Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Miro Nurmela

Aalto Data Repository

Research data management, sharing and publishing in the world of data intensive science

Master's Thesis
Espoo, December 18th, 2015

Supervisor: Associate Professor Keijo Heljanko
Advisor: Associate Professor Keijo Heljanko

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Aalto University
School of Science

ABSTRACT OF
MASTER'S THESIS

Author:	Miro Nurmela		
Title:	Aalto Data Repository Research data management, sharing and publishing in the world of data intensive science		
Date:	December 18th, 2015	**Pages:**	viii + 77
Major:	Software Systems and Engineering	**Code:**	T-110
Supervisor:	Associate Professor Keijo Heljanko		
Advisor:	Associate Professor Keijo Heljanko		

All fields of science are becoming data intensive. The decrease of computing price and the evolution of data collection methods have created novel research opportunities. This new data intensive paradigm puts a new kind of premium to research data, since it more than ever before forms the lifeblood of research. As a result the demand for publishing research data has increased from both funding bodies and the research community. These factors combined present novel challenges for research data management and publication.

This thesis sheds light on the current status of research data management, sharing and publishing. The primary contribution of the thesis is the examination of existing technical solution to these research data challenges. In addition requirements for successful research data solutions are proposed. The secondary contribution of the thesis is the research of the cultural atmosphere surrounding research data management, sharing and publishing.

Technical solutions for the three research data challenges were found mainly from within the open source community. Solutions like Dataverse, Invenio, Hydra Project and CKAN offer platforms for sharing and publishing data. Solutions like iRODS can be used to manage research data. These solutions serve their purpose, but there is no good integrated solution that would solve all three research data challenges. The lack of holistic solutions combined with the lack of culture and knowledge about research data management result in limited research data publishing and sharing.

Future work should, in addition to building an integrated solution for sharing, publishing and managing research data, aim to make the culture around research data management more open.

Keywords:	research data, repository, open data, open publishing, data policy, data lifecycle, data management
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

Aalto-yliopisto
Perustieteiden
korkeakoulu

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä:	Miro Nurmela		
Työn nimi:	Aallon tutkimustietosäilö Tutkimustiedon hallinta, julkaisu ja jakaminen tutkimustietokeskeisen tieteen ajassa		
Päiväys:	18. joulukuuta 2015	**Sivumäärä:**	viii + 77
Pääaine:	Ohjelmistojärjestelmät ja -tuotanto	**Koodi:**	T-110
Valvoja:	Professori Keijo Heljanko		
Ohjaaja:	Professori Keijo Heljanko		

Tutkimustieto on nykypäivänä keskeinen osa kaikkea tutkimusta. Tieteellisen laskennan hinnan lasku ja tutkimustiedon keräämismenetelmien kehitys ovat johtaneet uusiin tutkimusmenetelmiin. Tutkimustietokeskeinen suuntaus asettaa tutkimustiedon tärkeämpään asemaan kuin koskaan aiemmin, sillä ilman laadukasta tutkimustietoa parhaat tulokset jäävät saavuttamatta. Tämän johdosta tutkijayhteisö ja tutkimuksen rahoittajatahot ovat alkaneet vaatia tutkimustiedon julkaisemista. Näiden kehitysten johdosta tutkimsutiedon hallintaan ja julkaisuun täytyy kehittää uusia ratkaisuja.

Tämä opinnäytetyö valottaa tutkimustiedon julkaisemisen, jakamisen ja hallinnan nykytilannetta. Opinnäytteen pääpanos on näitä tutkimustiedon haasteita ratkovien teknisten toteutusten tutkimus, minkä lisäksi työssä ehdotetaan onnistuneen tutkimustietoratkaisun vaatimuksia. Opinnäytetyö tutkii myös tutkimustiedon julkaisemisen ja hallinnan kulttuuria ja käytäntöjä.

Opinnäytetyössä esitellään pääasiassa avoimen lähdekoodin ratkaisuja tutkimustiedon haasteisiin. Dataversen, Invenion, Hydra-projektin ja CKANin kaltaiset järjestelmät ovat alustoja tutkimustiedon julkaisemiseen. iRODSin kaltaiset sovellukset soveltuvat tutkimustiedon hallintaan. Nämä ratkaisut toimivat, mutta yhdistettyä ratkaisua tutkimustiedon hallintaan, julkaisuun ja jakamiseen ei ole. Kokonaisratkaisujen sekä tutkimustiedon hallintaan ja jakamiseen liittyvän kulttuurin puutteesta seuraa, että tutkimustietoa jaetaan vähän verrattuna sen määrään.

Jatkotutkimuksen tulisi keskittyä tutkimustiedon hallinnan, jakamisen ja julkaisemisen yhdistävän palvelun lisäksi etsimään ratkaisuja aiheeseen liittyvän kulttuurin ja tietotaidon parantamiseen.

Asiasanat: tutkimustieto, säilö, avoin data, avoin julkaiseminen, datapolitiikka, datan elämänkaari, datanhallinta

Kieli: Englanti

Acknowledgements

Thanks to my supervisor, Keijo Heljanko, for supporting me through the thesis.

This thesis would not have been possible without all the people that helped me by contributing their time and insights. The people I interviewd are credited in the footnotes of the thesis when their insights are being used. A huge thanks for them, especially Richard Darst, who helped with the formulation of the user stories.

Thanks also to my teacher colleagues of the ME310 course for letting me take time off from teaching to write this thesis.

Thanks to Jasmin for being there.

Espoo, December 18th, 2015

Miro Nurmela

Terminology

Research data	In the context of this thesis, research data refers to data that has been generated or used in scientific work
Research papers	Research papers is the umbrella term used in this thesis to cover traditional scientific publishing material, such as journal articles and conference papers
Research data management	Research data management refers to to the act of managing research data during a research project - this includes but is not limited to documenting, annotating and cleaning research data
Research data sharing	Research data sharing refers to sharing research data between two parties, either by sharing it privately or making the research data available
Research data publishing	Research data publishing refers to the act of making research data public for all the world to see
Metadata	Metadata is descriptive data about data that is used to give context and other additional information about the data itself
CSC	CSC is a Finnish provider of scientific computing and storage services
EUDAT	EUDAT is a EU level initiative that aims to bring research data management, sharing and publishing tools to European research institutions
DOI	DOI (Digital object identifier) is a commonly used persistent identifier scheme for research papers
Handle	Handle is a persistent identifier scheme
URN	URN is a Finnish persistent identifier scheme
Dataverse	Dataverse is an open source research data publishing platform originally developed in Harvard University
iRODS	iRODS is an open source research data management software
Etsin	Etsin is a metadata publishing tool for Finnish institutions
Avaa	Avaa is a Finnish service to open datasets for public use
ATT	ATT (Avoin Tiede ja Tutkimus) is a Finnish ministry led initiative to introduce openness to the Finnish field of science
PAS	PAS (Pitkäaikaissäilytys) is a Finnish project to develop long term archival of datasets
Hydra Project	Hydra Project (also referred to as Hydra in this thesis) is an extensible open source repository solution

Invenio	Invenio is a CERN based, now open source data repository solution
Zenodo	Zenodo is a public instance of Invenio with a custom user interface
GitHub	GitHub is a platform for collaborating on and sharing source code
B2Share	B2Share is a research data publishing service of the EUDAT initiative
B2Drop	B2Drop is a research data sharing and managing tool of the EUDAT initiative
Apache Solr	Apache Solr (also referred to as Solr in this thesis) is a tool that in the context of data repositories is often used to index databases to make them searchable
ACRIS	ACRIS (Aalto Current Research Information System) is a system to manage research information being implemented in Aalto University. It incorporates Elsevier Pure, a tool that has research data management and publishing features

Contents

Chapter 1

Introduction

The world of science is moving towards more and more data intensive research. Methods for gathering research data and analyzing the data are growing cheaper and cheaper while the knowledge of algorithms and statistical analysis keeps improving. This has made fields that were previously very data intensive, such as particle physics, accumulate even more data. On the other hand fields, such as social sciences, where the amounts of data have traditionally been small to quickly become much larger. This new world sets research data to a new kind of premium - it is the heart of research more than ever before in the past. [32]

With the increased value of research data managing that data becomes important. At the same time the world is becoming more and more connected, which enables research data to be transferred easily all over the world. A big part of research nowadays is done in groups that are scattered all around the world, which means that sharing research data with your partners becomes a real challenge. Research data might be too big to be sent via email or other traditional tools or it might contain data that can not leave a secure datacenter. [7, 29]

The advance of technology has also made it possible to share data with anyone, and since data generated by researchers is often done with public funding there is a logical argument to be made that publicly funded research data should be available for the public good as well. [7] This imposes challenges for the researchers, since in order to make research data public and useful for others appropriate metadata needs to be added to the data. This process is both time consuming and not necessarily required to do research, making it a low priority for researchers [57].

This thesis tackles the problems of the new, data intensive world of research with a focus on the technical implementations to research data publishing and sharing. The research questions are defined in Section 2. Section 2 also outlines the approach of this thesis and its main contributions.

In Section 3 scientific literature about research data management and sharing, open access publishing and other relevant fields are presented. The literature shows that open research data and openly accessible research papers further science, but there are many challenges that go into the process of opening them up. The problems of sharing research data are not only technical, but organizational and cultural matters have a big impact. The literature review also covers related areas, such as linked data, research data policy, research data curation and research workflows.

Section 4 positions the thesis in the Finnish field and presents a snapshot of the current level of research data management and sharing in Finland. The chapter contains interviews from the multiple stakeholders that deal with research data, questionnaires that were used to gauge the needs of Aalto University employees regarding research data and benchmarking of existing technical solutions. The findings echo the findings from the literature review in that there is little know-how or culture towards research data publishing and research data management is handled with very different approaches across individuals. Functional technical solutions exist and they seem to fill their roles, but a holistic solution that would take care of both managing research data during the research process and publishing it in the end is both missing and needed. The results from the multiple stakeholder groups interviewed also strengthen the view that simply technical solutions do not solve the research data management and publication problem - there is a need to teach research research data management and publishing to change the perceptions and culture surrounding them.

The positioning research leads to the experimental part of the thesis. Section 5 describes how we chose Harvard Dataverse from the group of existing technical solutions and how we use it to gain further insights about the technical implementations of research data publishing platforms. The user tests show that the mechanical process of using the research data publication platform is easy to learn and use, but there are some technical and usability improvements that could be made. Interactions with the users also brought up the point again that while the solution does seem to work, it would be hard for them to use since their research data is not primed for publication. In this context the solution working means that it can be used to upload, search and download research data.

The research done for this thesis is discussed in Section 6. In addition to evaluating the research methods of the thesis synthesis of the insights is presented. Survival strategies for institutions regarding research data management, publishing and sharing are proposed. This study has also yielded many requirements and points that need to be taken into consideration when designing or evaluating a research data management and publishing system. These requirements are presented for future use. Future work should include investigation into how a holistic solution for managing and publishing research data could be implemented as well as investigation on how the culture of research data management could be opened up.

The conclusions of the thesis are presented in Section 7. The section shows all the studied technical implementations in comparison with regards to research data publishing and management. With the examination presented in this thesis it is concluded that while the techincal solutions do sufficiently well in what they are designed to do, there is a need for a solutions that would combine both research data management and publishing. It is also concluded that in order to make these technical solutions work and spread, the culture of research data management and sharing has to be made more open.

Chapter 2

Problem Statement

2.1 Research questions

The primary research questions of this thesis are how research data can be shared and published using modern tools and how these tools work in practice. The rationale behind these questions is the increased value of research data. Advances in sharing technologies and the requirements for research data publication from the funding bodies and the research community make research data sharing and publishing is a current problem.

During the research it became clear that there are other non technical factors that affect the sharing and publishing of research data within the scientific community. In light of this the secondary research question of the thesis is what non technical matters affect the sharing and publishing of research data.

The non technical matters also concern research data management during the research process, which means that tools for research data sharing and publishing can not be examined without some examination of tools and practices for research data management. The thesis does not focus on research data management tools, but covers some of them and looks into the research data publication and sharing tools also from the angle of research data management during the research process.

2.2 The contributions of this thesis

The main contribution of this thesis is the technical examination of existing software solutions to research data sharing and publishing. The most widespread tools are benchmarked and from among them one solution, Harvard Dataverse, is examined in further detail for deeper understanding of the functioning of these systems. The existing research data publication platforms are fairly similar, which makes examination like this applicable to the other solutions as well.

In addition to the technical examination this thesis proposes a set of requirements that could be used as a basis of designing a research data management and publication system or validate an existing system. The requirements are derived from the background research and the user tests conducted.

This thesis also provides an overview of the cultural atmosphere that surrounds research data management and publishing. In light of the learnigs presented in the thesis survival strategies regarding research data management, publishing and sharing for research institutions in the new world of data intensive science are proposed and ranked.

2.3 The research approach

To address the research problems the following methods are used:

- Literature review,
- Expert interviews,
- Questionnaires,
- Tehcnical benchmarks and tests; and
- User tests.

Literature review is used to gain background knowledge of research data management and publishing. Expert interviews are used to position the thesis in the Finnish field - there is no point in working on something similar that is already being solved. The expert interviews also aim to see if the findings from the literature apply in practice.

Aalto University has is currently forming a policy about research data management and as a part of that project two questionnaires about the current needs of research data management have been conducted. The results of those questionnaires are presented as additional evidence for the needs and current status of research data management.

Existing technical solutions are benchmarked and examined in order to understand the current options available. In order to facilitate user tests one solution is going to be selected to act as a tool to conduct the user tests.

Since the questionnaires already exist to shed light on the current needs user tests were chosen as the method to extract most value towards designing a system for research data management and publishing. The reasoning is that since a system that relies this heavily on the users, in this case mostly the research scientists, the users should be heard first and foremost. Contextual interviews and tests with lead users were conducted to gain deeper understanding about the current tools for research data sharing and the current status of research data management. The goal of this user centric approach was to formulate a system that would not be just another information system that nobody uses but a system that would provide value for both the users that put their research data to the system and the users that would get research data out of the system.

Chapter 3

Background in scientific literature

3.1 Research on open access publishing

Open access refers to the online literature, research data and research papers in the context of this thesis, that are free of charge for the reader of the publication and openly available for reuse and redistribution [37, 60]. The research on open access publishing shows that publishing research papers openly correlates to increased citations to the research papers (ranging between 36% - 172% increase [18, 26, 27]). Open access publishing of research papers make the paper known and cited faster than non open access [27]. The increase in citations to research papers is magnified in papers with lower ranked journals, implying that open access publishing is a cost effective way to increase the chance of making research impact [69].

There are different ways to do open access to research papers. One division is the Golden Road and the Green Road, where the former refers to the journals themselves publishing with open access and the latter to so called self-archiving, where researchers publish their papers themselves [28]. Self-archiving can be done on the research paper's author personal website, a disciplinary archive or a institutional repository (either an institution wide or smaller unit, a department for example) [37]. Aalto University has an institutional repository for research papers and theses produced in Aalto University[1].

Open access publishing is growing in popularity. A study that explored the growth of open access publishing between years 1993-2009 reported an annual growth rate of 18% in open access journals and 30% increase in open access articles, shown in Figure 3.1 [42]. The study does not speculate on the reasons behind the growth of open access publishing, but mentions the fact that nowadays many researchers use free search engines such as Google Scholar[2] to look for sources and that makes research papers published in open access way to be more available than research papers that are not openly accessible. Open access has also introduced the concept of an open access mandate, which is a mandate or a policy adopted by a university, research institution or research funder. The open access mandate means that the researchers working under the mandate are required to publish their research papers[3].

