.5 Potential problems of data sharing

When the researchers were asked which problems they foresee in sharing their data, misuse of data and legal issues top both of the studies as illustrated in figure 4.29. There are however some differences as incompatible data types, lack of financial resources and technical infrastructure, all being more practical issues scored high in the PARE.insight study, while the third biggest problem foreseen by the researchers at UMB was the fear of losing the scientific edge of their research. In figure 4.29 the problems foreseen are ranged after importance according to the UMB study and the numbers form PARSE.insight are represented as a benchmark in grey.

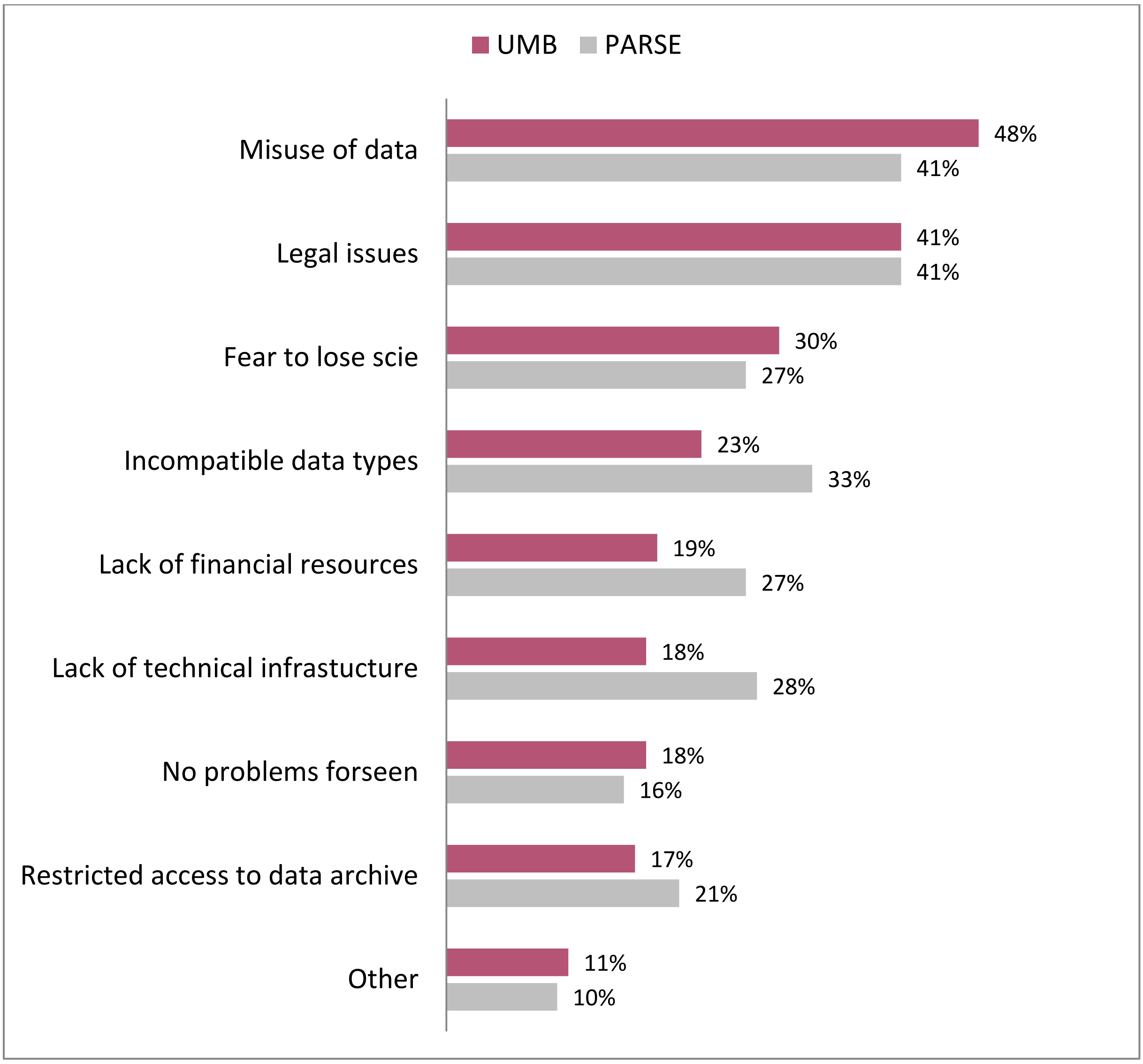


FIGURE 4.29: COMPARRISON OF POTENTIAL PROBLEMS IN SHARING OF RESAERCH DATA

The figure shows how the fears among the researchers questioned in the PARSE.insight study worry more about issues related to available resources and technical issues, while the UMB researchers are more focused on misuse and the scientific edge. In particular the last issue is curious as the institutions participating in the PARSE.insight study are known for high quality research and having the “scientific edge”, but the worry connected to this is less than the concern about technical infrastructure.

4.5.6 THREATS TO PRESERVATION OF RESEARCH DATA

Figure 4.30 shows that as with the reasons for preserving digital research data, the threats are given less importance by the UMB researchers. When the ratings, important and very important, are put together the differences are not that big, this repeats what is found in the previous comparison, that many of the same reasons are found important, but the level of importance differs in the two studies. Again it is also evident that the UMB researchers give less importance to the technical challenges such as sustainable hardware and long term preservation issues such as whether the current custodians of data cease to exist at some point. These two threats are regarded the most important ones by the researchers in question by PARSE.insight and are given a medium importance by the UMB researchers. Interestingly it is the fear that their own data will be misinterpreted or not understood that is considered the largest threat by the researchers at UMB followed by a fear that restrictions in access and use may not be respected in the future. It shows us that the concerns of the researchers asked by PARSE.insight are different from the concerns of the researchers at UMB, who are mainly concerned about “their own data” whereas the researchers questioned by PARSE.insight have the largest concern about the long time availability issues.