The total number of open access materials is hard to estimate nowadays, but the Directory of Open Access Journals[4] measures the amount of open

[1] https://aaltodoc.aalto.fi/

[2] https://scholar.google.fi/

[3] http://roarmap.eprints.org/

[4] https://doaj.org/

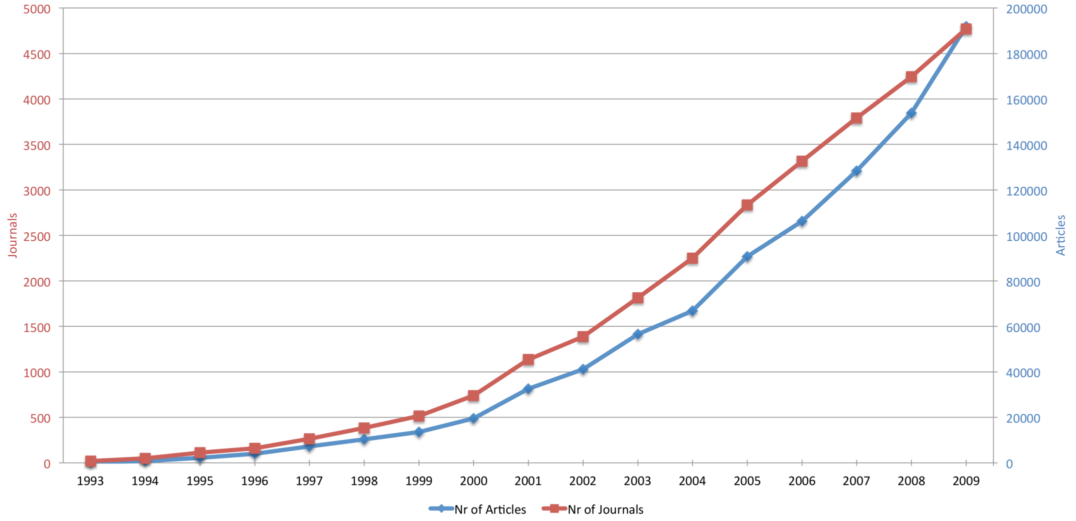


Figure 3.1: The growth of open access publishing between from 1993 to 2009 [42]

access journals world wide at 10 703 at the time of writing this thesis. The amount of open data journals and articles is growing faster than the amount of more traditional, non open data journals [42]. This is likely to correlate to the increased amount of open access mandates, since mandatory open access rules imposed to the researchers correlate to a four time increase in deposits to open access repositories [61].

It makes sense that open access publishing has become more popular. An usual metric to measure the impact and relevance of a researchers is the h-index [33], which takes into account the number of papers and the number of citations to those papers.

For further reference, we can recommend the work of Stevan Harnad as a good starting point to open access literature. He is referenced in [26–28]. If you are interested in data related to open access, it is available both on the Directory of Open Access Journals as well as in a public data repository about the growth of open access[5].

3.2 Research data open access publishing

The idea to make research data openly accessible has been around for a long time. As early as 1985 policies and practices to share research data to further research and prevent fraud were developed [19]. The Internet and the emergence of data intensive science have changed the the possibilities and quantities of research data. Research Data Alliance (RDA) has been founded in 2013 to address the growing need of research data publishing and sharing infrastruc-

[5] http://dataverse.scholarsportal.info/dvn/dv/dgoa

ture [4, 32]. Despite this, the practices of research data open access publishing are far from those of research paper open access publishing. Metrics, such as the h-index [33], do not exist for research data nor is research data publishing accounted for in the h-index.

Research papers should be self-contained in the sense that a if you know the area of research, you can read the research paper and understand it's main points and findings. Research papers often contain the processed results of the research data behind them as well. Research data, however, is rarely self contained and requires metadata to be useful for re use. It is also worth noting that datasets nowadays will not be used solely by people who are experts in the fields that the research data originated from nor will people work in geographically in the same locations anymore which makes metadata even more crucial for reusing the data [7, 29].

While research papers generally follow an established structure and can easily be published online in a PDF format, research data comes in many different forms and flavors. For example, phylogenetic trees used in evolution research look nothing like the brain images gathered in neuroimaging research. This imposes a technical challenge to the solutions that could be used to share research data - file formats are different, the file size may be range anywhere from few hundred kilobytes to multiple terabytes and because many fields do not share common practices on how to manage their research data even seemingly similar datasets might be incompatible between researchers [53, 64].

Peer reviewing, which is at the heart of making sure that research papers and thus research is valid, is yet to be fully introduced to the process of publishing research data. One study piloted a process of reviewing datasets the researchers had used [25]. The study found that many researchers indeed found datasets useful for their research and that the amount of people that responded to their online questionnaire was 573, 15.8% of the people who the survey was sent to. They are continuing to develop the review system, focusing on especially how and when to ask reviews from the dataset consumers in order to minimize the bother to the user.

As such, research data open access and publication is not prevalent in many fields of science. The practices of data sharing vary a lot between fields and even inside disciplines [7, 11]. This takes place even while all fields of science are becoming more and more data intensive and many fields, such as psychology, could reap great benefits from sharing research data [32, 65]. The following Sections 3.3 and 3.4 delve deeper into the benefits and challenges of sharing research data by publishing it with the open access paradigm. A paper by P Arzberger et al. [2] summarizes these challenges and benefits in an organizational level. A paper by Jean-Baptiste Poline et al. describes the situation of data sharing in the field of neuroimaging research in good depth and the discussion in the paper is also applicable to other fields of science [53].

The latest resarch shows that the acceptance towards sharing research data is growing. Challenges, like the quality of tools for research data sharing and management and the lack of training for research data management still exist but it seems that the world of reserarch is moving towards sharing research data. [63]

3.3 Challenges of sharing research data

Sharing research data poses many challenges, some of them organizational and managerial while others have to do with the nature of research data and the culture surrounding research data management, publication and sharing. The challenges are presented in more detail below [2, 62]:

- Technological issues - there is a lack of infrastructure to effectively publish and share research data.
- Institutional and managerial issues - the principles of open access require tailoring for institutions, since datasets, research funding and similar matters require local management.
- Financial issues - managing research data archiving and publishing infrastructure requires continuous financial investment beyond the scope of implementation and publication of the research projects results.
- Legal and policy issues - national and international law set limitations to sharing research data.
- Cultural and behavioural issues - in the end, research data sharing comes down to the actions of the researchers generating the research data and if the culture for it does not exist and the behaviour is not encouraged, there will be no sharing of research data.

These issues are highlighted in the questionnaire responses in [62] for reasons for not making their data not accessible to public, shown in Table 3.1.

The Tenopir paper [62] contains many more useful tables, showing for example, that many researchers (56.1%) either do not know or do not have metadata standards for their research data and that research data comes in many different categories without even going to the specifics in the fields of science.

Research data may be serious privacy concern especially in fields such as genomics or health care related research where research data could be connected to the participants of the study. There is a tension between sharing relevant and good quality data and protecting the privacy of the participants - more detailed data provides for a more rich research, but allows for an easier connection on to the participants [38]. Work is being done to facilitate safe and accurate sharing of data with privacy concerns. Both codes of conduct and technological solutions are required, especially since the data will outlive the participants of the study and the privacy must be preserved for the entirety of the data's lifetime [40, 47].

Table 3.1: Reasons for not sharing research data, questionnaire with 1329 respondents [62]

Reason	**Responses**	**Percent**
Insufficient Time	603	53.6%
Lack of Funding	445	39.6%
Do not Have Rights to Make Data Public	271	24.1%
No Place to Put Data	264	23.5%
Lack of Standards	222	19.8%
Sponsor does not Require	196	17.4%
Do not Need Data	169	15.0%
Other Reasons For Data Not Available	164	14.6%
Should not be Available	162	14.4%

A more implementation level problem to sharing research data is the fact that due to the legal issues or maybe the desire to work on your data before publishing it means that research data should be able the be shared more locally in addition to being open to the whole world. In practice, this means that there should be means to provide access to the research data to collaborating researchers or people within the research organization you are working in to enable collaboration while preventing access from the rest of the world [68].

A Chinese group of researchers conducted an experiment where they measured the efficiency of Chinese national data sharing platforms. Their findings are listed below [58]:

- The policies governing data sharing need to be improved.
- Metadata is poorly used.
- The whole datasets are not available and datasets are not in fact openly available but require permissions.
- The platforms do not offer all the necessary services related to research data sharing, such as data collection.
- The research efficiency of the platforms is low.

3.4 Benefits of publishing research data

When talking about the benefits of publishing and sharing research data two big points are generally made. Firstly it is thought that publishing and sharing research data pushes science forward by enabling more people to work on the data and subsequently accelerate the process of making relevant discoveries. It is also not to be forgotten that reproducibility is one of the founding pillars of scientific process and reproducing others' work without access to their data

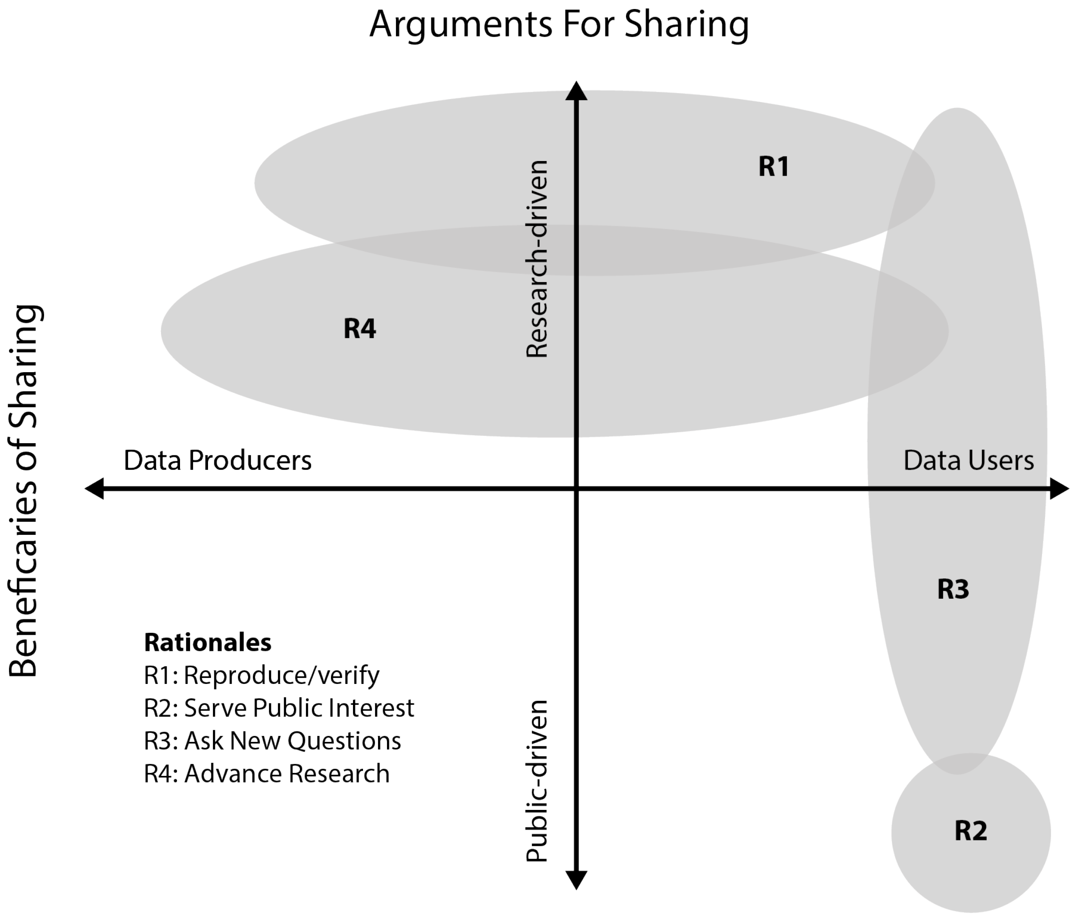


Figure 3.2: The different beneficiaries and rationales of data sharing - figure drawn after [7]

is very hard [36, 49]. Secondly it is thought that the publishing of research data makes your research papers better, giving them credibility and yielding more citations in the process. There are more subtle benefits as well that are discussed later in this section.

Publishing research data benefits different stakeholders involved in the data publishing game, as shown in Figure 3.2. The figure also lays out how the rationales of sharing research data affect the different stakeholder groups [7].

Publishing research data furthers science by allowing new people to ask new questions of the existing data [64], allowing for better reproducibility [36] and straight up widening the scope and depth of the research in question [20]. By making data publicly available the developers and people not included in the scientific community, such as political decision makers, can also benefit from the data being generated in the different fields of research [7].

One can take a different view on the matter of sharing research data as well - if one's results are strong and backed up by evidence, there should not be a barrier (other than the work related to the publishing and he possible legal

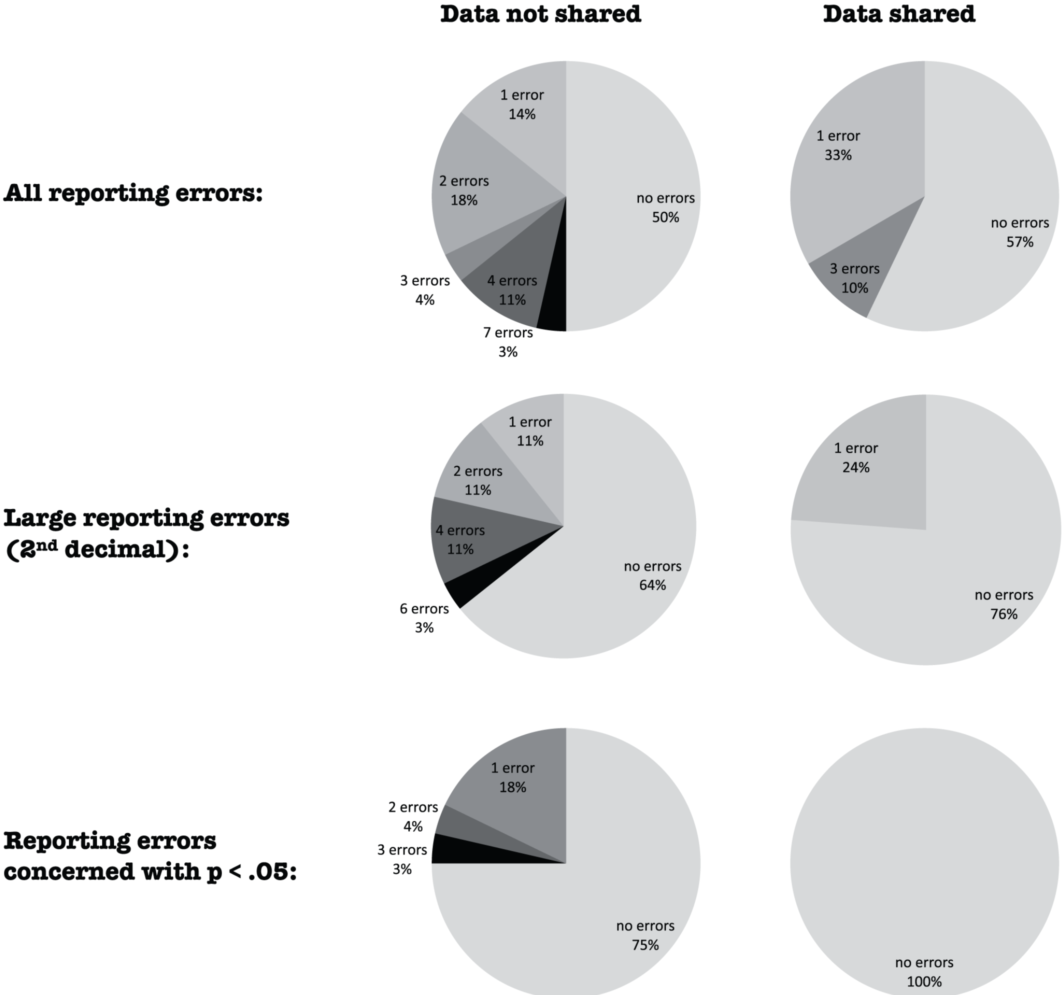


Figure 3.3: The difference in reporting errors between research papers with not public and public research data from the field of psychology [66]

issues) to publish one's data. A reason to not make data openly accessible is the fear of reanalysis and that people might find errors in your research. This was examined in a study in the field of psychology [66]. The findings of the study were that it was indeed so that research papers not publishing their data had weaker results. The comparison of amounts of errors is shown in Figure 3.3. In order to avoid this phenomenon better policies of sharing data are needed which would in turn make the quality of research better.