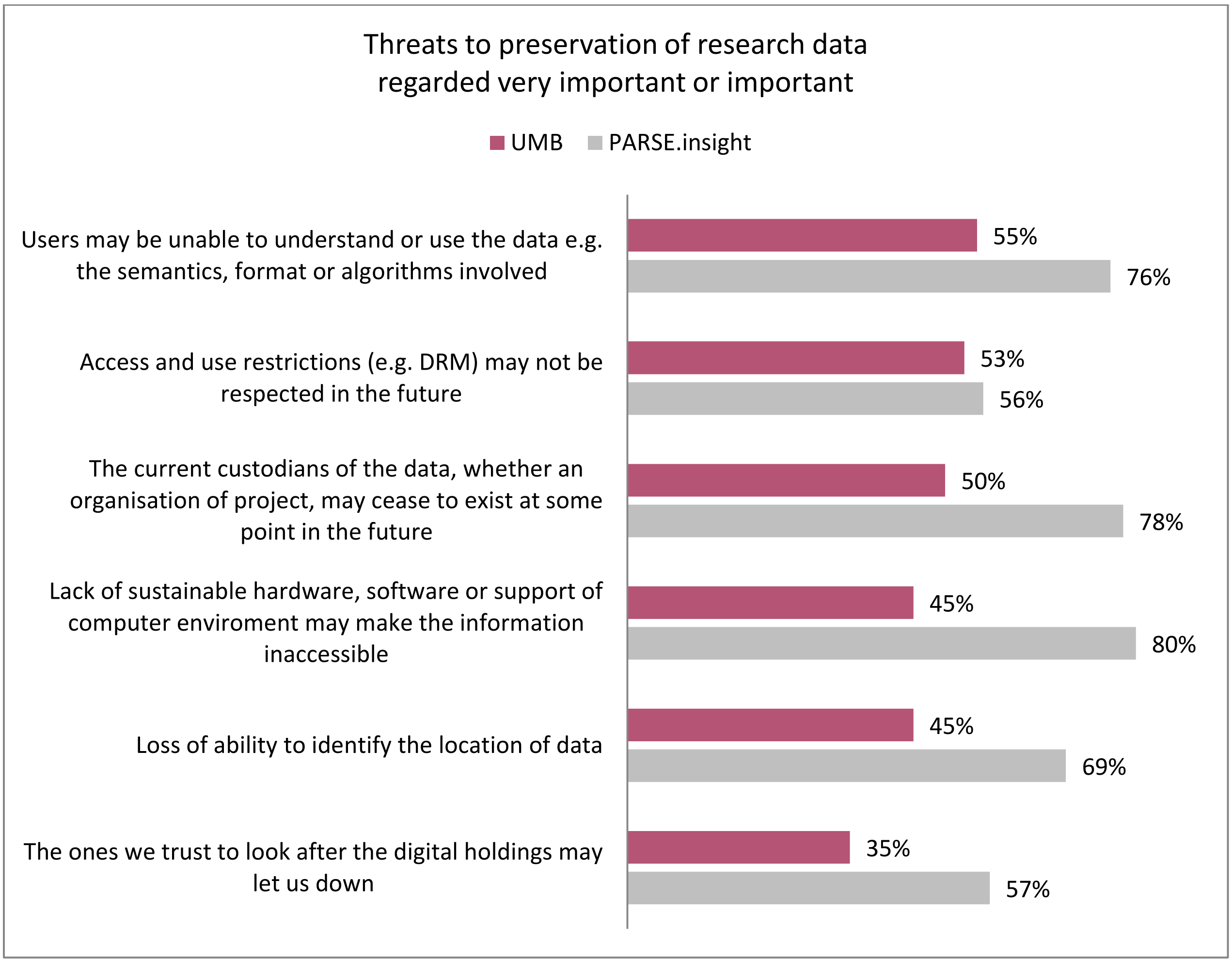


FIGURE 4.30: COMAPRRISON OF THREATS TO PRESERVATION OF DIGITAL RESEARCH DATA

4.5.7 The general differences between the findings at UMB and PARSE.insight

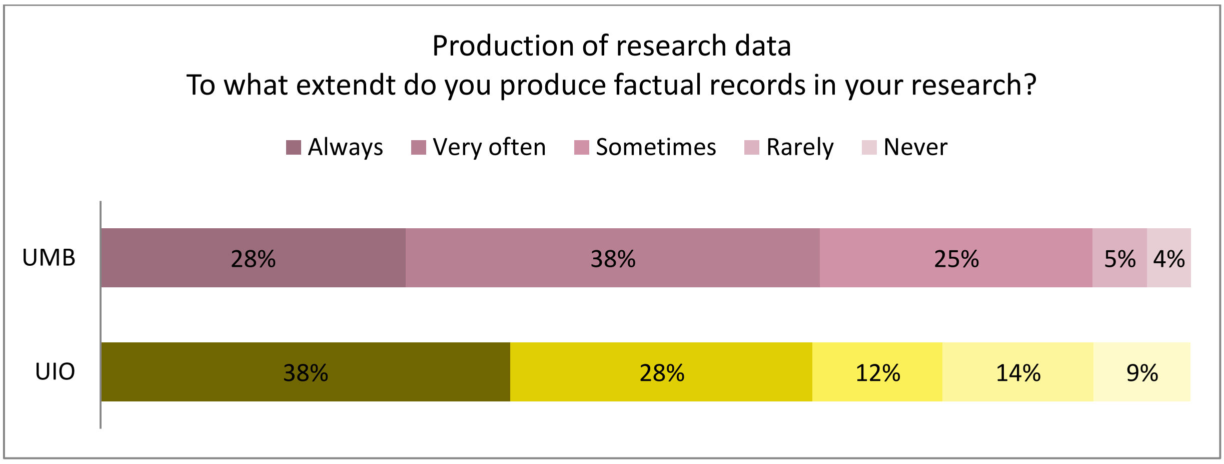
As described and illustrated above, the data collected at UMB differs from the data found in the study by PARSE.insight in many areas. The UMB researchers are more interested in reusing data from other researchers than the researchers questioned by PARSE.insight, but in all the other aspects the researchers form UMB show lower motivation than their fellow researchers at CERN, the Humboldt University and ESA. When asked about threats and potential problems the results are quite different. Not only do the UMB researchers express less concern or give less importance to the treats described than what is shown by the PARSE.insight study, it also emphasises different aspects. While the main concern of the researchers who answered the PARSE.insight study are towards practical issues such as hardware, economy and the challenge of long-term preservation, the researchers at UMB are mainly focused on the misuse of the data, whether access rights will be respected in the future, misinterpretation and legal rights regarding the research data. As such, one might say that the researchers from CERN, Humboldt University and ESA are more concerned about the superior issues of data preservation, while the researchers at UMB are more focused on their own data and how it can be misused or misinterpreted. The reason or reasons causing these differences are however not obvious. One possibility is the differences between the institutions, the PARSE institutions are the "top of their class", and so are the researchers working there. Being in this position and expressing less concern about the matters that would to some extent depend on them, such as the ability to give proper descriptions of the data for information not to be lost and data not to be misinterpreted can be seen as sort of self-confidence. They are possibly also more familiar with the issue of data sharing, in particular in the field of space observation the idea of sharing data came early, but the costs and the infrastructure needed to facilitate it, is at the same time extremely complex and expensive. When CERN speaks about proper storage and sharing of data (CERN mfl., 2009) it is with the perspective of the average time it takes to win a Nobel price, 25 years, also they are aware that the data gathered in many cases is not reproducible. This combination of uniqueness and prestige makes it reasonable to believe that the researchers are already dealing with the issue and challenge of data preservation in a different way than the researchers at UMB. Still this is again only an assumption based on the comparison of the data gathered in this survey. But as the comparisons within UMB also indicate differences in attitudes based on experience, it is the most obvious conclusion to draw of the differences found.

4.6 Comparison with questions used in a questionnaire at Oslo University

Some of the questions in the questionnaire were taken from a questionnaire sent out to researchers at the University of Oslo (UIO) in 2009 as a response to the OECD guidelines. The results were presented in an article in "Forsker Forum" in 2010 (Grøttum, 2010). As that questionnaire was in Norwegian, the questions had to be translated. Most of the questions by Grøttum used the Likert scale to express level of agreement or support of various issues making the translation more complicated, mainly as appropriate terms for the scale also had to be found. Due to this and the need to compress my questionnaire, some questions were reformulated into statements where the researcher would express level of agreement. This could have had an influence on the differences found in the responses for some of the questions. In one question the options given originally were found not to cover all evident alternatives and as a result two additional options where added. Other factors worth mentioning that could have affected the responses are differences in time, and in the disciplines of the researchers.

4.6.1 Production of relevant research data

One of the first questions in the questionnaire refers to the OECD guidelines and their definition of relevant research data, “factual records used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings” (Gurria, 2007) and asks to what extent the researcher produces this kind of data in their research. Apart from the translation, no modifications were made here. Figure 4.31 compares the responses from UIO and UMB. What we find then is that more researchers at UIO replied to “Always produce this kind of data” with 37% versus 28%. If we add the respondents replying “very often” the responses become however quite even. Different from UIO, UMB has a large group of respondents replying to produce this kind of data “sometimes”. Only 9% of the respondents from UMB reply that they “rarely” or “never” produce this kind of data versus 23% from UIO. This indicates that the number of researchers dealing with research data on a regular basis at UMB is slightly higher in comparison to UIO.



FIGUR 31. COMPARRISON OF PRODUCTION OF RESEARCH DATA FITTING THE OECD DESCRIPTION

4.6.2 Channels used to make data openly accessible

The researchers were also asked which channels were used in their field for making data openly accessible. When comparing with UIO it has to be noted that the researchers from UMB was given two options more: Own website, which in the UIO questionnaire was included in the institutions website, and “I don`t know”. When comparing the results in figure 32 it is notable that a high number of respondents form UIO replied that other channels are used. It is however likely to be that respondents selected this option when they were not certain of the answer, as 18% percent of the respondents from UMB replied that they don`t know. Further NSD (Norwegian Social science database) is more commonly used by respondents from UIO. UIO has in comparison with UMB, large faculties of Social sciences and Humanities, while UMB has more of a Life Science profile. What is however remarkable is that more than twice as many of the respondents from UMB replied that the journal/publisher web page is used. It is at this stage only possible to speculate on this big difference being caused by the difference in time, in the research fields represented or if other reasons are causing the journal/publishers to score high on the results from UMB, while only average from UIO. It would however be a possible issue to investigate further in later research.

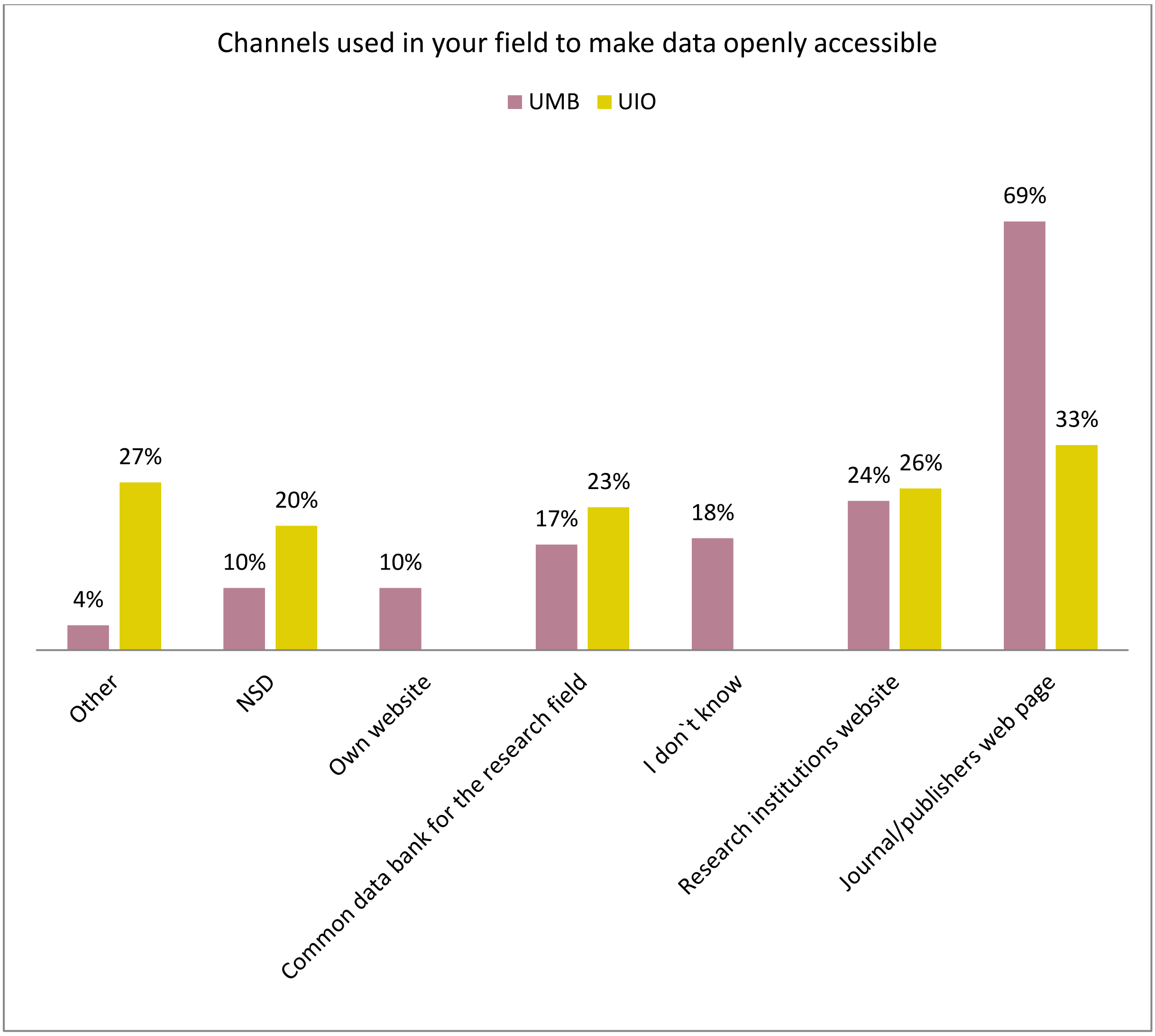


FIGURE 4.32: COMPARRISON OF CHANNELS USED TO MAKE DATA OPENLY ACCESSIBLE

As the figure shows publishers are more involved in the publication of research data for the researcher at UMB than UIO. The reason for this can probably be found by looking back at the difference found between disciplines, telling us that the publication of research data by publishers depend much on the subject.

4.6.3 Important issues for sharing data

The importance for the researcher to have exclusive rights to their research data in a temporary period is often pointed out in the literature, as one of the criteria`s the researcher underlines before agreeing to share their data. This is confirmed by the results from this questionnaire, and in comparison with UIO, shown in figure 33, it is clear that the researchers from UMB are slightly more focused on this aspect than their fellow researchers at UIO. The group finding this issue to be very important is however, as illustrated in the figure, about the same size.

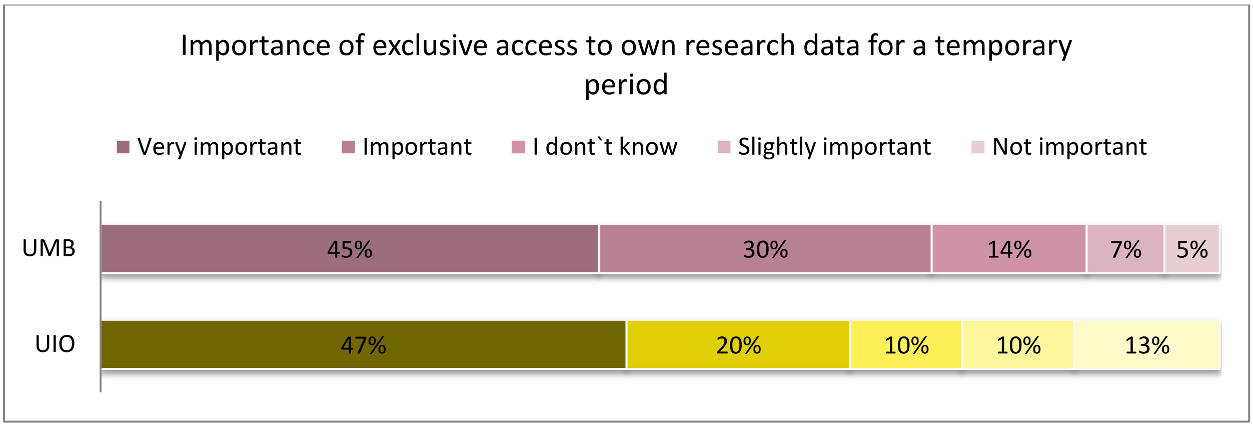


FIGURE 4.33: COMPARRISON OF ATTITUDES TOWARDS EXCLUSIVE RIGHTS

In order to discover what influence e-science would have on the different research fields, the researchers were asked to express their attitudes towards a statement regarding whether the sharing of research data would strengthen the research in their field. The formulations on the two questions ended up being slightly different, both as a result of translation and the need to compress the questions in a questionnaire that was already quite long. The results found in figure 34 show however an evident increase from the UIO questionnaire to the UMB questionnaire. While only 35% of the UIO respondents believed that sharing of research data would have a larger impact on strengthening the research in their field, 82% of the researchers from UMB agreed that sharing of research data would have this impact. This difference indicates that the researchers are of the same understanding as the Microsoft Research group claiming that sharing of data will have the strongest impact on the Natural Sciences. Still it is not something to state with 100% certainty as the differences partly can have been caused by differeneces in the formulations and the Likert scale when the translation was done. Still it gives strong reason to belive that the resaerchers in the Llife Sciences belive that sharing of research data wil have a stronger positive impact on their field than the average researchers from different diciplines represented at UIO.

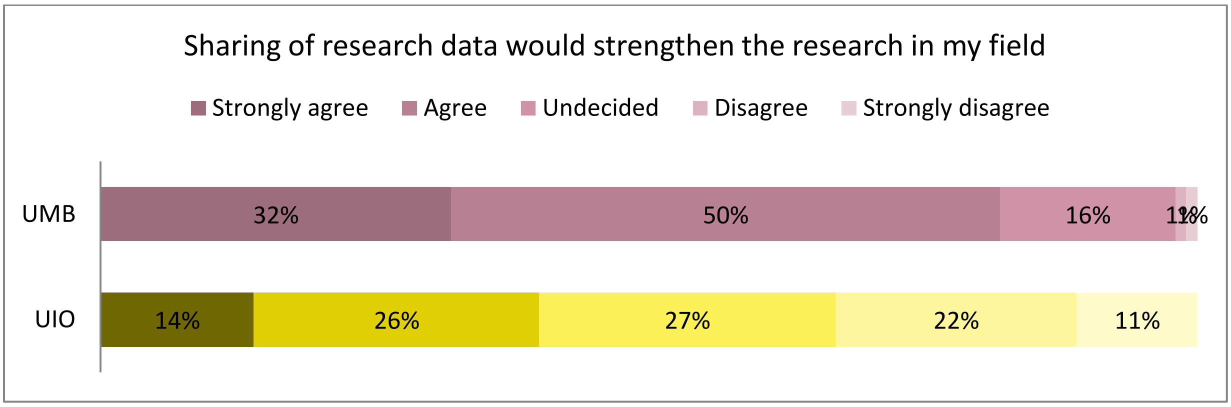


FIGURE 4.34: COMARRISON OF ATTITUDES TOWARDS THE IMPACT OF DATA SHARING

4.6.4 The general differences between the findings at UMB and UIO

When comparing the results from UMB to the results from UIO in 2009, there are some differences both regarding data production, attitudes towards data sharing and the channels used for data publication. As the universities have quite different profiles this could be one explanation, by investigating the participation of the different faculties in the UIO questionnaire it can however be identified that the two faculties with the larger participation on UIO are the faculty of medicine (29%), and the faculty of mathematics and natural sciences (26%). Third and fourth come the faculty of humanities (14%) and social sciences (12%). Another possible factor is the difference in time between the studies, from 2009 until 2012 in which much has happened in the field of data sharing and the raising of awareness might also affect the attitudes towards the issue. These are however only speculations made to raise awareness that various factors can impact, cause or explain the differences between the two universities.

5. DISCUSSION

In the following chapter the previous results and findings presented are discussed according to relevant issues or factors.

5.1 FINDINGS REGARDING EXPERIENCE WITH SHARING AND REUSE OF RESEARCH DATA

Most of the research data produced at the university is what the OECD standard describes as sharable data (Gurria, 2007). This shows us that there is a big potential for developing a structure for the exchange of research data. There are however barriers such as developing an infrastructure for sharing that can handle a large number of data formats in a proper and secure way and the need to train researchers in using metadata and metadata standards so that the context of the data is not lost. The results from this study show that a lot of sharing is already taking place, mainly through informal channels such as colleagues and personal networks. But there is also an unexploited potential for further sharing, as most of the researchers who do not currently reuse data from others would like to do so in the future. The differences here are large and depend on years of experience, indicating that it takes several years as a researcher before a functional network for data sharing is in place. As a result of not having the network of researchers in place, more of the researchers early in their careers, experience not being able to get access to the data they would need in their research.