In order to get research data published and openly accessible the researchers generating the data need to put it to the public domain. The benefits of sharing, however, are more clear to the consumers of the published data and preparing data for publication requires work from the person publishing the data. Traditionally academic credit and credibility has been tied to the number of articles published and citations of those articles and public data does not play a role in that. However, publishing datasets can have a positive impact on research papers' citation rate bringing the associate credit with it indirectly [51].

In a study about cancer microarray clinical publications it was found out that research papers that had their research data published received a 69% increase in citation rate. In the same study the 48% of the research papers that had the associated research data included received 85% of the aggregate citations [51].

In addition to providing value to the researchers publishing their papers in the form of citations, one study [52] study suggests that the investment to a research data repository gives a generous return of investment to the institution building and using the repository. The study makes a comparison of spent money - 400 000 dollars in original research resulted in 16 papers, whereas the cost of running a data repository of biological research data sets for a year would cost the same 400 000 dollars but contribute more than 1000 papers within the following four years.

Making research data openly accessible lessens the risk of data fraud taking place. Data fraud, which constitutes acts such as using fictitious data or tampering with the data in order to support your conclusions, could be prevented or made more difficult by sharing all research data. This matter is discussed in a paper by Peter Doorn et al. [15] following a series of incidents where fraudulent research data was used as a basis for research.

3.5 Research on the validity of increased citation rate with publishing research data

While many research papers imply a causation between publishing research papers in an open access way and an increased citation rate, there are studies that take this claim critically and examine whether there are other causes that would explain the perceived citation advantage. In one study [12] the research methods of studies that show the increased citation rates with open access publishing do not take all the matters that go into citation analysis into account and proposes that more sophisticated and rigorous statistical analysis to prove the causation would be required.

Another study notes that open access research papers get more downloads and readers, but no more citations. The citation advantage might be an artifact of other causes or caused by uncontrolled experiments [13]. The paper proposes that the artifact could be the self selection bias - authors choose their best work to be open access, thus contributing the increased citation rate to the quality of work instead of the open access.

While these studies are fewer than the ones pointing to the increased citation rate in open access publishing, it is important to note that citation analysis is not a simple manner and the methods to show the benefits of open access publishing should be carefully designed. There are studies refuting the self selection bias, so clearly there is still work to be done in the area [22].

3.6 Different ways of sharing research data

Openly publishing and sharing research data takes place on different levels of organizations. The general ways to publish research data openly are the following:

- Institutional data repositories, discussed in [11].
- Disciplinary repositories, example in [17].
- International repositories, example in [45].
- National repositories, example in [9].

A more practical look into the different implementations to these as well as some practical benchmarking will be presented in Section 4.8.

3.7 Adherence to data publishing requirements

Many funding bodies and journals require the publication of research data related to the research papers. In Finland the Academy of Finland has also made its policy such that it demands research data to be made as public as possible.[6] Adherence to this requirement has been studied and the results are that despite the demand for sharing research data it is not being shared commonly. In one study 10 datasets were requested from a journal that explicitly requests sharing datasets and only one dataset was received [57]. In a bigger study the authors were able to gain 63 datasets out of 249 possible datasets from a journal that also requires datasets to be shared for reanalysis [65]. In a study of 500 hundred papers, of which 149 were not subject to any research data publishing requirements, only 47 papers had submitted the complete raw data online. Of the 149 who were not required to do publish research data did not publish anything [1]. The 500 hundred papers in the last study were selected from 50 different journals by selecting the 10 research papers with the highest impact factor.

The reasons behind not sharing research data were different. In the smallest study [57], where the researchers were able to contact the people who refused to share, the reasons for not sharing are the following:

- Two email addresses listed on the original research paper were no longer valid and once one of them was reached she was on maternity leave and could not help.

[6] `http://www.aka.fi/avoin_julkaiseminen/`

- Two of those who did not share did not give a reason for not sharing - on further inquiry one of them responded that he was not aware of the research data publishing requirements.
- One stated that he was too busy and compiling the research data would be too much work.
- One had changed institutions and no longer had jurisdiction over the data and the people in the publishing institution responded that sharing would have been too much work.
- Three did not answer the inquiry at all at first - on further inquiry one of them replied that he was in favor of sharing data in general but wanted to conduct more research on the research data himself first.

In the biggest study the authors postulated that factors such as the 6-12 months between the publishing of the papers and the authors' investigation the research data publication policies might have changed, though this is noted to be unlikely [1].

3.8 Sharing research workflows

Since reproduction of results is a big draw towards sharing research data sharing research workflows has been a subject of research. As an example, you could have a look at my$_{\text{Experiment}}$[7] or Galaxy[8]. The former is a holistic approach to allow researchers from all fields to share workflows, whereas the latter focuses on the field of genomics. This approach promotes the transparency of science and enables easier reproduction of data. From the scientific literature side you can take a look at [23, 55].

Along with sharing research workflows the idea of executable papers has been thrown around.[9] The idea of executable papers is to provide a platform where along with the research paper some way to execute relevant objects to the paper are included. Relevant objects might include plotting graphs or trying theorems proposed in the research paper with different outputs. Some of the solutions to also allow the viewing of primary research data [24, 48]. The main goal of executable papers is to allow the results of the research paper to replicated easily.

3.9 Sharing Big Data

The focus of this thesis and the discussion in this literature review has been in the traditional research data, but Big Data is becoming more prominent in

[7] http://www.myexperiment.org/home

[8] https://galaxyproject.org/

[9] http://www.executablepapers.com/

all fields of science. Managing Big Data adds more complexity to the data management processes and infrastructure [14]. Sharing Big Data is also a wholly different concept than sharing more traditional data, since you cannot download Big Data datasets on your personal computer. The problem becomes more about access controls and providing a secure infrastructure to access the Big Data when collaboration and publication of Big Data is concerned.

An interesting twist to the sharing research workflows and Big Data is the Hadoop plug-in for Galaxy, which allows Big Data computation to be integrated to the genomics research workflows [50].

3.10 Research data curation

Publishing and sharing research data requires work that is all away from the researchers other research work. Moreover, after the research data has been published it no longer concerns the researcher, but someone has to look after it. University libraries have already assumed the role of curating digitally published research papers and it would make sense for them to be included in the process of research data publication and archival as the curation and archival experts [31, 34]. The sharp increase in the amounts of data as well as the growing digitalization of research output is going to force libraries to find their role as well as researchers to find a way to collaborate with the data management experts to make the research data publication and archival as successful as possible.

3.11 Implementations in the literature

Some of the technical solutions for research data management (including management during the research projects and publishing data) have been documented in scientific literature. They are presented here and a more practical look into some of them is presented in section 4.8.

iRODS (Integrated Rule-Oriented Data System) is open source software designed for research organizations to manage their research data. iRODS virtualizes the storage hardware and offers a programmable rules system to enable automatic data management. For more information, see their website[10] and the book about it from [54].

CKAN, an open source data publishing platform, is a tool that can be used to publish digital research data. It can be found online[11] and a paper describing it in use in a academic context can be found in [67]. The article found the CKAN system to be robust open source software suitable for publishing research data but not managing the research data lifecycle.

[10] `http://irods.org/`

[11] `http://ckan.org/`

Dataverse is an open source data repository software solution which you can find online[12]. Similarly to CKAN, Dataverse focuses on publishing research data but has more emphasis on getting citations on the data and making it that way more comparable to publishing research papers. You can read the background of Dataverse from [39].

The Hydra Project[13] is another open source repository solution. What differentiates Hydra from the previous implementations is that it is quite flexible - the user can define all the data models and the content types that go into the repository. A research paper about the system is at [3].

Invenio is a CERN born open source reseach data repository. It began as a CERN documentation server, aiming to be a publishing platform for the so called ”grey literature” - research output produced outside commercial and academic distribution channels - which has been a CERN way of publishing. It has since grown and been adapted worldwide as a backbone of many repository solutions. [8]

The implementations mentioned here have been open source solutions. Elsevier, an academic publishing company, has also integrated research data management tools to their publication management solutions. Currently their solution is known as Pure and it integrates to other Elsevier solutions. [10]

3.12 Linked data

Open access research data discussion would not be complete without a mention of linked data. Linked data refers to data published in the Internet that is connected with typed links, is machine readable and linked to and from other data [6]. Linked data does not necessarily refer to research data, but the analogue of online accessible data is clear. The ideas that go into making linked data possible are similar to the ones that research data requires - standards for expressing data in a unified manner and enabling all kinds of data to be published and linked are at the core of linked data. Linked data is tightly connected with the concept of semantic web [5].

Linked data is studied in Aalto University, and one of the studies describes a system to publish linked data [21]. While not directly about research data, bringing research data to be a part of the linked data available online would be an interesting addition to the plain publishing of research data. The Finnish linked data initiative as well as the international organization have online resources for those interested[14].

[12] http://dataverse.org/

[13] http://projecthydra.org/

[14] http://www.ldf.fi/index.html, http://linkeddata.org/

Chapter 4

Positioning the Thesis

Master's theses do not live in a vacuum. To position the thesis and provide the best possible outcome for Aalto University we have made an effort to find out what is the current state of research data management, publishing and sharing as well as what are the current challenges and projects in the Finnish landscape. While the literature review in Section 3 covered the overall view of the current status of publishing and sharing research data and research papers, the goal of this thesis is to provide practical value towards implementing research data management, publication and sharing systems. Practical insights lie within people working in the area. The tools used to position the thesis were interviews, benchmarking existing solutions and a questionnaires.

4.1 Researcher interviews

Scientists and researchers are the core user group of any publishing or sharing system since they are the ones generating the data to the system and possibly populating the system. The research also shows that a successful repository system requires user engagement [46]. Scientist and researcher interviews were conducted within different research groups and researcher in Aalto Otaniemi campus. The goal of these interviews was to learn how data management is taught and used in Aalto University and what are the scientist and researchers perceptions on sharing and publishing research data. Previous experiences with sharing and using others' data were also gathered.

The interviews presented here were conducted in the beginning phase of this thesis - later, more in depth interviews were conducted to test a possible solution. This discussion can be found in Section 5.

In an interview with a research group[1] the findings fell in line with the findings from literature. As of writing of this thesis, data management is not systematically taught for the researchers and there is no culture or experience in data sharing. Upon questioning it became clear that the data in the possession of the researchers would have required a lot of cleaning and metadata additions prior to publishing - this is no surprise considering the fact that the datasets they were using were not designed to go public in the first place. Though lack of metadata practices and lack of publishing infrastructure were also brought up.

[1] M. Nurmela and the Complex Networks group at Aalto University (`http://becs.aalto.fi/en/research/complex_networks`), personal communication, July 31st, 2015

Some members of the research group had shared some of their datasets through public cloud services such as Google Drive[2] with collaborators and used others' datasets. This also raised the point that in order to use others' datasets they had been asked to cite the papers that used that dataset, underlining the the undeveloped culture of research data sharing. Some members of the group saw big advantages in making research output public, especially from the angle of reproducibility, but also raised concerns about the privacy issues that for example telecommunication data would cause. The members of the group were also concerned about about the size of their data, since using the existing solutions to share hundreds of gigabytes worth of data would prove cumbersome.

One member of the group had taken a role of a mentor with the data management issues, teaching the others to use databases and version control software to handle their data and code. The other members of the group, interviewed separately, noted that an effort had been made towards better policies in data handling. The group pointed out that personalized assistance and word of mouth were an efficient way to learn ”boring” things like data management. Own previous experience and learning by doing seemed to be the main way people had learned to, for example, comment their code or arrange their research data into databases so that accessing them would require less time and managing code would be easier.

Discussions with the complex networks groups also brought up the point that even though some data could be published for all the world to see, some data should be only accessible to people within Aalto University and some should even have a more fine grained access control set to them.

In a separate interview with a brain image researcher[3] similar concerns arose. Brain imaging data is large and that makes sharing it hard. Brain images are also considered personal data and publishing them is problematic. The researcher expressed interest in sharing and using others' datasets, but lacked the tools to do so.

4.2 Science IT staff

If a system to publish and share research data would come to be it would have to be maintained and run by people other than the researchers, since the job of the researchers is to do research and not to maintain software systems. With this in mind the people managing the scientific IT systems are a key component to building a successful research data management, publication and sharing platform.

In interviews with a scientific IT systems specialist[4] the findings again aligned with the findings from literature. There is a lack of metadata standards

[2] `http://drive.google.com/`

[3] M. Nurmela and E. Glerean, personal communication, August 13th, 2015

[4] M. Nurmela and M. Hakala, personal communication, July 1st and 7th, 2015

that would make data storage and management uniform across institutions and disciplines. Finland has projects going on related to open science[5] and CSC[6] is building scientific computing environments for Finnish institutions (research, library, archival, education and such). According to the specialist collaborating with all these ongoing projects would benefit both parties and also allow Alto University to develop systems that are not just point solutions. This would also make sense from a financial point and practical point of view, since research is nowadays done in collaboration with other institutions and working together would enable that.

From the point of view of IT system specialist the ideal system for research data includes the whole lifecycle of the data. This entails education on how to manage the data from the creation to the publishing and infrastructure that can be tailored to fit the different user needs. Research data is comes in many forms so a solution that is not flexible would be unsuitable.

University of Jyväskylä has implemented a Dataverse data repository system[7] as well as an iRODS[8] system to manage and store research data. In an interview[9] they noted that nowadays universities need a platform to publish research data. Jyväskylä is among the first in Finland to implement a data policy in practice, offering an infrastructure to manage data during the research projects and publish the results in the end (even though Dataverse and iRODS are not currently easily compatible with each other, though some work has been done for that[10]).

The Jyväskylä experts told that while the Dataverse system was easy to install and modify to accept Jyväskylä University credentials, it still had a long way to go before it was universally accepted within the university. At the time of writing this, the Jyväskylä Dataverse has been running for approximately a year and it contains a very small amount of datasets. Some research groups, however, are using it manage their internal datasets. As to how to get the more datasets into the system, they planned to continue educating about it and making it that way a part of researchers' routine.