Consequently most of the researchers do not apply adequate metadata to their data, although the exceptions are those who currently share their data openly and through this they are required to use metadata and often also apply metadata standards. When data is exchanged through colleagues the information needed to interpret the data is not necessarily written down and added to the research data as context, it is therefore more likely to get lost, leaving the research data more at risk for misinterpretation. As stated by Heuer at CERN: "When we talk about data preservation it is always data and metadata that is the important point, you need to preserve the data but also the knowledge" (CERN mfl., 2009). In the current system at UMB, the research data is stored, but the knowledge or metadata is not preserved with it, as the storage is not meant for reuse, retrieval or for the data to be understandable, or have a value on its own in the future, but is rather a structure imposed for future control in suspicion of data fabrication. The literature is clear on the fact that only trough stable, longtime preservation initiatives where metadata is stored along with the data for future retrieval and understanding can you have actual preservation of the data. Today the researchers store and retrieve research data through completely different platforms, both of them with a clear potential for improving if formalization and context is added to the data. In order to facilitate a structured exchange of research data, standards must be established and applied to the data. This issue seems to still be a challenge for the researchers who have little or no knowledge of metadata standards, and only a minority of the researchers have established routines for registering administrational or technical metadata to the research data they create. Standards for metadata are something only few researchers at the university have some familiarity with.

5.2 Findings regarding attitudes towards sharing and reuse of research data

The most important reasons found for data preservation are the possibility for reanalysis and possible stimulation for the advancement of science. At the same time the researchers fear misuse or misinterpretation and legal issues, along with the privacy of their informants. They also worry that a system for data preservation will require much of their time, which could have been spent on doing research. Compared with the responses from PARSE.insight it is found that the researchers questioned there worry more about practical issues, whereas the concerns for the researchers at UMB are more related to how others will deal with the data they collect. The reasons for these differences are not clear, they do however show different approaches to data sharing, one focused on the practical challenges of creating a repository, the other about juridical and ethical challenges related to making their data openly available. Still a large number of the UMB respondents believe that the building of a national infrastructure for data preservation would help to guard against some of these threats.

5.3 Years of experience as a factor

The results show that in particular within some areas, years of experience have an impact on how the researchers respond. The least experienced researchers are positive both about reusing data from their own discipline and from other disciplines; they are also strong supporters of the arguments for the preservation of research data. When it comes to sharing of research data under “Routines for own data storage and collaboration” the new researchers are most open, but those who have been doing research for between 5 and 10 years are the most skeptical. The researchers tend to get less skeptical as they gain more experience. This tendency is however difficult to explain. Is it is a curve that is constant, starting with research being open minded and positive to the possibilities of sharing, then getting closed with a “fear of people stealing their ideas” before they again become more open as they gain confidence as researcher. This could have to do with different generations having different attitudes towards the same issue independent of their work. Another possible assumption is however that it reflects the ambitions researchers hold at different stages in their careers.The comparison with the PARSE.insight study can also be used to some extent to illustrate a different place in their careers as researchers. The data from the questions where the differences are most evident, are however not presented in the Insight report, and therefore not available. The differences found in the comparable material is however that in the PARSE.insight study the researchers are more concerned about issues related to available economical resources and technical issues, while the UMB researchers are more focused on misuse, legal issues and the scientific edge. This difference can be caused by similar differences, the data is however not enough to explain, but only to point out differences, and by this release new questions.

In order to find a certain answer to what causes the researchers to express quite different opinions and habits according to years of experience, it would be necessary to do an investigation, either by locating studies on the issue of researchers' attitudes, or by gathering more data on this specific issue. This study shows however that there are differences both in how researchers work according to years of experience, and with regards to which attitudes they hold towards the sharing of their own research data:

-New researchers tend to use several channels for finding and accessing digital research

-The researchers most active in reusing data from others are those with 10 to 20 years of experience.

-Researchers with 5 to 10 years of experience are better at assigning metadata to their digital research than their colleagues.

-Half of the researchers with less than 5 years experience said that they would like to share their research data in the future

The first two issues are likely to be connected with the time it takes to establish a network of fellow researchers who can exchange data. New researchers are not so likely to have this established yet, and are therefore forced to search for data in various other channels, while more experienced researchers know where to look or who to contact in order to find the data they need, explaining the higher reuse of data among researchers with more than 10 years of experience.

It is however harder to explain why researchers with 5 to 10 years of experience are better at assigning metadata to their research data and it would be interesting to know if these are habits the researchers will take with them later in their career, and as such represents a change of habits among researchers, or if it is a habit they lay behind as they gain more experience. The questions dealing with concerns and fears towards data sharing shows that the fears among this group also differ to some extent from their colleagues at UMB, and at the same time the attitudes are most similar to the ones found in the PARSE.insight study. One can only say that hopefully these answers show a changing in the routines of assigning metadata among the researchers who are not familiar with an "analog" way of doing research. The explanation for why PhD students have not adapted the same routines might be that they are not yet dealing with the issue of preservation as they are still most likely in the middle of their first research project. The numbers also show that the PhD students are the ones most positive towards sharing their data. This tells us that the ambitions to share and provide open access are fairly high for this group and that to start up a data repository approaching the graduating PhD students and offer them a possibility to safely store their data, and get potential accreditation by future reuse, would be welcomed among this group.

These differences, according to years of experience, say something about changes in the way of working through their careers as researchers. It indicates that it takes more than 10 years to have an optimized collegial network for sharing, and until the researchers have this they search in various channels to find the data they need, often not finding it.

Researchers with 5 to 10 years of experience are the best ones for assigning metadata, still with a long way to go before standards for metadata are used among all of them. These researchers are likely to have all the research data they have collected in digital form and work methods adapted to digital platforms. As a result of this they experience a stronger need to assign metadata in order to navigate in their own research material. Their habits are far from optimized, something standards, routines and training can help with, still they are attempting to store information with their data, and by this also trying to store the context in which the data was collected. A possible recommendation would be to integrate data preservation into the PhD programs so that researchers would learn how to preserve data along with doing their first data collection as researchers. As the responses show, many of the PhD students are open to sharing their data, a sharing that can only be optimized with the usage of standard metadata formats.

5.4 Academic discipline as a factor

Even if UMB is a university focused on the Life Sciences, there are still huge variations in how the research is carried out, These differences between the sciences are also reflected in how the researchers respond to some of the questions. In order to best present these differences some of the departments that have results varying much from the media that is presented, discussed what kind of research they produce in the department.

5.4.1 IKBM

More than 70% of the researchers at IKBM have had requests from publishers to include relevant digital research data, while none of the respondents from INA have experienced this. This tells us not only that the researchers at IKBM most likely have more experience with data sharing, but also that the publishers used by these researchers find that research data somehow strengthens the impact or credibility of the research. Furthermore, the results from this study show that the IKBM researchers are better than any of their colleagues at assigning metadata to their research data. One of the researchers from IKBM even stated that the data he publishes will not be accepted if they do not fulfill existing standards. A quote that not only shows that the researcher is familiar with the value of metadata standards, but also that data publication is a used way of publishing research, follows. To better understand what type of research is done in this department a presentation form the IKBM web page is presented:

> *"In IKBM the researchers work with biotechnology and food science by integrating chemistry, biochemistry, microbiology, molecular biology, mathematics statistics and technology in order to create new knowledge" («Forskningen ved IKBM», 2012)*

These are all science disciplines that to a large extent are dealing with numerical data. In one of the first questions presented it came out that almost 60% of the respondents from IKBM produced raw data, making it the third most common data format in the department after office documents and images, as well as one of the departments that most commonly produces raw data in research.

5.4.2 Noragric

The department of International environmental and development studies (Noragric) presents an opposite scenario to IKBM, being the department at UMB with a clear social science approach. These differences were also found in the data collected, where the results form Noragric often stood out as different from the media. In the web pages of UMB the research done in Noragric is presented as the following:

> *"Noragric carries the academic responsibility for international environment and development studies at UMB and is the university's department of interdisciplinary and social sciences. Noragric is responsible for the coordination of UMB's institutional collaboration with countries in South and Eastern Europe" («Forskningen ved Noragric», 2012)*

Interdisciplinary work is mentioned in this description and is also found in the data as Noragric is the only department at UMB where data from other disciplines is reused more frequently that data from their own discipline. Among those who do not currently reuse data there is a will particularly to reuse data from their own discipline. Examples of interdisciplinary research can be found both in the development part, where researchers at Noragric use competencies from other disciplines to create development projects around the world. Further environmental studies is often mentioned as one of the disciplines where data sharing across disciplines is believed to have a great impact. Even if the researchers from Noragric are shown to be active in reusing data, they are not equally active in making their own data sharable by adding sufficient metadata. Most of the respondents do not add any kind of metadata to the research data they collect. This can possibly be connected to two different reasons: 1. As few researchers currently share their data, they are not forced to learn appropriate usage of metadata and 2. The form of much social science data such as data collected through interviews makes it to some extent understandable and also without metadata, making the adding of metadata less urgent for the researcher in order to be able to understand their own data in the future. When asked about the importance of having exclusive rights to own data for a temporary period almost 80% point this out as very important to them compared to the other departments who have responses around 40-50%.This indicates the importance for the researchers at Noragric to be able to use and publish from their data, before making it openly available to others.

These two departments are only two examples of how different traditions for sharing can be. Other departments that would have shown similar differences are INA and ILP. Noragric and IKBM stand however as good examples of the complexity of a university. This also illustrates how important sharing is in different ways for different disciplines. The researchers at Noragric are largely depending on finding available data from other disciplines, and as most sharing takes place through colleagues one can only imagine the effort it sometimes must take to establish a large collegial network for data exchange that goes beyond one discipline. The IKBM researcher often needs to make their data sharable in order to be able to publish their results. The issues and challenges of data exchange are in this way highly relevant to both departments; it does however affect the researchers differently. Still a support system for assisting the researchers in assigning appropriate metadata standards and using stable platforms for data exchange would, without doubt, be of great use for researchers at both of these departments, as well as other departments at UMB.

5.5 The Life Sciences as a factor, particularities for UMB

The comparisons with previous studies in UIO and by PARSE.insight indicate that some issues relate particularly to UMB. The comparison with PARSE.insight shows that the researchers at UMB are less concerned about technical and practical issues, but share the concern for copyright issues and potential misuse. It tells that the researchers at UMB have a strong faith in their technical staff to develop functional systems and maintain them well as well, as they have little concern about the costs and the possibility for founding. It might also reflect the amount of data that needs storage, the researchers at UMB was not asked how much data they produced but institutions like ESA and CERN produce huge amounts of data that naturally makes the costs a bigger concern than for a smaller university. The comparison with UIO shows however that more researchers at UMB produce what is described by OECD as sharable data. This is most probably a reflection of the sciences represented at UMB. Furthermore, the usage of journal/publishers web page for making data openly available stands out as a particularity for UMB being more than twice as common there as opposed to UIO. Also when it comes to believing that sharing of research data would have an impact on their research field, the UMB researchers are very positive. Together it makes it possible to conclude that researchers in the life sciences in Norway believe that sharing of research data would have a strong impact on their field. Already today much research data is shared through journals and publisher web pages, but as much of the researchers produce sharable data in their research, there is the potential for the sharing of more data.

5.6 Current sharing as a factor

As previously discussed in the results, IKBM shows experience with data sharing and this has a clear impact of the researcher’s usage of metadata and understanding of the importance of metadata standards. Among the researchers who share their data the usage of metadata is higher than for any other group as 75% assign some kind of metadata to their research data and more than 20% are familiar with some kind of metadata standard, twice as many as the media. This clearly illustrates the importance of metadata and metadata standards for the researchers to openly share their data and not only make it available through colleagues, as the information about the data needs to be added.

As the literature review showed, the metadata issue is found to be of great importance when sharing research data and the findings confirms this.

The researchers who currently share their data find the most important reason for sharing data to be that the research is publicly founded and the results therefore should become public property. The issue of property is complex both when it comes to research and when it comes to data, as several researchers argue that they have copyright on what they produce («Opphavsrett», 2012), even if it is produced with public founding. At the same time data is not protected by the copyright law as it is not found to be "creative work" but should be reproducible(«Åndsverksloven», 2009). Also databases are protected, and the grey zones, when it comes to data copyright, are many and complex. It therefore makes it important what the researchers themselves think about the rights to the data they collect and use in their research.

6. Conclusion

As the discussion above shows several factors have an impact on how the respondents relate to the different issues touched upon. In this chapter conclusions are drawn and put into the context of the research questions and the aims and objectives of this study. Furthermore, a set of recommendations based on the findings will be made, regarding both practical issues and the attitudes the researchers hold. Finally recommendations and suggestions for further research will conclude the chapter.

6.1 Conclusion to the research question

The aim of this study was to better understand how researchers deal with their own data and to what extent they reuse data from others in order to make recommendations for the optimization of these processes. It was formulated in the following research question:

What are the attitudes towards and experiences with sharing and reuse of scientific/research data among researchers in the life sciences in Norway?

As the research question has a double focus aiming both to detect the experiences and attitudes with sharing and reuse of scientific data, the conclusion to the research question is divided in the same way. Finally, a conclusion regarding which issues are particular to the Norwegian University of Life Sciences, is drawn.

6.1.1 Conclusions regarding current reuse of data

1. **There is a frequent reuse of data in the life sciences, and research data is primarily found through colleagues.**

As the discussion and analysis show there is a frequent reuse of data among researchers and data is primarily found through informal channels and social networks which again take time to establish. As younger researchers tend to use several channels in order to find the data they need, researchers with more experience use the networks they have to retrieve necessary data.

2. **The current reuse of data is far from optimized.**

The experience of not getting access to the research data they need is however a familiar problem for most researchers. Furthermore researchers who do not currently reuse data would like to do so.

Both of these findings lead to the conclusion that there is a clear potential for improving data sharing in the life sciences.

3. **For researchers who make their data openly available the principles of open science is their primary reason to do so.**

The responses from researchers who make their research data openly available today show that the reason these researchers find to be most important is that when the research is publicly funded, the results should become public property and therefore properly preserved. The principle of scholarly information as a public good is discussed in the literature presented by Borgman (2007, s. 35)as the core principle of open science.

4. **Most of researchers have little or no understanding of the usage of metadata.**

The results clearly show that most of the researchers have little or no understanding of the usage of metadata standards and about half the researchers do not assign any metadata at all to the research data they produce. The exception is researchers who have some experience with sharing the data they collect through formalized channels. As the literature shows, proper preservation of research data requires that information about the data must be stored along with it.

6.1.2 CONCLUSIONS REGARDING ATTITUDES TOWARDS SHARING AND REUSE OF SCIENTIFIC DATA

1. **Exclusive rights for first publication and accreditation for reuse must be assured for researchers to be willing to share the data they collect.**

Independent of discipline and years of experience, almost all the researchers find these two criteria to be crucial in order for them to be willing to share their data. In order to assure this system for publishing data along with publication it must contain functions for data citing. Further training of researchers in proper data citation must be done, so that the data producer gets the deserved accreditation.

2. **The attitude among the researchers towards making data openly available depends much on where the researchers are in their careers**

As discussed earlier the results clearly show different attitudes depending on years of experience. One theory presented is that the attitudes towards sharing change according to the ambitions the researchers hold. This has led to a search in the literature in the field of knowledge sociology in order to better understand how researchers relate to issues such as career and recognition. No clear description of changes in the ambitions depending on the career level has been found. However Robert Merton in his essay, “Behavior patterns of scientists”, describes the importance of early recognition of researchers to be as essential to discovery as intelligence (Robert King Merton, 1973, s. 327), and she further states that "science is crowded and accelerated. There is no sitting down alone anymore and letting apples fall down" (Robert King Merton, 1973). Both of these statements support the theory that the ambition to become famous is a key motivation for the researchers. It does not however explain further if there are certain steps in their careers where this is more important, and if this can be placed in context with changing attitudes towards sharing. The conclusion and findings here therefore remain that the attitudes towards sharing depends on where the researchers stand in their career, and I can only recommend others to investigate further the connection between attitudes and career level.

3. **Misuse or legal issues are the problems most of the researchers expect to meet in the future**

The findings show that the greatest concern the researchers at UMB held towards the future of data sharing was fear of misuse and legal issues, and compared to the PARSE.insight results, the UMB had little concern regarding financial resources, technical infrastructure or incompatible data types. This shows us that the UMB researchers are less concerned about practical issues, than the researchers at the institutions investigated by the PARSE group.

6.1.3 CONCLUSIONS REGARDING ISSUES PARTICULAR TO THE LIFE SCIENCES AND UMB

1. **Researchers in the life sciences believe that sharing of research data will have a much stronger impact on their field than researchers from other sciences.**

The comparison with UIO shows that the researchers at UMB are much stronger in their support of the claim that sharing of research data would strengthen the research in their field. This conclusion supports previously presented literature assuming that data sharing will have the strongest impact on the natural sciences, as this assumption is clearly shared by the researchers, and journals and publishers in these disciplines have a much stronger focus on including research data.

2. **The usage of journals and publishers web pages for making research data openly available is particular to the life sciences and UMB.**

The comparison with UIO shows that usage of journal and publishers' web page for making data openly available is more common in the research fields represented at UMB. As the usage of publishers were more than double, compared to UIO it tells us that publication of research data on journal/publishers web page is a peculiarity for the Life Sciences.

6.2 Recommendations

One of the aims of this research was to make recommendations on how to implement a repository for research data at the university. The types of recommendations that can be given are divided into two groups, the first one dealing with practical issues and the second one dealing with recommendations regarding the attitudes the researchers have towards the sharing of research data.

6.2.1 RECOMMENDATIONS REGARDING PRACTICAL ISSUES

The findings show that most of the researchers questioned had little or no experience with data sharing, those who did have experience with this however also had a better understanding of challenges and technicalities. In order to reach out and teach all the researchers the routines for preservation and storage of research data, a basic training and support system for usage of metadata, and preferably metadata standards, must be put in place. The libraries hold unique expertise on the field of metadata and therefore make a natural collaboration partner when establishing data repositories. Establishing of such repositories does however require the librarians, the researchers and the computer scientist at the university to work together in the phase of establishment. Due to differences between the sciences in the data collected it does make sense to make discipline specific repositories. The potential in combining data from different sciences is however a strong argument for establishing a national structure after the ANDS model from Australia. In Norway Cristin, already being responsible for the joint portal to institutional repositories in Norway Nora, makes the most obvious candidate for such a national responsibility for a joint search on research data.