As of writing of this thesis Jyväskylä is more involved in development of the iRODS system, having developed a system called Kanki to facilitate collaboration between researchers [41]. The Kanki system is a desktop interface to the iRODS system that allows users to easily access and modify data stored in the iRODS data grid. The Kanki system is not about publishing research data but data management during a research project. iRODS also has federation capabilities, meaning that two iRODS instances could be integrated such that you could access the data from the other system. When writing this Jyväskylä

[5] http://avointiede.fi/

[6] https://www.csc.fi/

[7] https://dvn.jyu.fi/dvn/

[8] http://irods.org/

[9] M. Nurmela, I. Korhonen and A. Auer, personal communication, August 19th, 2015

[10] https://irods.org/wp-content/uploads/2014/06/Odum-DFC-iRODS-Boston.pdf

and CSC were planning to start testing the federation capabilities. Following up on that on a later date would be interesting.

4.3 Project managers on research data related systems

Building and running software systems requires commitment not only from the primary users discussed in Sections 4.1 and 4.2 but the management that supports them. The priorities of the managerial types might not lie in the usability or maintainability of the system, but rather in managing costs and minimizing risks.

Interviewing a manager from Aalto side it was made clear that there is a desire to minimize the systems we have to build and maintain ourselves - since CSC exists to provide scientific computing and storage resources, why should we not use them? Non established or new fields of science could have value from a local repository, but those that have international or discipline specific repositories could use those as well.[11] Managing research data is seen as a problem and a consensus best solution has not emerged.

The Finnish National library is in charge of long term preservation of relevant objects in the Finnish research. They are building a long term storage solution[12], data management plan tool[13] and managing the Finnish unique identifier service[14]. They also manage a Finnish cultural repository[15].

The most important thing to for the project manager at the National Library was that metadata associated with the data has to be good - otherwise archival, management and reuse is impossible. Making metadata work within an institution requires commitment from all levels of management and tools to facilitate that. It is also important to note that even if two systems within different organizations would be technologically perfectly compatible, the bottlenecks might stem from different policies in different organizations and the bureaucracy that comes with it. This thought lessens the burden for all technology to be perfectly compatible.[16]

4.4 Librarians

Publishing research data requires expertise in digital publishing and metadata creation experience. University libraries are experienced in publishing digital

[11] M. Nurmela, A. Sunikka, personal communication, July 17th, 2015

[12] http://avointiede.fi/tutkimus-pas

[13] http://portti.avointiede.fi/tutkimusdata/tuuli-tyokalu-tutkimuksen-datanhallinnan-suunnitteluun

[14] http://www.kansalliskirjasto.fi/fi/julkaisuala/urn.html

[15] https://www.finna.fi/

[16] M. Nurmela and E. Keskitalo, personal communication, August 21st, 2015

research papers (open access or restricted access) and making metadata descriptions about digital and physical publications. This makes librarians and libraries and essential part in bringing research data to the open publishing world.

In an interview with the people responsible of the digital publishing at Aalto University Library[17] the essential role of librarians as the classifiers and describers of the data was brought up. Professional data handlers can do very good metadata descriptions, even if they are missing some of the domain knowledge related to the research data. The librarians also handle the relationships to the publishers - though what is the role of traditional scientific publishing authorities in the future when organizations can publish datasets and even research papers papers easily on their own remains an open question.

Aalto University library runs the Aaltodoc service[18] which contains full text materials on research papers and theses published in Aalto University. The system runs on on DSpace[19], an institutional repository software for publishing digital objects. DSpace focuses on publishing research papers, but the person in charge of the system reckoned that it probably could be modified to host small datasets as well. This would require some additional work of course, so a better way could be to link the research papers the relevant datasets in the corresponding systems.[20]

The National Library of Finland is in charge of implementing long term storage and archival of important datasets and other research material. From that point of view and also research data in general the biggest challenges are not technical - software systems to store data and manage it exist, but making it so that institutions themselves commit to managing and storing data is the bigger challenge. And once different institutions are able to manage their own data, the collaboration between the institutions' systems is likely to be more difficult in the policy and bureaucracy sense. There are also many unresolved questions related to long term storage. Who decides what datasets are used for the long term storage? What kind of metadata long term storage requires in addition to the metadata already in the original dataset? What is the most suitable file format for long term storage, since tools and software used to create it outdated relatively soon? The work to figure out these things and the Finnish long term archival project is going on during the writing of this thesis.[21]

From curation point of view persistent identifiers are very important for all kinds of research outputs. Since research data publishing is a new phenomenon, it's still undecided what kind of identifier should be used with it.[21]

The universities already have datasets stored within their systems and one challenge would be to get these datasets public as well. Manual work on that would be futile, which is why a system that would extract the existing data

[17] M. Nurmela and A. Rousi, personal communication, September 30th, 2015

[18] `https://aaltodoc.aalto.fi/`

[19] `http://www.dspace.org/`

[20] M. Nurmela and J. Nevala, personal communication, August 27th, 2015

[21] M. Nurmela and E. Keskitalo, personal communication, August 21st, 2015

from universities' systems automatically would be useful.[22]

4.5 Course organizers

In addition to research, the mission of universities is to teach. With the world of research moving to the direction of more and more data intensive science there is a need universities need to adapt and offer students the possibility to study with relevant datasets. Aalto University has started offering a minor in Data Science in order to cater to this need.[23]

In an interview with the people in charge of the Aalto Data Science minor it became clear that even though data intensive science is taught, there is no Aalto infrastructure for the teaching. Datasets and the computing power are acquired from vendors which had caused some awkward arrangements since the access to the outside resources had to be controlled more tightly than just on Aalto level. These issues can be worked out though and from the point of teaching it would be nice to have data available from within the university, data and computing resources could also be acquired elsewhere.[24]

The people in charge of teaching also could offer insight to the question about the basic skills that go to data analysis. The question is interesting since in order to leverage the fact that science is becoming more and more data intensive scientists need to possess skills both to analyze their data and manage it in a sufficient way. The skillset that is required for data handling and management is largely programming, since naturally the analysis is computerized and things such as cleaning and preparing data for analysis is most efficient when done programmatically. On the other hand, the data analysis part asks for skills in algorithmics and statistics. When we take into account that the Aalto Data Science minor, for example, is open for students from all fields of science, it seems that we need to be teaching a new set of skills to almost all students that want to take part into the new wave of data intensive science.

4.6 CSC

CSC is the national computing service provider for Finnish institutions. They offer both computing power and disk space for Finnish institutions and they are a state-owned non-profit organization.[25] Their role is interesting when it comes to research data management and publishing, since they already offer computing services to the relevant institutions in Finland. After all, one goal of publishing and sharing research data is to enable collaboration and involving

[22] M. Nurmela and J. Kesäniemi, personal communication, September 2nd, 2015

[23] `http://studyguides.aalto.fi/minors-guide/2015/en/sci/sci-minors-for-all-aalto-students/analytics-and-data-science.html`

[24] M. Nurmela, J. Bragge and P. Malo, personal communication, August 6th, 2015

[25] `https://www.csc.fi/csc`

the one institution in Finland that offers services to all the other institutions makes sense in that regard.

One of CSC's services related to research data management is the IDA service[26] that is specifically designed to store research data and the related metadata. CSC also maintains Etsin[27], a service to host metadata related to research data as well as links to the location of the data. Etsin itself does not contain datasets. These efforts fall under the national project of Open Science[28] (Avoin Tiede ja Tutkimus in Finnish) that promotes the openness of research and science. On the top level these initiatives have been put into motion by the Ministry of Education and Culture[29].

The IDA service, however, has not been very widely adopted as a place to store and share research data. This has to do with IDA's user interface and the fact that the iRODS backend is not designed for publishing information. Issues with policies and permissions also hinder the publication process.[30] In addition to the technical matters there is of course the fact that institutions, such as universities, do not have a very well developed culture for research data management or publishing which also contributes to this.

CSC also implements a service called AVAA[31], which contains spatial datasets such as open street maps for Finland and data from weather stations around Finland. It seems that it is not designed for all kinds of datasets or it has not been adopted for other uses.

4.7 Research data management questionnaires

Separately from this thesis two studies were conducted in Aalto University about research data management and sharing by Ilari Lähteenmäki and Aalto IT services [43, 44]. The goal of the studies was to find out the current needs as well as the current status of research data management within Aalto University. The studies were carried out as questionnaires and the questionnaires were distributed to all branches of Aalto.

The first questionnaire [44] divided the research process in the context of research data into four phases, which are the defining of the research scope, the research, publishing the research data and archiving the research data. 87 people submitted answers to the study from all the schools in Aalto. The breakdown respondents is shown in Figure 4.1.

The biggest challenges reported by the questionnaire were with the actual research process and the archiving of data, as outlined by Figure 4.2. During the scope defining phase of the research the research data challenges mentioned

[26] http://avointiede.fi/ida

[27] https://etsin.avointiede.fi/

[28] http://openscience.fi/

[29] http://www.minedu.fi/OPM/?lang=en

[30] M. Nurmela and S. Westman, personal communication, August 14th, 2015

[31] http://avaa.tdata.fi/en/

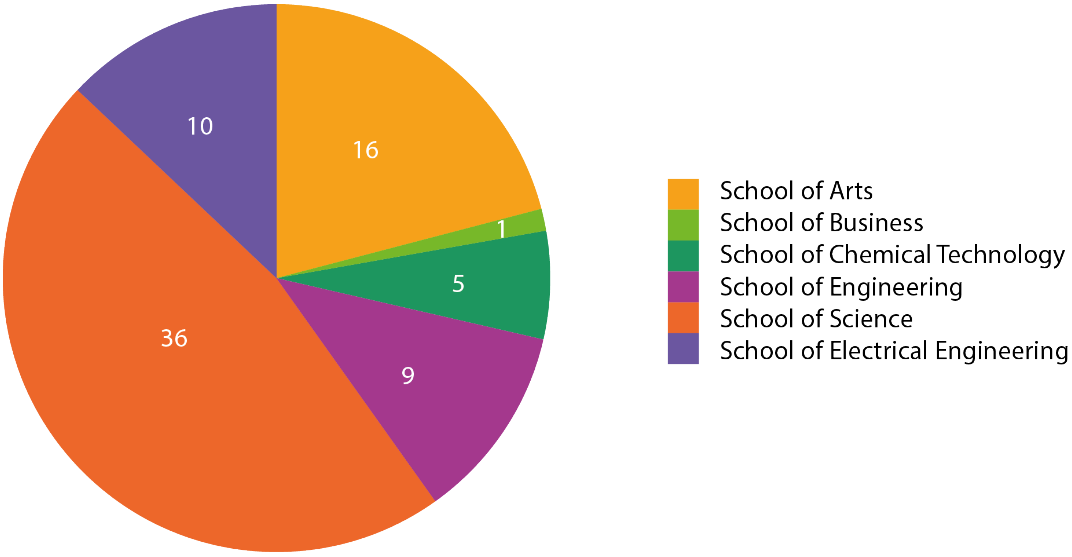


Figure 4.1: The respondents to the first questionnaire, divided by the school of the respondents [44]

were storage, availability and version control. The actual research phase contained many challenges, biggest of which were the lack of storage space, the size and amount of files generated (the current system could not handle them) and the challenges with data availability. The problems with sharing research data came from the lack of sharing infrastructure, version control, storage space (quantity and persistence - no way to persist digital files) and other restrictions, such as privacy concerns or classified data. The single biggest problem with archiving research data was that there is no infrastructure for archiving research data.

The greatest need for services derived from the first questionnaire was reported in storing data, metadata, finding data, archiving, sharing and backing up data. The challenges where the requirements span from are both technological and policy related. In the list below the the answers are compiled by percentages:

- The name of the folder/file or the location of the folder/file is forgotten or poorly described, approx. 35%.
- Ownership of the data, approx. 30%.
- Not enough disk space, approx. 30%.
- Version control, approx. 30%.
- Non functional or corrupted equipment, 29%.
- Sharing data with partners and collaborators, 24%.

- Unwanted deletion of data, 22%.
- Complicated user interface, approx. 20%.
- Forgetting password, 17%.
- Access rights, approx. 12%.
- Failed backup, 8%.

The second questionnaire [43] has 368 respondents, divided into different branches of Aalto as shown in Figure 4.3.

The second questionnaire found that the most important things for research data management and sharing system would be that the research data could be worked on in collaboration, big files that you cannot attach to emails could be shared and that there would be version control for the files. Research data

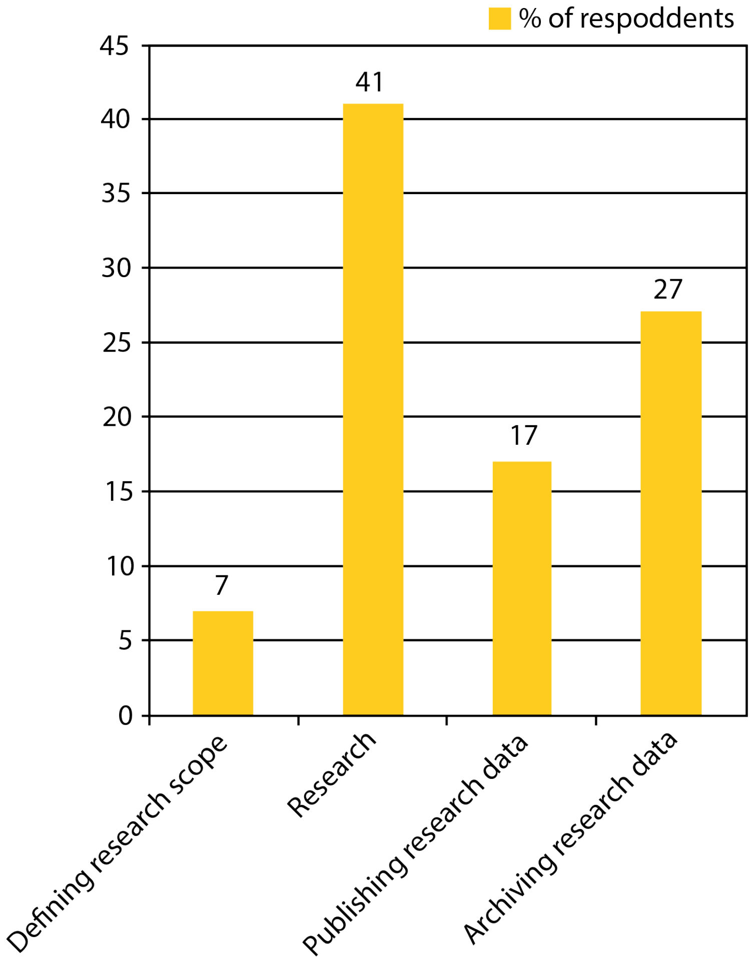


Figure 4.2: Where the biggest challenges lie with research data management, divided by the phase of research process [44]

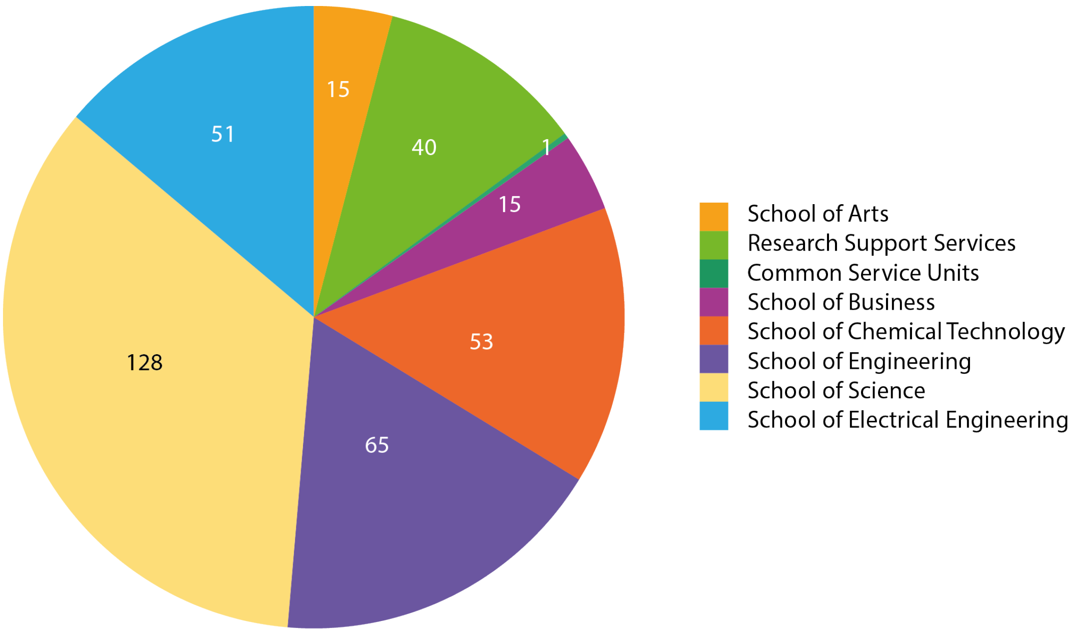


Figure 4.3: The affiliations of the respondents to the second questionnaire [43]

would mostly be shared between Aalto University staff, but sharing with collaborators outside Aalto, sharing with students as well as personal file storage were required of the system. Figure 4.4 shows who the research data needs to be shared with. The research data should be accessible with personal computers in addition to Aalto workstations.

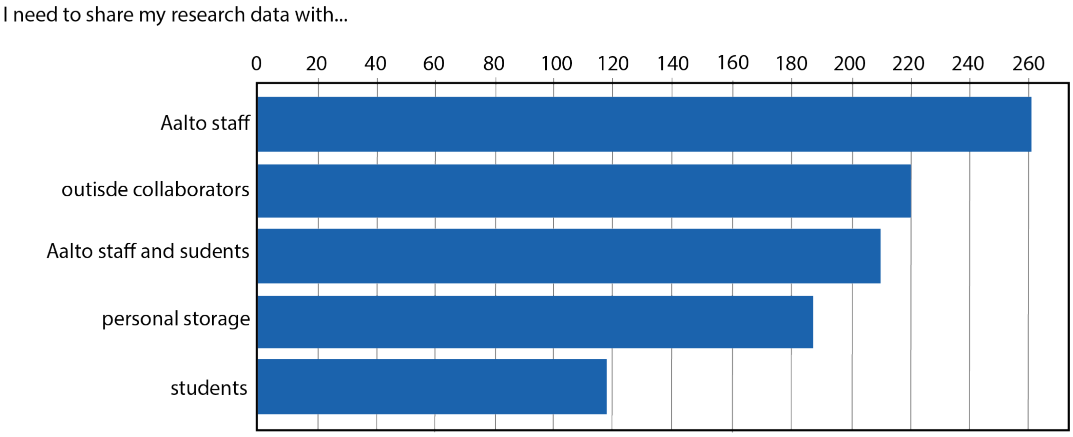


Figure 4.4: The different parties research data has to be shared with - horizontal axis is the amount of respondents [43]

The types of files people would need to save are very varied - the most popular type of files are Microsoft Office style files (text documents and spreadsheets) and PDF files. Different image, video, sound, code and other formats are brought up in the questionnaire, even virtual machine images. The amount of required storage space varies from 10Gt to over 500Gt, as shown in Figure 4.5. The most common storage time for research data is measured in years, since that is the scope of research projects and data management is required throughout the project. The distribution of required storage times is shown in Figure 4.6.

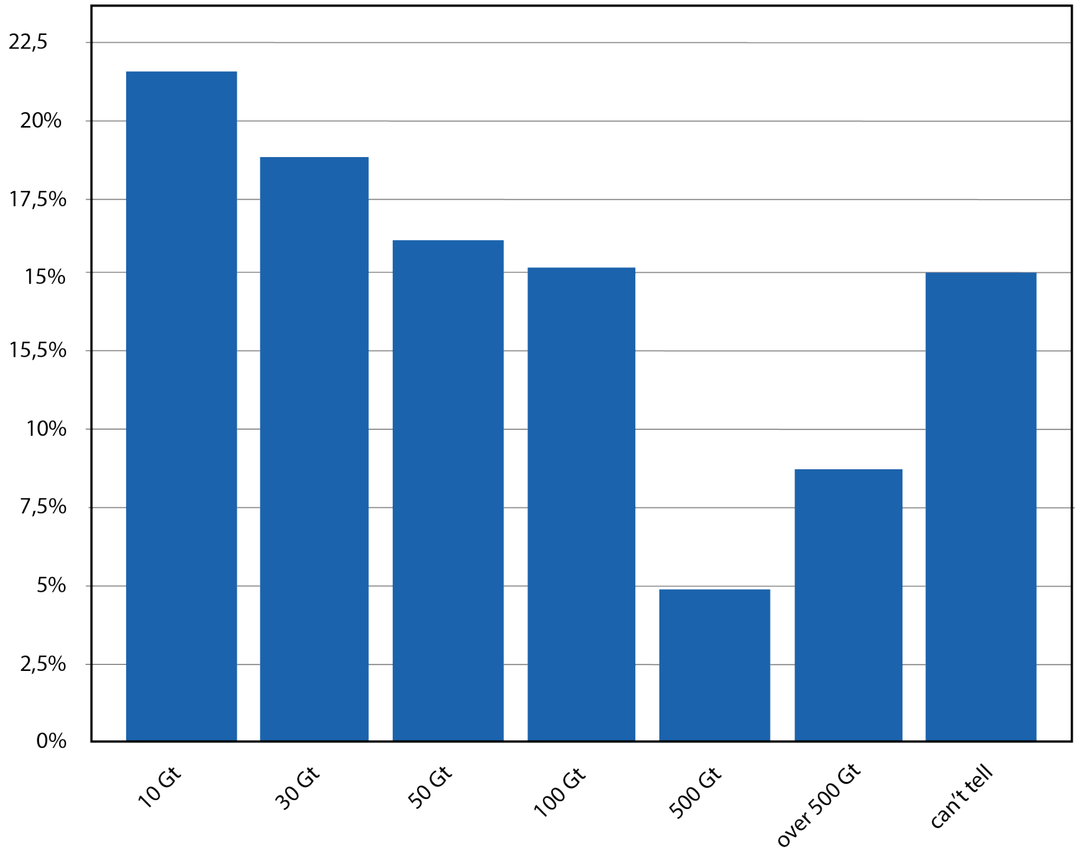


Figure 4.5: The required storage capacities according to the respondents - vertical axis is the percentage of respondents [43]

Most of the respondents of the second questionnaire handle their research data every day and their research data is confidential. Not all confidential data needs to be shared outside Aalto University, but most respondents had to do that from time to time.

Most of the respondents use existing cloud services, such as Google Drive or Dropbox to do research data management during their projects. The systems are used both out of necessity (there is no existing Aalto infrastructure, for example, to have non-Aalto people working on the same datasets or files) and because they are easy to use and people have experience from using them from outside of work. For some, this is not an option, since some data has to be stored in Finland and the cloud providers cannot comply to that request.

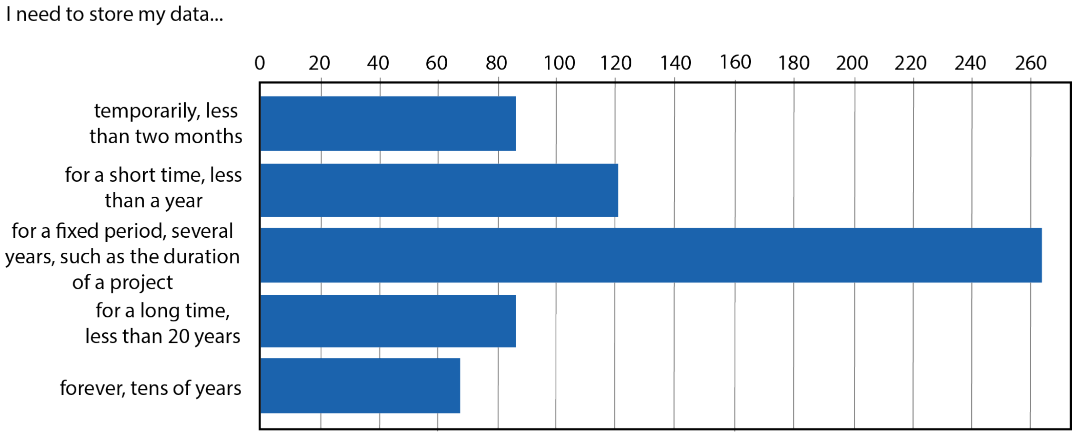


Figure 4.6: The required storage time of datasets - the horizontal axis is the amount of respondents [43]

These studies show that there are indeed a need already in place for a system to manage and share research data in Aalto. In addition training and policies to support data management through the lifecycle of data are required. Existing cloud based solutions, such as Google Drive and Dropbox are the benchmark that people use to compare existing and future solutions to. This sets a bar for the solutions that would be implemented to address research data management and publication challenges.

4.8 Benchmarking existing solutions

Technical solutions to publishing and managing research data exist already, many of them open source and free. Some of the solutions were mentioned in the literature review in Section 3.11. This section describes those systems and some other in more practical detail. The benchmarking was done using online demo versions of the software, reading the related documentation and in some cases examining the source codes and installation processes that they required.

For a more visual overview, see Section 7 for an overview table of existing solutions. The Section 5 contains a more in depth look into the Dataverse system and the installation processes of Hydra Project and Zenodo.

This is not a comprehensive listing of repository systems. Many institutions have implemented their own systems and other, non-institute related systems exist as well. The benchmark solutions presented here have been chosen due to their wide adoption.

Harvard Dataverse is a Harvard University originated open source system for research data publication. This system was benchmarked by both reading the relevant documentation as well as making a local installation for testing purposes - more on the test installation on Section 5. It is implemented with

Java and is available online at GitHub[32]. Dataverse uses Apache Solr[33] to index its PostgreSQL[34] database to facilitate faceted search. For server software you can use Glassfish[35] or Apache[36].

Their website[37] reports that the Dataverse software would be installed in 12 universities around the world. This number does not include, for example, the Jyväskylä Universiy Dataverse which makes it reasonable to assume that there are other Dataverses out in the world. The main Harvard Dataverse[38], which is the biggest Dataverse instance around, contains 59 652 datasets that contain 288 174 files as of writing this thesis.

Dataverse offers a wide range of features. It generates citeable DOIs for datasets published in it and it also allows the administrator of the system to use Handle identifiers[39] out of the box as well. It has a publishing workflow that allows for saving of drafts and review before publishing. Faceted metadata based fulltext search makes datasets discoverable and metadata can be added both to the dataset level and the file level. The access control system built into Dataverse allows permissions to be granted to individuals and groups alike. The access controls can also be integrated with Shibboleth[40], which is in wide use across research institutions across the globe. For a full list of features, source code and other relevant information see the links in the footnotes.

Dataverse allows manual upload of datasets and an API to explore datasets and upload them programmatically. Once published, new versions of datasets can be uploaded and all the versions remain visible. Dataverse also includes the the TwoRavens[41] application for visualizing tabular datasets.

Dataverse is a functional system for publishing research data. It is not a tool for managing research data during the research project, but it could be used even during that time to share research data with collaborators. It also does not offer long term archival options.

CKAN[42] is an open source data portal software that aims to make data available over the Internet. CKAN instance was not installed for the purposes of this thesis, but a demo CKAN cite[43] was tested in addition to the documentation being read. CKAN is implemented in Python and it uses Apache Solr to to index a PostgreSQL database, much like Dataverse.

CKAN offers full text search of dataset metadata and versioning of datasets. CKAN also allows for robust customization and data visualization as well as

[32] `https://github.com/IQSS/dataverse`
[33] `http://lucene.apache.org/solr/`
[34] `http://www.postgresql.org/`
[35] `https://glassfish.java.net/`
[36] `https://httpd.apache.org/`
[37] `http://dataverse.org/`
[38] `https://dataverse.harvard.edu/`
[39] `https://www.handle.net/`
[40] `https://shibboleth.net/`
[41] `http://datascience.iq.harvard.edu/about-tworavens`
[42] `http://ckan.org/`
[43] `http://demo.ckan.org/`

uploads through a web interface and an API. The access controls to the CKAN are coarse grained, public sharing of datasets or sharing withing an organization. CKAN is also promotes its extensibility, and indeed it has been adopted, for example, as the frontend of the CSC's services[44]. What separates CKAN from the other repository solutions is that it offers extensive geospatial features.

CKAN is designed for publishing and sharing data and not for managing data during the project it is being created.

Invenio[45] is a CERN originated digital library management software. It started as the CERN documentation server, hosting over 1 000 000 bibliographic records and is now available at the public domain[46]. Zenodo[47] is a CERN run public instance of Invenio with a thin UI layer on top. The EUDAT B2Share service[48] also runs on Invenio. Invenio installation was tried using the source code of Zenodo to get a better grasp of the system.

Invenio is built with Python and it implements similar features as the other systems discussed earlier in this section. Fulltext metadata search, uploads through an API and web UI and other repository features are present. Zenodo is integrated with GitHub and it gives the option to publish source code as a citeable scientific entity. Invenio itself has the capability to implement different persistent identifier methods - Zenodo has chosen to use DOI and EUDAT B2Share used Handle out of the box.

Invenio is like Dataverse - a tool to publish digital assets but not for managing data during research.

A side note from EUDAT is that they have also implemented a Dropbox-like service called B2Drop[49], which is targeted to researches that want to share research data during the research project. EUDAT project source code for both B2Share and B2Drop is also available online[50].

The Hydra project is an open source digital repository solution. The Hydra solution was also tested during the writing of this thesis by installing and setting up the system in addition to reading the related documentation and trying available instances.

The Hydra project is a Ruby on Rails[51] application built on top of the Fedora repository software[52]. It uses a project called Blacklight[53] for the discovery platform and similarly to Dataverse it uses Apache Solr to index data stored in the Fedora repository. The code for the Hydra project is available

[44] https://github.com/kata-csc

[45] http://invenio.readthedocs.org/en/latest/

[46] https://github.com/inveniosoftware

[47] https://zenodo.org/

[48] https://b2share.eudat.eu/?ln=en

[49] https://eudat.eu/services/b2drop

[50] https://github.com/EUDAT-B2SHARE, https://github.com/EUDAT-B2DROP

[51] http://rubyonrails.org/

[52] http://www.fedora-commons.org/

[53] http://projectblacklight.org/

at GitHub[54] and the developer community maintains a wiki online[55]. The Hydra website lists 29 universities and institutions that have adopted the Hydra software, mostly in the USA.