6.2.2 RECOMMENDATIONS REGARDING ATTITUDES TOWARDS SHARING AMONG THE RESEARCHER

As previously stated there is a great concern among the researchers regarding legal issues. By consulting specialists on copyright and creating clear guidelines for the researchers on how to deal with research data with regards to copyright, these fears can be limited. Further appropriate licenses should be developed so that the researchers themselves can regulate how the data they collect can be reused. To cope with the concern of misuse, clear lines regarding copyright and the judicial aspect, along with guidelines for how to make the data understandable in the future and the issue of metadata being presented under practical issues are all important.

6.3 Directions for future research

Data sharing and e-science librarianship is a growing research field, still with space for several studies to be conducted. In this study the perspective has been mostly from the researchers and their role in data sharing, in order to gain a better understanding of how the researchers work and share information. Theories from other sciences can be applied. Then in particular, other studies of researchers from the sociology and sociology of knowledge were looked at to better understand how the career steps affect the researchers' attitudes. Also the norm of reciprocity from social psychology can be applied for further investigation of the mutual benefit of sharing. The field of data sharing and the principles it is based on makes a natural connection with the libraries, being promoters of open access to research and knowledge for centuries. To improve the understanding of how these principles stand in the research communities would to some extent also indicate the importance of the academic library as an institution.

As research data repositories are still being developed and implemented, research detecting what works and what does not would be of great use for those having plans of making one. There is a need to present studies of best practice and failures in Norway, as no such studies have been done up to this point. Collecting and systemizing expertise from institutions with experience in data sharing is a necessary first step for others to follow.

Also material adding to the debate of data curation in general is needed as this is a field with growing importance and with several visions, but less concrete data on effects. Questions arising here are: How can data citation be applied and work? Do data repositories improve the reuse of data? And is data made available in repositories being reused for other purposes? Hackathons arranged to promote the reuse of public data are gaining popularity all over the world. The question as to what extent research data is used in prototypes developed at these events? Does data sharing promote global collaboration on research? As these questions indicate, there are a large number of angles to take for further investigation of research data sharing.

REFERENCES

Allard, S. (2012). DataONE: Facilitating eScience through Collaboration. Journal of eScience Librarianship, 1(1), 3. doi:10.7191/jeslib.2012.1004

Association of Research Libraries (ARL) :: Unpacking the NSF Requirement. (2010). Retrieved May 2012, from http://www.arl.org/rtl/eresearch/escien/nsf/nsfbackground.shtml

Babbie, E. R. (2007). The Practice of Social Research. Thomson Wadsworth.

Baker, K. S., & Bowker, G. C. (2007). Information ecology: open system environment for data, memories, and knowing. Journal of Intelligent Information Systems, 29(1), 127–144.

Borgman. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, Massachusetts: MIT Press.

Borgman, C. L. (2011). Local or global? making sense of the data sharing imperative. Presented at A decade in internet time: symposium on the dynamics of the internet and society, University of California, Los Angeles. Retrieved May 2012, from http://works.bepress.com/cgi/viewcontent.cgi?article=1254&context=borgman

Castells, M. (2010). The rise of the network society: with a new preface. The information age: economy, society, and culture (Bd. 1). Chichester: Wilwey-Blackwell.

CERN, Le Diberder, F., Heuer, R., Mele, S., & Diaconu, C. (2009, December 7). CERN Document Server: Symposium on Data Preservation in HEP. Streaming video. Retrieved February 4, 2012, from http://cdsweb.cern.ch/record/1227010

Choudhury, S. G. (2008). Case Study in Data Curation at Johns Hopkins University. Library Trends, 57(2), 211–220. doi:10.1353/lib.0.0028

Choudhury, S. G. (2009). E-Science at John Hopkins University. I D. B. Marcum & G. George (Eds.), The Data Deluge: Can Libraries Cope with E-Science? (1. ed., p. 93–98). Santa Barbara, California: Libraries Unlimited.

Creamer, A., Morales, M. E., Crespo, J., Kafel, D., & Martin, E. R. (2012). An Assessment of Needed Competencies to Promote the Data Curation and Management Librarianship of Health Sciences and Science and Technology Librarians in New England. Journal of eScience Librarianship, 1(1), 4.

Davenport, T. H. (1997). Information Ecology: Mastering the Information and Knowledge Environment (1. ed.). New York: Oxford University Press, USA.

Dixon, P. (2011, July 25). Fulfilling potential - the library in the digital age. Lecture presented at Digital library learning lecture in users and usage of digital libraries, Settignano.

Dreyer, M., Bulatovic, N., Tschida, U., & Razum, M. (2007). eSciDoc–a Scholarly Information and Communication Platform for the Max Planck Society. German e-Science Conference, Seq.

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata Principles and Practicalities. D-Lib Magazine, 8(4). doi:10.1045/april2002-weibel

e-Science Portal for New England Librarians - a librarian´s link to e-science resources. (2010, 2012).e-Science Portal for New England Librarians. Retrieved May 10, 2012, from http://esciencelibrary.umassmed.edu/

European Strategy Forum on Research Infrastructures. (2010). Strategy report on research infrastructures - Roadmap 2010 (p. 23). Retrieved from http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf#view=fit&pagemode=none

European Union. (2010). Riding the wave - How Europe can gain from the rising tide of scientific data - Final report of the High Level Expert Group on Scientific Data A submission to the European Commission (p. 38). Hentet fra http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Forskning ved UMB | Forskningen ved Institutt for internasjonale miljø- og utviklingsstudier (Noragric). (2012). UMB. Retrieved May 12, 2012, from http://www.umb.no/forskning/artikkel/forskningen-ved-institutt-for-internasjonale-miljo-og-utviklingsstudier-noragric

Forskning ved UMB | Forskningen ved Institutt for kjemi, bioteknologi og matvitenskap (IKBM). (2012). UMB. Retrieved May 12, 2012, from http://www.umb.no/forskning/artikkel/forskningen-ved-institutt-for-kjemi-bioteknologi-og-matvitenskap-ikbm

Forskningsrådet. (2008). Verktøy for forskning - Nasjonal strategi for forskningsinfrastruktur (2008 - 2017) (p. 72). Oslo.

Friedlander, A. (2009). Head in the clouds and boots on the ground: science, cyberinfrastructure, and CLIR. I D. B. Marcum & G. George (Eds.), The Data Deluge: Can Libraries Cope with E-Science? (1. ed., p. 77–90). Santa Barbara, California: Libraries Unlimited.

Galaxy Zoo: Hubble. (2010). Retrieved May 7, 2012, from http://www.galaxyzoo.org/

Gastinger, A. (2012a). Challenges of e-science and virtual research environments for academic libraries in Norway (p. 125–128). Presented at Bobcatsss Amsterdam 2012, Bad Honnef, Germany: Bock+Herchen Verlag.

Gastinger, A. (2012b, mars 21). E-Science og forskningsinfrastruktur - hvorfor fagbibliotekene bør engasjere seg. Presentert på det 73. norske bibliotekmøtet, Stavanger. Retrieved from http://www.bibsys.no/files/out/bibliotekmoter/2012/presentasjoner/arr4_gastinger.pdf

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. ACM SIGMOD Record, 34(4), 34–41.

Greener, I. (2011). Designing Social Research: A Guide for the Bewildered. Sage Publications Ltd.

Grøttum, P. (2010). Deling av forskningsdata : Universitetsforskerne ble ikke spurt. Forsker Forum, 41(3), 40.

Gurria, A. (2007). OECD Principles and guidelines for access to research data from public funding. OECD. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Harms, P., Smith, K., Aschenbrenner, A., Pempe, W., Hedges, M., Roberts, A., Ács, B., mfl. (2011). The Quality in Quantity-Enhancing Text-based Research. Data Driven E-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010), 265.

Hey, T., & Hey, J. (2006). e-Science and its implications for the library community. Library Hi Tech, 24(4), 515–528.

Interim insight report - first insights into digital preservation of research output in Europe. (2009). (p. 20). Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf

Jankowski, N. W. (Eds.). (2009). E-Research: Transformation in Scholarly Practice (1. ed.). New York: Routledge.

Kowalczyk, S. T., & Shankar, K. (2011). Data sharing in the sciences. Annual Review of Information Science and Technology (Bd. 45). Information Today, Inc.

Lesk, M. (2008). Recycling Information: Science Through Data Mining. International Journal of Digital Curation, 3(1). Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/71/50

Lippincott, J. K. (2009). Library and information technology support of e-science in the western context. I D. B. Marcum & G. George (Eds.), The Data Deluge: Can Libraries Cope with E-Science? (1. ed., p. 63–76). Santa Barbara, California: Libraries Unlimited.