Hydra project is very flexible - it is used around the world for institutional repositories, museum websites, cultural heritage storage and other similar projects. Extensibility comes with a price, since an out of the box Hydra installation requires the user to configure all the types of data you would need for your chosen repository. Many implementations do exist, since Hydra has been adopted around the world, but despite the fact that these solutions are open source they are still tightly connected to the institutions that implemented them, demanding still work if they were to be adopted to a new institution. Some Hydra versions have dataset versioning implemented.

Hydra project is aimed to archiving and publishing data. It is not designed as a tool for managing data during research projects.

Aalto University has adopted Elsevier Pure[56] as a tool for researchers to manage their publications. The system allows for linking datasets to your profile, but the version that is running at Aalto University as of writing of this thesis does not support actually uploading datasets to a repository. This brings up an integration point for future development, since it would make sense that when you upload your datasets to a repository it should show up in your researcher profile.

iRODS[57] is a data management software used by many research institutions to manage research data. It is open source and available on GitHub[58] along with the many libraries and software components that come with it. iRODS virtualizes storage hardware, which in part constitutes to the amount of different software packages that go into it, since different hardware requires different drivers and many contributors have implemented their own packages for iRODS. For the purposes of this thesis an instance of iRODS was not installed, but experts and users of the system were interviewed and documentation was read along with the code.

In addition to virtualizing the storage hardware iRODS offers data discovery by compiling metadata about the files and folders in the system to a metadata catalog. The iRODS rules engine allows the users to program workflows and automated events to the system. Newer versions (since iRODS 3) of iRODS allow iRODS systems to be federated, which means that two iRODS systems from different origins can share data easily.

iRODS is implemented in C++ and it offers client APIs for multiple programming languages. iRODS also scales from single personal computers into large data grids.

In Finland iRODS is in use at least in the University of Jyväskylä and in the

[54] `https://github.com/projecthydra`

[55] `https://wiki.duraspace.org/display/hydra/The+Hydra+Project`

[56] `https://www.elsevier.com/solutions/pure`

[57] `http://irods.org/`

[58] `https://github.com/irods`

systems of CSC (the IDA system mentioned in Section 4.6 uses iRODS). The iRODS systems in Finland focus on enabling researcher collaboration, which is the role iRODS is primarily designed for. The iRODS architecture is not designed for sharing data, but could be used as a part of a system that shared data.

GitHub[59] has been mentioned a few times as the place where the source code of many of the projects mentioned here resides. It is a close analogue to sharing research data, since it is a platform to share formatted information between collaborators and anyone who might be interested in the source code. Being a widely adopted platform it also contains, for example, lists to datasets[60] available around the globe as well as datastreams[61] that you could use as a basis of applications serving greater public and good. The Zenodo connection also makes source code citeable.

GitHub hosts, according to their own statistics as of writing of this thesis, 30.1 million code repositories. While the main appeal of GitHub is to allow software to be written collaboratively (GitHub could be seen as the collaborative extension to Git[62]), the version control software. GitHub also hosts some data and visualizes that data. What really sets GitHub apart from the research data sharing platforms is the social aspects it contains. GitHub has powerful issue tracking systems, commenting capabilities and it promotes good coding style and contribution by rewarding the contributors with recognition.

GitHub is a Ruby on Rails application with Git and the required C daemons integrated to it. Git makes it also tool for managing code during the research project, making GitHub a solution for the entire lifetime of source code. There is also GitLab[63], that enables institutions to set up their own GitHub-like systems.

[59] https://github.com/

[60] https://github.com/caesar0301/awesome-public-datasets

[61] https://github.com/HazeWatchApp

[62] https://git-scm.com/

[63] https://about.gitlab.com/

4.9 Solutions in Finland

In Sections 4.1 - 4.6 the current situation in Finland was touched from different perspectives. The Table 4.1 below summarizes the learnings from different actors in Finland as well as some additional findings. This list is not comprehensive but does contain the biggest institutions. It is is likely that other institutions and the institutions listed below are working on more matters. All Finnish universities maintain a digital publishing archive for papers and theses published in those universities - they are not listed below. The footnotes are on the next page.

Table 4.1: Actors in Finland and their actions as of writing of this thesis

Actor	**Work currently underway**
Ministry of Education and Culture	The Open Science and Research Initiative[64], which entails services for publishing research data and metadata (mentioned before in Section 4.6). They also organize training to facilitate open research and research data in Finnish institutions. The Initiative also runs the Tuuli project, that aims to help researchers to make data management plans that are required nowadays by funding bodies.
CSC	As the national provider of scientific computing services CSC works with all the major Finnish institutions. It is also the implementing force in many of the Ministry level projects. CSC is also the the Finnish contact point to EUDAT, the European Union level research data services.
University of Helsinki	University of Helsinki has formulated its data policy[65] and is working towards implementing the infrastructure required to support it.[66]
University of Jyväskylä	Unviersity of Jyväskylä has an own Dataverse instance and they are using iRODS as their research data management tool. They are also developing a desktop interface for iRODS to enable researcher collaboration.
University of Tampere	University of Tampere operates the Finnish Social Science Data Archive[67], an resource center funded by the Ministry of Culture and Education to store datasets from the field of social science.
National Library of Finland	The National Library manages the URN identifier scheme that can be used to get persistent identifiers to datasets and other scientific material in Finland. It's also involved in the long term storage project (PAS). It also runs Finna[68], a digital archive for Finnish museums, archives and libraries.
Aalto University	Aalto University is forming its research data policy.

4.10 Outcomes of the positioning research

It is clear that research data management and publishing is not just a technical problem. Perfectly fine technical solutions for publishing, sharing and managing research data exist. However, a system that would integrate these and make the whole lifecycle of research data - from creation to archiving - does not exist. The challenge of managing, sharing and publishing research data brings together experts from multiple fields (researchers themselves, support staff in librarians and science IT and the managers who oversee these transaction to name a few) and this collaboration has not been figured out yet. Without proper technical solutions there definitely can not be proper research data management, but policies and culture have to be developed as well. There is little culture or reward to going through the trouble of properly managing and publishing research data.

It also seems that the problem should not be solved independently by all the institutions in Finland, let alone in the world. Successful research data publishing solutions are open source software projects, that offer the base solution for free and the institutions can then become part of the developer community simultaneously improving the software and adapting it to their own needs. Similarly iRODS, the tool for managing research data, is open source. In addition to using similar tools as their peers different institutions should consider also forming their policies and research data management guidelines to be compatible. After all, one goal of sharing and publishing research data is to make collaboration easier and make science better.

The insights gained from this positioning research and the learnings from the literature review in Section 3 and the next Section 5 will be discussed together in Sections 6 and 7.

[64] http://openscience.fi/frontpage

[65] https://www.helsinki.fi/sites/default/files/atoms/files/datapolicy_final_en.pdf

[66] M. Nurmela and V. Tenhunen, personal communication, November 9th, 2015

[67] http://www.fsd.uta.fi/en/index.html

[68] https://www.finna.fi/

Chapter 5

Prototype Solution

In Section 4.8 existing solutions for the problems of publishing, sharing and managing research data are presented. It is clear that since these solutions exist, there is no point in reinventing the wheel and implement a new system. Instead we have implemented a local installation of some of the systems presented in the benchmarking section. From the test installations we have chosen the one that works the best and used that to run tests on potential users of the system. The tests are used to gain insights on what the finalized system should look like. It is also notable that the existing solutions are remarkably similar as noted in as noted in Section 4.8, so using any one of them would give applicable results. The prototype solution focuses on publishing and sharing of research data. The other option would have been to focus on the research data management during the research project, but research projects last longer than the span of this thesis and the results gained from that avenue of research that would likely be quite superficial. The lack of culture and practices are a factor for both publishing and and managing research data. The lacking of research data management culture is due to the lack of education and need for it, whereas publishing research data is a moderately new phenomenon and the culture is still being formed. This makes it a more novel subject of study. Increased demand for research data sharing and publishing would also force research data management practices to be developed further.

We ended up choosing the Harvard Dataverse solution to be the prototype for our purposes. The following sections detail the rationale behind this choice, the technical details of the system and the tests that were conducted using the system along with the learnings.

5.1 Rationale behind selecting Dataverse

As a part of the benchmarking the existing solutions and in order to select the right tool to run tests on users we tried installations of a Hydra head (Hydra head is a Hydra instance in the Hydra Project terminology), a Zenodo instance and a Harvard Dataverse instance. These three were chosen because they represent different technologies and are widely adopted as tools for publishing research data.

Setting up a Hydra head is fairly simple using Ruby Gems[1]. Setting up the basic Hydra head does not get you far, however, since after setting up the installation you need to define your data model and almost everything else on

[1] https://github.com/projecthydra/hydra-head

your repository.
This setup cost makes Hydra a very versatile framework. It is being used on many places beyond just research institutions, such as museums and image repositories[2]. Many of these systems are built on Hydra solution bundles[3] which are also are open source. Installations of a clean Hydra head and the version run by Penn State University[4] were tried.
The conclusion about Hydra heads was that while the system is modern and quite easy to install, the setup of the system made it too time consuming to setup a testing prototype in a reasonable time frame. The system is flexible and if you wanted to build your own customized repository solution Hydra would be suitable for that. The Penn State implementation was heavily branded and tweaked for their purposes, making it hard to make it work for prototyping purposes. Blacklight[5], the frontend library used by the Hydra project, is quite good and makes for easy to use and efficient frontends.
We tried installing Zenodo system locally from the source code[6], but could not get the build process to work correctly as of writing of this thesis. It was later found out that the Zenodo system, which is built upon the Invenio archiving software, is notoriously hard to install according to the people who originally built it[7].
Due to the problems with the installation the Zenodo system was ruled out at the prototyping phase.
Harvard Dataverse is easy to install with the installation instructions as both a development version from the source code[8] and a production version with the installation bundle[9].
The easy installation immediately gave a functioning software repository to conduct tests with and that lead to the decision to use Dataverse as the prototype to test current data repository solutions and gain user feedback to supplement the other research.
Though implemented in different technologies, the functioning of the existing research data repository systems is quite similar. All of them offer form based dataset uploads, full text searches and some forms of access control. Many of them are even built on same technologies, such as Solr indexing software[10] or postgreSQL[11].
The similarity of the systems as well as the fact that there is no global consensus on what repository software is the best in business hints that you could use

[2] https://wiki.duraspace.org/display/hydra/Partners+and+Implementations

[3] https://wiki.duraspace.org/display/hydra/Hydra+Solution+Bundles

[4] https://scholarsphere.psu.edu/

[5] http://projectblacklight.org/

[6] https://github.com/zenodo/zenodo

[7] M. Nurmela and D. Lecarpentier, personal communication, September 30th, 2015

[8] http://guides.dataverse.org/en/latest/developers/index.html

[9] http://guides.dataverse.org/en/latest/installation/

[10] http://lucene.apache.org/solr/

[11] http://www.postgresql.org/

any one of them in your organization. From this angle it also makes sense to use one of them to gain user insights and figure out how the systems should be developed in order to satisfy the user needs better.

5.2 Users of the system

As examined in Section 4, a research data repository systems have many stakeholders. Identified key stakeholders are presented in the following:

- Researchers,
- University courses,
- Research groups,
- Librarians,
- Students,
- IT staff; and
- Other interested parties.

The requirements of these different stakeholders are boiled down to user stories, which are presented in Appendix A.

5.3 Prototype system description

Harvard Dataverse is a Java application. Other technologies employed are Apache Solr[10] for indexing the database to facilitate search, postgreSQL[11] for database and Glassfish or Apache for serving the web pages[12]. Dataverse also has in built support for the R statistical computing language[13] for running simple statistical analyses on the data and for data visualization using the TwoRavens tool [35].
The class model of Dataverse[14] is described in Figure 5.1.
In the heart of Dataverse is the division of content into Dataverses. The closest analogue to a Dataverse is a normal folder in a typical file system - Dataverses can contain other Dataverses, but in the place of files Dataverses contain datasets. Datasets, in turn, contain the files that make up the dataset. The Dataverse split of the system also allows for fine grained access control, since Dataverses can be shared with no one, with single users or user groups.
Users can use the Dataverse with either the web user interface or the API offered by Dataverse. The dataflow and interplay of the different components of the Dataverse application[14] is shown in Figure 5.2.

[12] https://glassfish.java.net/, https://httpd.apache.org/

[13] https://www.r-project.org/

[14] https://github.com/IQSS/dataverse/tree/4.3/doc/Architecture

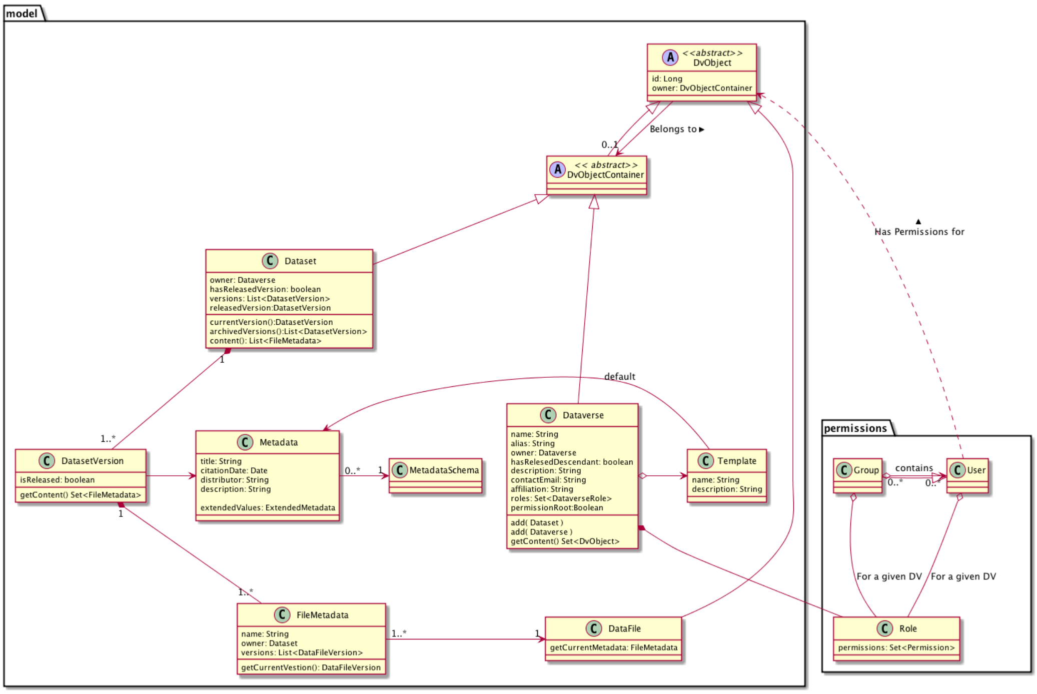


Figure 5.1: The Dataverse class model[14]

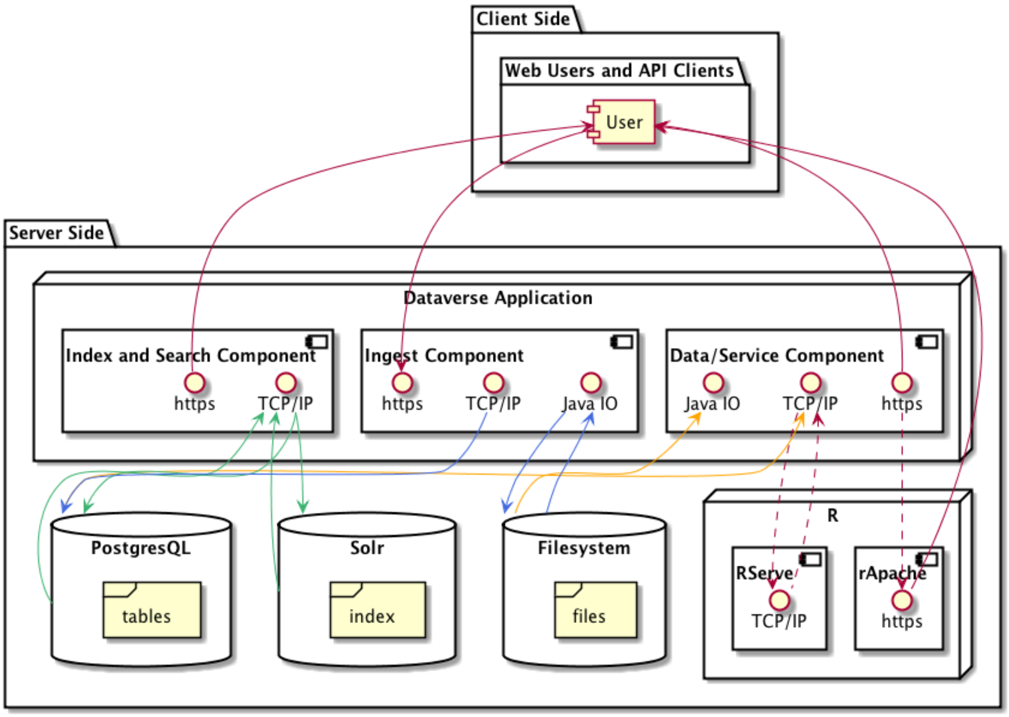


Figure 5.2: The Dataverse application dataflow[14]

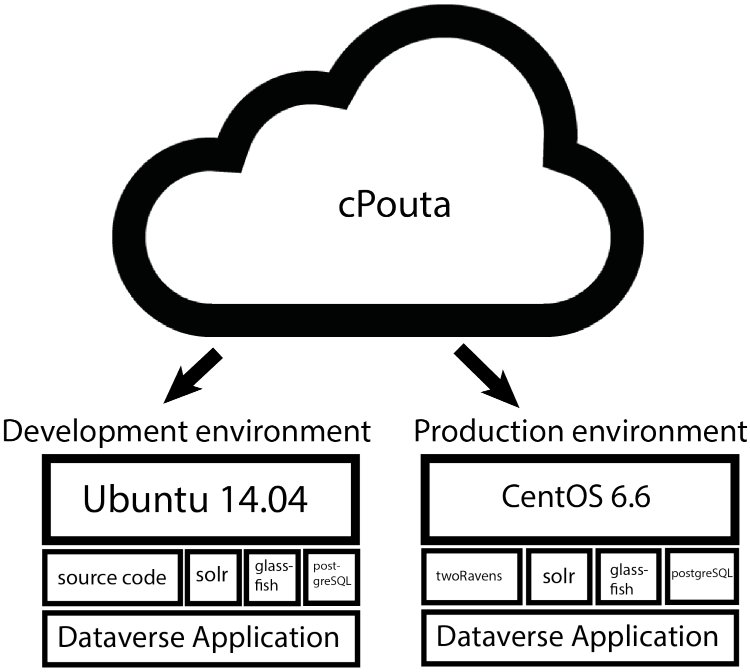


Figure 5.3: The different installations in the cPouta environment

The Dataverse Java application encompasses the Client Side and Dataverse Application in Figure 5.2. Ingest refers to uploading datasets to the system and the Index and Search and Data/Service components serve the user the desired content, be it search results or data to be downloaded. The rApache component handles the data visualization and runs the TwoRavens tool, and RServe is used for the statistical computation when the user requests that.

Two versions of the Dataverse system were installed - one from source code[15] and one from the installation bundle provided by the developers of Dataverse[16]. The installations were run in the CSC cPouta environment[17], which is an OpenStack instance[18]. The source code installation is referenced from now as the development installation. It was used to get a feel of the code and the system quality and the access to the source makes debugging weird situations easier. The installation from the installation bundle is referred to as the production installation. The installation bundle should provide a more stable system than the branch of development code that was forked for the development installation. Figure 5.3 shows the different installations in the cPouta environment.

The development installation is installed on top of Ubuntu 14.04 and it works, even though the installation instructions propose the use of Red Hat based

[15] https://github.com/quarian/dataverse

[16] https://github.com/IQSS/dataverse/releases/tag/v4.2

[17] https://research.csc.fi/cpouta

[18] https://www.openstack.org/

systems. The TwoRavens application is omitted from the development installation, since the data visualization is not the core functionality of the research data repository software. The production installation is done on CentOS 6.6, which is a derivative of Red Hat Linux. The installation was first tried on CentOS 7.0, but the differences between CentOS 6.x and 7.x made it so that the installation would not work - CentOS 6.6 was settled to be the final environment of the development installation.

As for the information security of the prototype solution, the development installation is not set up with any firewall rules or other security, since its purpose is to get a feel of the system. For the production installation firewall rules are set using the security groups functions of OpenStack. Inbound TCP/IP traffic was only allowed to port 8080, which Glassfish was listening to. When it comes to research data, security is important, and since the technologies such as Glassfish and postgreSQL are well known software and their default passwords and ports are well known setting up firewalls and changing those passwords is imperative. Additionally, with Dataverse, firewalling the port that Apache Solr uses is important, since it circumvents the user credentials and it could be used to retrieve any indexed information in the system.

To replicate the test systems follow the installation instructions[19] and the description here.

5.4 System test setup

In order to extract value from the prototype Dataverse it needed to be tested with actual users. Of all the stakeholders groups presented in Section 5.2 the research scientists is the most important one, since without them there is no research data and without them using the research data repository there is no public research data. Section 4 summarizes learnings from the other stakeholder groups. Section 4 also contains results from surveys conducted on research data management and sharing to provide context also from the point of view of researchers.

To test this system we worked together with the Complex Networks Group[20] and the Speech Group[21] of Aalto University. Two kinds of tests were designed and implemented. The first test was a contextual interview - which means an interview and observation conducted in the user's normal working environment. The contextual interviews were conducted both with the development installation of the Dataverse as a tool for discussion and as conversations and observations about the current state of research data management and the current working practices. The second test was conducted with two lead users and the production Dataverse installation. The users were asked to input their datasets to the Dataverse and fill in the appropriate metadata.

[19] `http://guides.dataverse.org/en/latest/installation/`

[20] `http://becs.aalto.fi/en/research/complex_networks/`

[21] `http://research.ics.aalto.fi/speech/`

The contextual interviews focused on the current methods of research data management and sharing. The users were asked to describe their practices and how they had come across to them (taught by the university, learned on their own or some other methods). When applicable, the users were asked to show their current setups for research data management and sharing. When the development Dataverse was used the users were asked to upload a test dataset to the Dataverse and walk the interviewer through the thought process. In addition the interviewees were asked to use the actual Harvard Dataverse[22] to find datasets relevant to their field of study and talk through the process. All these interactions were also observed to find out usability and other issues that might arise during the exchanges. In total 10 members of the Complex Networks group were involved in the contextual interviews.

The lead users were granted access to the development Dataverse and were briefly instructed to the different functionalities of Dataverse. The instructions were left vague in order to make them read the relevant user guides and provide feedback on how easy the system was to use after only very brief instructions. After roughly a month's time the lead users were debriefed and interviewed about their experiences with the Dataverse system.

The goal of these tests was to understand the current status of research data management and publishing, how a research data publication system would fit this status and how well does the research data publication system fill its designated role. The usability of the system is also a point of interest.

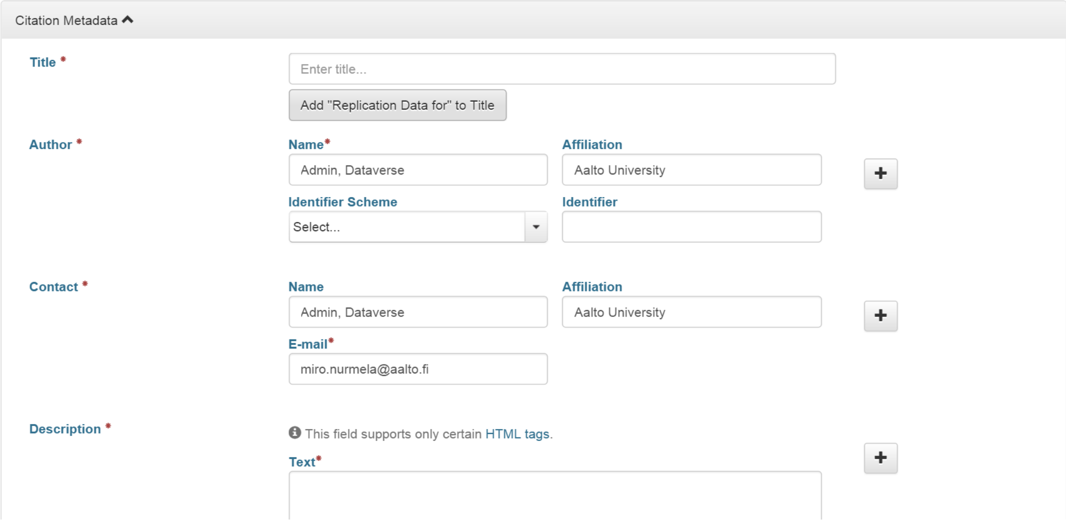
In addition the installation and maintenance of the prototype installations would give insight on how the system would be to maintain and how it would work from a technical point of view.

5.5 Outcomes of the prototype and the tests

The contextual interviews and lead user tests yielded results on technical matters, user interface and experience matters and on the current status of the research data management and publishing.

The results of the tests showed that the technical implementations for research data sharing fill their role. In the contextual interviews it was clear that the manual uploading and searching worked fine - the users were able to pick up the process of searching and uploading datasets quickly. The upload process is form based, shown in Figure 5.4 which is similar to many form based upload pages you can find online. Files for datasets can be dragged and dropped to the user interface or the file browser of the computer can be used. The users were able to find the suitable method for them from the two options. But even though the upload process is mechanically easy, the upload form contained a part where the user was supposed to insert keywords of the dataset being uploaded. The keywords come with the concept of vocabulary that was not

[22] https://dataverse.harvard.edu/



Citation Metadata

Title *

Enter title...

Add "Replication Data for" to Title

Author *

Name*: Admin, Dataverse

Affiliation: Aalto University

Identifier Scheme: Select...

Identifier:

Contact *

Name: Admin, Dataverse

Affiliation: Aalto University

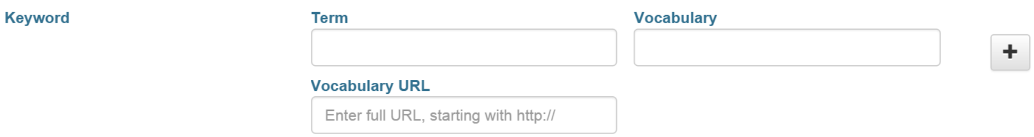
E-mail*: miro.nurmela@aalto.fi

Description *

This field supports only certain HTML tags.

Text*

Figure 5.4: A screen capture of the dataset upload form



Keyword

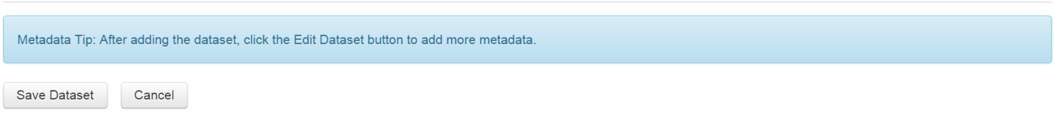
Term:

Vocabulary:

Vocabulary URL: Enter full URL, starting with http://

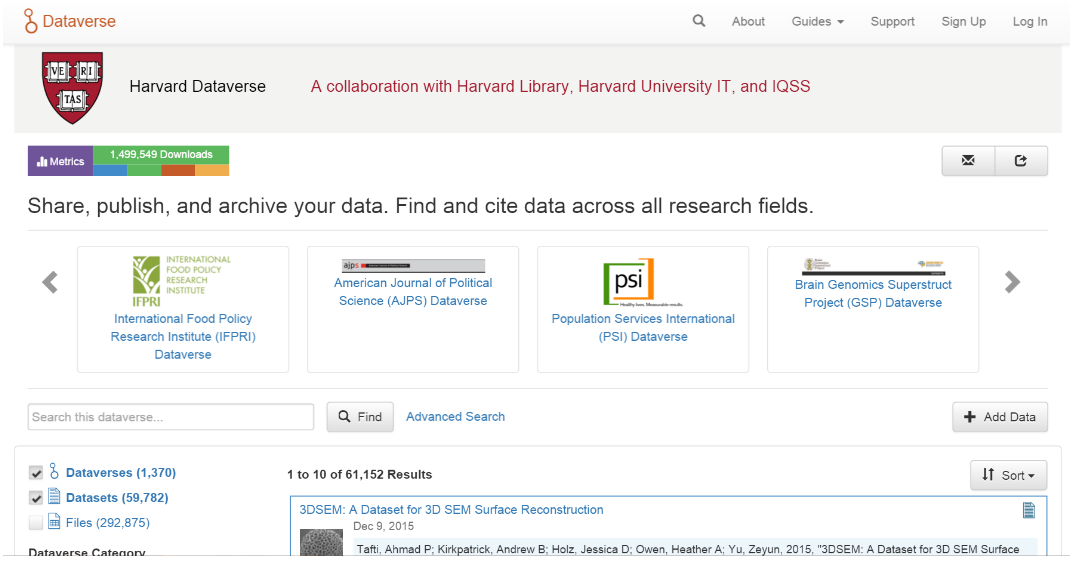
Figure 5.5: A screen capture of the confusion inducing vocabulary in the upload form

known to the users. The point of confusion is shown in Figure 5.5. In this context vocabulary refers to a set of words that have been agreed upon within a field of science in order to standardize communication within the field.
What is lacking of the upload process was the extended metadata that would be important for the reuse of the data. The upload form contains the basic metadata that is required to make the dataset searchable. Additional metadata, such as details about the process of gathering the data, needs to be filled in after the initial upload. The upload form contains a reminder, pictured in Figure 5.6, to go add the metadata later. This approach has pros and cons, since it makes the upload process easier but makes the metadata likely lacking. Most users did not notice the hint to go add the metadata later and those who did notice it did not know where the metadata addition would be done.
The search functionality is placed front and center in the user interface of the Dataverse main page, as shown in Figure 5.7. Users easily locate the search bar and are able to start searching for relevant datasets. The search results leave users lacking, however. This is not entirely the fault of the search system, since the users found the names and short descriptions of the datasets poor at describing the datasets which made finding relevant datasets hard. Using the advanced search, which all the users did not find, helps narrow down the



Metadata Tip: After adding the dataset, click the Edit Dataset button to add more metadata.