Lougee, W., Chowdhury, S., Gold, A., Humphrey, C., Humphreys, B., Luce, R., Lynch, C., et. al. (2007). Agenda for developing E-science in research libraries - final report and recommendations to the scholarly communication steering committee, the public policies affecting research libraries steering committee, and the research, teaching, and learning steering committee. (s. 26). Retrieved from http://www.arl.org/bm~doc/ARL_EScience_final.pdf

Lougee, W. P. (2009). E-science and research libraries: an agenda for action. I D. B. Marcum & G. George (Eds.), The Data Deluge: Can Libraries Cope with E-Science? (1. ed.). Santa Barbara, California: Libraries Unlimited.

Luce, R. E. (2009). Grand challenges and new roles for the twenty-first-century research library in an era of e-science. I D. B. Marcum & G. George (Eds.), The Data Deluge: Can Libraries Cope with E-Science? (1. ed., p. 3–16). Santa Barbara, California: Libraries Unlimited.

Lynch, C. (2007, august). The Shape of the Scientific Article in The Developing Cyberinfrastructure. CTWatch Quarterly. Retrieved March 29, 2012, from http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/

Marcum, D. B., & George, G. (Eds.). (2009). The Data Deluge: Can Libraries Cope with E-Science? (1. ed.). Santa Barbara, California: Libraries Unlimited.

Merton, Robert K. (c1979). The sociology of science: an episodic memoir. London: Feffer & Simons.

Merton, Robert King. (1973). The sociology of science: theoretical and empirical investigations. Chicago: University of Chicago Press.

Microsoft Research. (2005). Towards 2020 Science (p. 86). Retrieved from http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., mfl. (2000a). Collection-Based Persistent Digital Archives - Part 1. D-Lib Magazine, 6(3). doi:10.1045/march2000-moore-pt1

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., et. al. (2000b). Collection-Based Persistent Digital Archives - Part 2. D-Lib Magazine, 6(4). doi:10.1045/april2000-moore-pt2

National e-Science Centre definition of e-Science. (2012). Retrieved February 10, 2012, from http://www.nesc.ac.uk/nesc/define.html

Neylon, C. (2011a). Building the perfect data repository…or the one that might get used. Science in the Open. Blog. Retrieved May 9, 2012, from http://cameronneylon.net/blog/building-the-perfect-data-repository-or-the-one-that-might-get-used/

Neylon, C. (2011b, November 22). I need to publish more and read less! How new platforms will enable you to publish more effectively while reducing information overload. Presented at The 6th Munin conference, Tromsø. Retrieved from http://www.ub.uit.no/baser/ocs/index.php/Munin/MC6/paper/view/14

Neylon, C. (2012). They. Just. Don't. Get. It…. Science in the Open. Blog. Retrieved May 10, 2012, from http://cameronneylon.net/blog/they-just-dont-get-it/

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. Organization Science, 5(1), 14–37.

NSF. (2012, February). Cyberinfrastructure for the 21st Century Science and Engineering - Advanced Computing Infrastructure - Vision and strategic plan. National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf

Opphavsrett - en introduksjon. (2012).torvlund.net. Blog. Retrieved May 13, 2012, from http://www.torvund.net/index.php?page=opph-innl

Pickard, A. J. (2007). Research Methods in Information. London: Facet Publishing.

Piorun, M. E., Kafel, D., Leger-Hornby, T., Najafi, S., Martin, E. R., Colombo, P., & LaPelle, N. R. (2012). Teaching Research Data Management: An Undergraduate/Graduate Curriculum. Journal of eScience Librarianship, 1(1), 8. doi:10.7191/jeslib.2012.1003

Research Council of Norway. (2012a). Tools for research - Part I - Norway´s national strategy for research infrastructure 2012-2017 (p. 20). Oslo.

Research Council of Norway. (2012b). Tools for research - part II - Norwegian roadmap for research infrastructure 2012 (p. 12). Oslo.

Ringdal, K. (2007). Enhet og mangfold: samfunnsvitenskapelig forskning og kvantitativ metode. Bergen: Fagbokforl.

Robson, C. (2002). Real world research: a resource for social scientists and practitioner-researchers. Oxford: Blackwell.

Scientific Data Management at UIUC. (2012). Retrieved April 3, 2012, from http://dais.cs.uiuc.edu/dais/sdm/sdm.php

Simberloff, D., Barish, B., Droegemeier, K., Etter, D., Fedoroff, N., Ford, K., Lanzerotti, L., et. al. (2005). Long-Lived Digital data collections: enabling research and education in the 21st century (p. 85). Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf

Soehner, C., Steeves, C., & Ward, J. (2010). E-Science and Data Support Services A Study of ARL Member Institutions.

Stallman, R., & Free Software Foundation (Cambridge, Mass.). (2002). Free software, free society : selected essays of Richard M. Stallman. Boston MA: Free Software Foundation.

Szigeti, K., & Wheeler, K. (2011). Essential Readings in e-Science. Retrieved February 10, 2012, from http://www.istl.org/11-winter/internet2.html

Taxt, R. E. (2011). Åpenhet i forskning Hva, Hvorfor og hvordan? Presented at IT-forum UiB, Universitetsbiblioteket i Bergen. Retrieved from http://webcache.googleusercontent.com/search?q=cache:F7BoJW0W8LUJ:www.ub.uib.no/felles/dok/2011/IT_forum_Taxt_070411.ppt+%C3%A5pne+forskningsdata&cd=4&hl=no&ct=clnk&gl=no

Thaesis, & van der Hoeven, J. (2010). Insight into digital preservation of research output in Europe.

The Fourth Paradigm Blog | Nature Publishing Group. (2010). Blog. Retrieved March 22, 2012, from http://blogs.nature.com/fourthparadigm/

Walters, T. O. (2009). Data curation program development in US universities: The Georgia Institute of Technology example. International Journal of Digital Curation, 4(3).

Wright, A. (2007). Glut: Mastering Information Through The Ages (annotated edition.). Joseph Henry Press.

Åndsverkloven Lov om opphavsrett til åndsverk m.v. (2009).Lovdata.no. Norwegian laws. Retrieved may 13, 2012, from http://lovdata.no/cgi-wift/wiftldles?doc=/app/gratis/www/docroot/all/nl-19610512-002.html&emne=opphavsrett*&

APPENDIX 1. INVITATION LETTERS

SUBJECT: EXPRESS YOUR OPINION ABOUT OPEN RESEARCH DATA (ORD) - MASTER STUDENT FROM HIOA TRYING TO COLLECT THE OPINIONS ON ORD AT UMB

Hi

My name is Live Kvale, I am master student of Digital Libraries at HIOA (http://dill.hioa.no/) I am interested in the future of E-Science. In my thesis I will investigate the attitudes towards open sharing of research data among researcher at UMB, I would therefore ask you kindly to spend 15 minutes on answering my questionnaire: [LINK]

In December I did one month internship at UBMB, and was then invited by Andreas Åkerstrøm to present my research project and questionnaire at the meeting in FON 12.01.12. The slides from the presentation can be found her: http://prezi.com/dn6pvqr-eywr/apne-forsknings-data/

Thank you for your time and do not hesitate to contact me if you have any comments or questions.

Best regards

Live Kvale

livekvale@gmail.com

SUBJECT: REMINDER! EXPRESS YOUR OPINION ABOUT OPEN RESEARCH DATA (ORD) - MASTER STUDENT FROM HIOA TRYING TO COLLECT THE OPINIONS ON ORD AT UMB

Hi

Two weeks ago I sent out and email asking you to answer on a questionnaire about research data, many have answered and I would like to thank those who answered for their time. However I would love to have some more responses, you are probably all aware of that the responds-rate matter.

So those of you who did not answer last time would you please spend 15 minutes on my questions? And those of you who did answers, could you encourage your next-door researchers to answers as well? (next time you run into them over lunch or in the corridor)

[LINK]

Thank you for your time and do not hesitate to contact me if you have any comments or questions.

Best regards

Live Kvale

livekvale@gmail.com

SUBJECT: THANKS FOR PARTICIPATING IN QUESTIONNAIRE ABOUT OPEN RESEARCH DATA (ORD) - MASTER STUDENT FROM HIOA TRYING TO COLLECT THE OPINIONS ON ORD AT UMB

Thanks to all those of you who took time to answer my questionnaire!

[LINK]

The questionnaire will be closed in a five days after being active for 1month collecting opinions from about 20% of the researchers and PhD students selected as my sample. If you are curious to know more about my thesis I am keeping a blog as a notebook at http://sharingandaccess.wordpress.com/. Here you will find both the results from the questionnaire and my thesis as soon as it is ready.

Thank you all again for participating and do not hesitate to contact me if you have any comments or questions.

Best regards

Live Kvale

livekvale@gmail.com

APPENDIX 2. THE QUESTIONNAIRE

Open Research data

The questionnaire is built on a questionnaire made by PARSE.insight to discover researches needs for longtime data storage and has been modified for use on UMB.

The questionnaire is in English but you are free to answer in Norwegian on the open questions

Thank you for participating!

Live Kvale

Your identity will be hidden

Read about hidden identity. (Opens in a new window)

1) Which department/center are you connected to?

Dept. of Animal and Aquacultural Sciences, IHA

Dept. of Chemistry, Biotechnology and Food Science, IKBM

Dept. of Ecology and Natural Resource Management, INA

Dept. of Economics and Resource Management, IØR

Dept. of Landscape Architecture and Spatial Planning, ILP

Dept. of Mathematical Sciences and Technology, IMT

Dept. of Plant and Environmental Sciences, IPM

Dept. of International Environment and Development Studies, Noragric

Aquaculture Protein Centre, APC

Animal Production Experimental Centre, SHF

Centre for Plant Research in Controlled Climate, SKP

Centre for Continuing Education, SEVU-

Centre for Land Tenure Studies

Centre for Integrative Genetics, Cigene

Norwegian Centre for Bioenergy Research

Other, please specify:

2) How many years have you been involved in research?

< 5 years

5 - 10 years

10 - 20 years

> 20 years

3) What is your current role in research? (multiple answers possible)

I am a PhD candidate

I am a researcher

I am a lecturer with some research tasks

I am a research group leader or manager

I am a research director

Other (please specify)

The OECD guidelines define the relevant research data as "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings".

4) To what extend to you produce this kind of data in your research?

Never - Rarely - Sometimes - Very Often - Always

5) Please indicate which of the following digital research data you produce: (multiple answers possible)

Standard office documents (text/word documents, spreadsheets/excel, presentations/ppt)

Network-based data (web sites, e-mail, chat history, etc.)

Databases (DBASE, MS Access, Oracle, MySQL, etc.)

Images (JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc.)

Structured graphics (CAD, CAM, 3D, VRML, etc.)

Audiovisual (multimedia) data (WAVE, MP3, MP4, Flash, etc.)

Scientific and statistical data formats (SPSS, FITS, GIS, etc.)

Raw data (device specific output)

Plain text (TXT in various encodings)

Structured text (XML, SGML, etc.)

Archived data (ZIP, RAR, JAR, etc.)

Software applications (modelling tools, editors, IDE, compilers, etc.)

Source code (scripting, Java, C, C++, Fortran, etc.)

Configuration data (parameter settings, logs, library files)

Blogg, Wiki, twiter or other web 2.0 tools

PDF documents

Other (please specify)

Some research areas has own traditions for making data available on different places, please indicate what is valid for your field.

6) Which channels are used in your research field to make research data openly accessible?

NSD (Norsk samfunnsvitenskaplige datatjeneste)

Journal/publishers web page

Common data bank for the research field

Own website

Research institutions website

Other (please specify)

I don't know

7) Please express your level of agreement with the following statements:

Strongly Agree - Agree - Undecided - Disagree - Strongly Disagree

7.1 To have exclusive access to my own research data for a temporary period is important

7.2 Sharing of research data would strengthen the research in my field

7.3 The researcher needs to have a possibility to regulate how the research data can be (re)used commercially by usage of licenses

7.4 Open access to research data strengthens the credibility of the research

8) If you have additional comments to the statements in question 7 please add them here:

9) Which of the following standards or guidelines that are used in digital preservation are you familiar with? (multiple answers possible)

OAIS (Open Archival Information System)

Dublin Core

PREMIS (Preservation Metadata: Implementation Strategies)

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)

OAI-ORE (Open Archives Initiative Object Reuse and Exchange)

METS (Metadata Encoding and Transmission Standard)

NISO (National Information Standards Organization)

None of the above

Other (please specify)

10) Do you as a rule assign any additional information (metadata) to your digital research data? (multiple answers possible)

Administrative information (e.g. creator, date of creation, filename, provenance)

Technical information (e.g. encoding type, description, file format, settings, software utilities)

No

Other (please specify)

11) Are existing standard important for you to share your research data?

12) Which of the following applies to the digital research data of your current research: (multiple answers possible)

My data is openly available for everyone.

My data is openly available for my research discipline.

My data is openly available for my research group / colleagues in research collaboration.

Access to my data is temporarily restricted.

My data is available for a fee.

My data could be made available with appropriate changes (e.g. anonymous clinical data)

I do not share my data, but I would like to do so in the future.

I do not share my data and I do not want to share it in the future.

13) Regarding your current research, do you collaborate with researchers from other projects?

Yes, with researchers from other projects in my discipline

Yes, with researchers from other projects outside my discipline

Yes, both within my discipline and cross-disciplinary

No

14) Increasingly, awareness is growing that data should be shared as well as publications. Do you experience or foresee any of the following problems in sharing your data? (multiple answers possible)

Fear to lose scientific edge

Incompatible data types

Restricted access to data archive

Legal issues

Lack of technical infrastructure

Misuse of data

Lack of financial resources

No problems foreseen

Other (please specify)

15) Do you presently make use of research data gathered by other researchers in your discipline?

Yes - No

16) Do you presently make use of research data gathered by other researchers in OTHER disciplines?

Yes - No

The following criteria must be fulfilled for this question to be shown:

(If Do you presently make use of research data gathered by other researchers in your discipline? equals No)

17) Would you like to make use of research data gathered by other researchers in your discipline?

Yes - No

The following criteria must be fulfilled for this question to be shown:

(If Do you presently make use of research data gathered by other researchers in OTHER disciplines? equals No)

18) Would you like to make use of research data gathered by researchers from OTHER disciplines?

Yes - No

19) How do you locate and access digital research data? (multiple answers possible)

Via colleagues

Via institutional database and search facilities

Via general search engines (Google, Yahoo, etc.)

Via data centers or archives (World Data Center, DANS, National Archives, etc.)

Via social media and online networks

Via formal literature (articles and book chapters)

N/A (I do not use data from others)

Other (please specify)

20) Did you ever need digital research data gathered by other researchers that was not available?

Yes - No - I don't know

21) How do you publish your research results usually? (multiple answers possible)

Books

Journal of publisher (available by subscription)

Open access journal

Institutional repository/open access archive

Grey literature (e.g. technical reports, white papers, preprints)

Conference proceedings

Websites (e.g. Research Blogging)

Other (please specify)

22) Do you think it is useful to link underlying digital research data with formal literature?

Yes - No

23) Do you want to be credited when your underlying digital research data is used by others?

Yes - No

24) Do journals to which you typically submit your work require you to include relevant digital research data (i.e. data used to create tables, figures, etc.)?

Yes - No

25) Online research colaboratories are virtual collaborations between researchers for sharing research data and new insights, or for debating about hot topics in science. Do you make use of an online research colaboratory?

Yes - No

26) Do you think that the information generated by these online colaboratories should be preserved (e.g. chat logs, wiki´s)

Yes - No

27) Please indicate how important you think the following reasons for preserving digital data are:

Very important – Important - Slightly important - Not important

27.1 If research is publicly funded, the results should become public property and therefore properly preserved.

27.2 It will stimulate the advancement of science (new research can build on existing knowledge).

27.3 It may serve validation purposes in the future.

27.4 It allows for re-analysis of existing data.

27.5 It may stimulate inter- disciplinary collaborations.

27.6 It potentially has economic value.

27.7 It is unique.

28) How important/relevant do you regard the following threats over the next 10 years?

Very Important – Important - Slightly Important - Not Important - I don't know

28.1 Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved

28.2 Lack of sustainable hardware, software or support of computer environment may make the information inaccessible

28.3 Evidence may be lost because the origin and authenticity of the data may be uncertain

28.4 Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future

28.5 Loss of ability to identify the location of data

28.6 The current custodian of the data, whether an organization or project, may cease to exist at some point in the future

28.7 The ones we trust to look after the digital holdings may let us down

29) When submitting data to an external facility, are you required to:

Yes – No - Don`t know - Other

comply to a standard data format?

transfer copyrights?

supply additional information such as manuals, software?

30) To which of the following facilities would you be willing to submit digital research data in the near future? (multiple answers possible)

Digital archive of organization

Open digital archive of organization

Digital archive (data center) of your discipline

Publisher (data combined enclosed with scientific paper).

External web service for online backup

I do not want to submit digital research data to external facilities

Other (please specify)

The following criteria must be fulfilled for this question to be shown:

(If To which of the following facilities would you be willing to submit digital research data in the near future? (multiple answers possible) equals I do not want to submit digital research data to external facilities)

31) If you do not want to submit data to an external facility, why not? (multiple answers possible)

I do not know of any digital archives (repositories or data centers) to which I can submit data.

I do not believe my digital research data is secure at a data center, journal site or repository.

I am not aware of any submission procedures.

I do not want to run the risk of anyone else being able to access and use my digital research data.

The submission procedures are too complicated and therefore it takes too much time to submit data.

Submitting digital research data costs money and therefore is not attractive to me.

I am afraid my data will be misused (wrong interpretation, biased statements, etc.)

I am afraid I will lose my copyrights.

Don't know.

Other (please specify)

32) Do you think that a national infrastructure for data preservation and access should be built to help guard against some of these threats?

Yes

No

I don't know

33) Do you think the following initiatives would be useful for raising the level of knowledge about preservation of digital research data?

Very useful – Useful - Slightly useful - Not useful - Don't know

Workshops on preservation of digital material

User-oriented training sessions on digital preservation

International knowledge platform/forum on digital preservation

Development of guidelines/manuals on how to preserve digital data

34) May the data collected in this survey be shared with related projects in the field of digital preservation?

Yes - No