Figure 5.6: A screen capture of the metadata reminder



Dataverse
About
Guides
Support
Sign Up
Log In
Harvard Dataverse
Metrics
1,499,549 Downloads
International Food Policy Research Institute (IFPRI) Dataverse
American Journal of Political Science (AJPS) Dataverse
Population Services International (PSI) Dataverse
Brain Genomics Superstruct Project (GSP) Dataverse
Search this dataverse...
Find
Advanced Search
Add Data
Dataverses (1,370)
Datasets (59,782)
Files (292,875)
Dataverse Category
Sort

A collaboration with Harvard Library, Harvard University IT, and IQSS

Share, publish, and archive your data. Find and cite data across all research fields.

1 to 10 of 61,152 Results

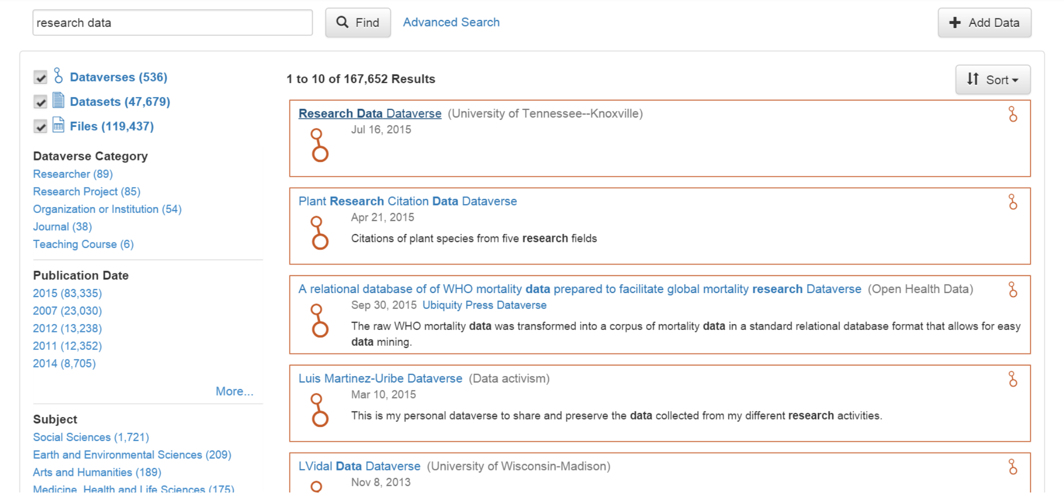
3DSEM: A Dataset for 3D SEM Surface Reconstruction
Dec 9, 2015
Tafti, Ahmad P; Kirkpatrick, Andrew B; Holz, Jessica D; Owen, Heather A; Yu, Zeyun, 2015, "3DSEM: A Dataset for 3D SEM Surface

Figure 5.7: A screen capture of the Harvard Dataverse main page

options. The options in the advanced search were noted to miss details. Details the users were missing were dataset size and metadata related to their field. It was noted that the system would work very well if you knew what you are looking for, for example a scenario where your fellow researcher would have shared the identifier of the dataset.

The order in which the results are displayed is also unclear to the users. The search results page offers the chance to sort the results, but that option was missed by all users during the tests. The default setting is to sort the results by relevance, but it is unclear to what relevance refers to in this context, since the default search searches all searchable fields in datasets, Dataverses and files. The fact that Dataverses, datasets and files are by default mixed in the search results also does not help in finding relevant datasets. There is the option to filter the files, datasets or Dataverses out of the search results, but the option was fairly commonly missed by the users. It might be due to the novelty of the terminology (Dataverse, especially, is a novel concept for the users). An example of search results is shown in Figure 5.8.

Dataverse implements helpful hover texts on the terms, as shown in Figure 5.9. This feature went unnoticed by all users at first, but most of them would find the feature by accident and find it helpful. It would help if the helpful hover



research data

☑ Dataverses (536)
☑ Datasets (47,679)
☑ Files (119,437)

Dataverse Category
Researcher (89)
Research Project (85)
Organization or Institution (54)
Journal (38)
Teaching Course (6)

Publication Date
2015 (83,335)
2007 (23,030)
2012 (13,238)
2011 (12,352)
2014 (8,705)
More...

Subject
Social Sciences (1,721)
Earth and Environmental Sciences (209)
Arts and Humanities (189)
Medicine, Health and Life Sciences (175)

1 to 10 of 167,652 Results

Research Data Dataverse (University of Tennessee--Knoxville)
Jul 16, 2015

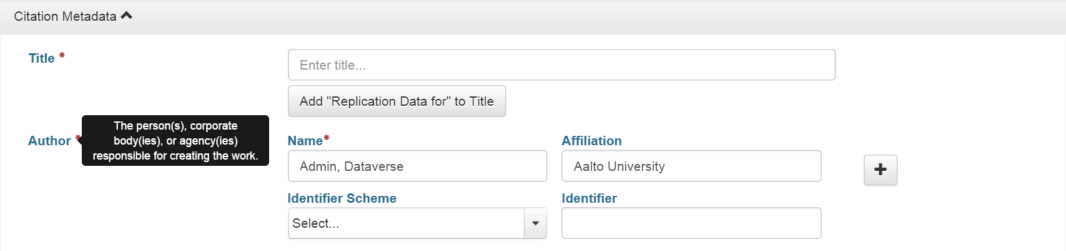
Plant **Research** Citation **Data** Dataverse
Apr 21, 2015
Citations of plant species from five **research** fields

A relational database of of WHO mortality **data** prepared to facilitate global mortality **research** Dataverse (Open Health Data)
Sep 30, 2015 Ubiquity Press Dataverse
The raw WHO mortality **data** was transformed into a corpus of mortality **data** in a standard relational database format that allows for easy **data** mining.

Luis Martinez-Uribe Dataverse (Data activism)
Mar 10, 2015
This is my personal dataverse to share and preserve the **data** collected from my different **research** activities.

LVidal **Data** Dataverse (University of Wisconsin-Madison)
Nov 8, 2013

Figure 5.8: A screen capture of an example search results page



Title *
Enter title...

Author *
The person(s), corporate body(ies), or agency(ies) responsible for creating the work.

Name * : Admin, Dataverse
Affiliation: Aalto University
Identifier Scheme: Select...
Identifier:

Figure 5.9: A screen capture of the hover texts implemented in Dataverse

text would be indicated, for example, by a question mark symbol next to the item to be explained.

What was common in both the contextual interviews and the lead user tests was that the concept of publishing research data is novel for all the users. None of them had published their research data online and few had used datasets from others, but in those cases the datasets were always acquired by request and then delivered using existing systems, such as Google Drive or email. When questioned on if they could make their research data public, for example with a tool like Dataverse, many expressed that their research data had some privacy concerns or other limitations to sharing or that it would take a considerable amount of time to apply relevant metadata to make the datasets presentable. Source code from some projects had been published in GitHub.

The contextual interviews and the lead user tests also shed light on the current ways of managing research data. Network drives were used by the users and their research groups to share all the data within the group, but the there were no clear guidelines on how to describe dataset metadata or how to organize the file structure on the network drives. This meant that it was at times hard for the researchers to find relevant things from the drives and for the new members

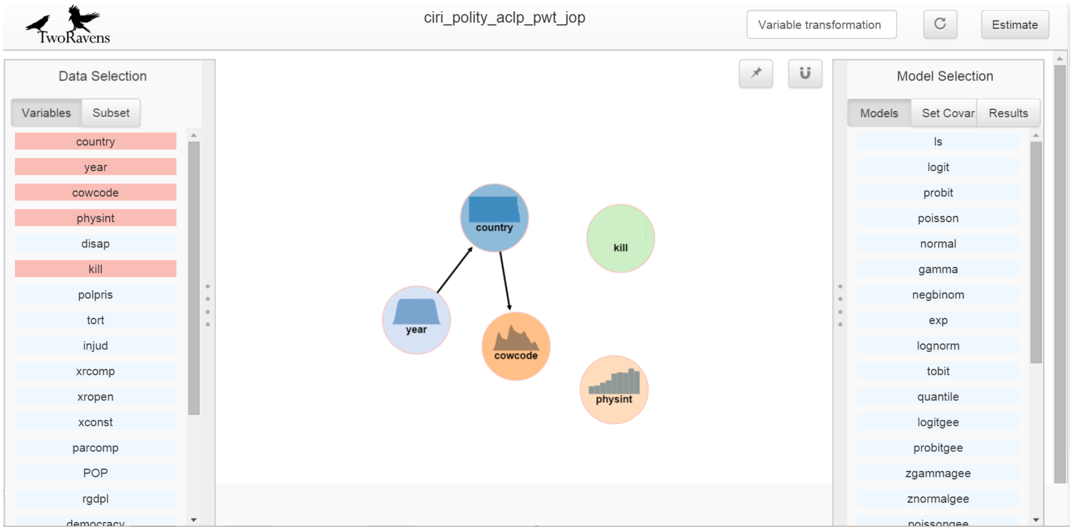


Figure 5.10: A screen capture of Data visualization implemented in Dataverse

of the groups to get acquainted with the system. It turned out that currently research data management is not taught to the researchers, but knowledge is gathered from peers and mistakes you make on your previous projects.

The research data visualization, which came up as an interesting feature to have in a research data publication platform during both the initial interviews and the contextual interviews is implemented in Dataverse. The user feedback on it, however, was that the visualization is very confusing and despite the users being well versed in statistical methods could not understand what the information displayed on the screen meant and how to get valuable information out of it. Figure 5.10 shows a view of the visualization tool.

The lead user tests mirrored the results from the contextual interviews when it came down to the manual upload process of the datasets. Despite the lack of previous knowledge on how to use the system or on how to publish research data the lead users were able to learn the system and upload their datasets. After working with the system for a longer time the need for computerized uploading process became clear. For example, the user from the Sound Group had hundreds of video files that could be uploaded to the publication system, but it would not make sense to spend the time to do that manually. In addition some of that data was being generated every day, so if that was to be published daily it would require an automated system. The Dataverse offers a an API and a guide for using it[23]. The user found that the guide with only example commands was not enough to make the use of API easy. Coincidentally there is research about research video data sharing which could be useful in the future when video data needs to be shared [59].

After using the system for a longer time the users noted that the system could

[23] http://guides.dataverse.org/en/4.2/api/

help them access their datasets from outside the university's network. The network drives and similar approaches in used nowadays limit the access to the systems to the university network, which hinders both the ability of the researchers to work on it from different computers as well as the ability to share the information with collaborators from other places. There are, of course, ways to access data within the university network with VPN or other similar technologies, but that adds complexity to the process.

The longer period of use also brought up the fact that there is information in how people store their data. For example, the folder structure of the network drive might split files into subfolders by date of data collection. This serves as implicit metadata, since there is no metadata file that tells that this piece of data is from a certain date, but a human using the system would understand it. Dataverse offers a chance to upload .zip file to the system and this way preserving folder structures. In the case of the Sound Group the video files are so large and there are so many of them that uploading a huge file would make the reuse of that file very hard, not to mention the uploading and downloading the file.

The installation and maintenance of the prototype Dataverse installation had both flaws and good things to it. The basic installation following the installation guide was easy, but adding in components like the TwoRavens or Shibboleth authentication system was documented less clearly to the point that it was not worth the effort to install the Shibboleth system for the prototype solution. The TwoRavens installation process seemed to work, but the system did not end up working in the prototype solution, so the TwoRavens was tested with the actual Harvard Dataverse.

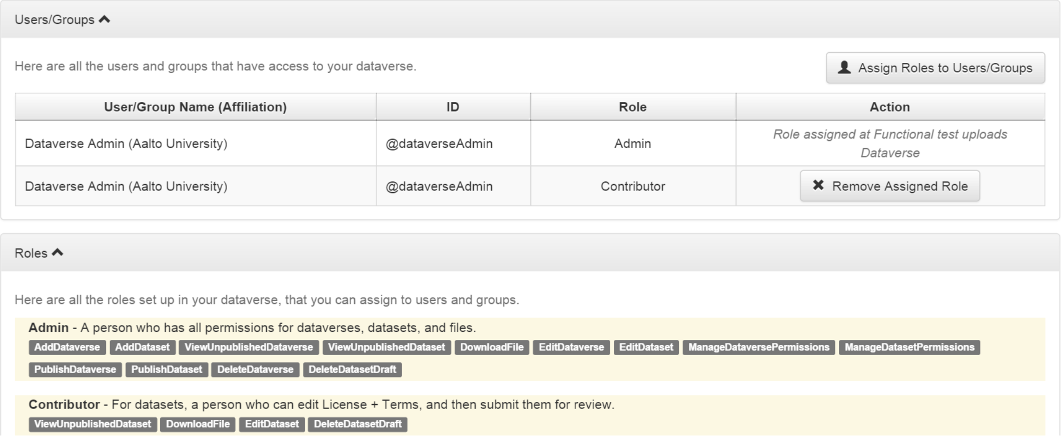
Codewise the Dataverse implementation is clean - it follows good object oriented programming practices. Functionality is split into different classes and the methods are short and well named. Dataverse has unit tests implemented as well as a Jenkins environment[24] for automated integration tests. The test coverage is not very high - only 5%[25].

User management with the Dataverse tools was easy. Dataverse offers a default set of roles that cover the needs of research data publication process well. Custom roles can also be made if the default ones are not enough. Figure 5.11 shows the user management view of a single dataset. Permissions can be changed for individual or groups, allowing the administrator of the dataset to allow different parties to see the dataset before it is published. Publishing will make it visible for all, but managing the permissions before publishing the desired functionality of private publishing could be achieved.

During the test on the development installation we came across an interesting situation when the partition where the database of the Dataverse became too full and no new datasets could be uploaded to the system. This did not crash the system - the Dataverse could still be accessed and datasets be downloaded,

[24] `https://build.hmdc.harvard.edu:8443/`

[25] `https://coveralls.io/github/IQSS/dataverse`



Users/Groups

Here are all the users and groups that have access to your dataverse.

User/Group Name (Affiliation)	ID	Role	Action
Dataverse Admin (Aalto University)	@dataverseAdmin	Admin	*Role assigned at Functional test uploads Dataverse*
Dataverse Admin (Aalto University)	@dataverseAdmin	Contributor	✖ Remove Assigned Role

Roles

Here are all the roles set up in your dataverse, that you can assign to users and groups.

Admin - A person who has all permissions for dataverses, datasets, and files.

AddDataverse AddDataset ViewUnpublishedDataverse ViewUnpublishedDataset DownloadFile EditDataverse EditDataset ManageDataversePermissions ManageDatasetPermissions PublishDataverse PublishDataset DeleteDataverse DeleteDatasetDraft

Contributor - For datasets, a person who can edit License + Terms, and then submit them for review.

ViewUnpublishedDataset DownloadFile EditDataset DeleteDatasetDraft

Figure 5.11: A screen capture of the dataset user management view

but new datasets could not be uploaded. Moving the database to a new partition did not alone solve the problem since the failed dataset uploads had left partial files to the Dataverse partition that prevented the Dataverse from caching the new uploads. Once the fragmented remains of the failed uploads were removed the and Dataverse had disk space to cache new datasets the system started working again normally. It is good that the system was robust enough to handle the insufficient space on the partition, but not cleaning up the failed uploads caused frustration and the Dataverse documentation did not help in resolving the issue.

Interesting note on Dataverse is that it gives DOIs to the published datasets. This functionality is not documented, but digging into the code and the logs from the system it turns out that that Dataverse uses the EZID[26] service to supply DOIs for datasets. What is strange is that when you set up your own Dataverse you can just publish data and the DOIs will be given for you. This is strange because registering DOIs is not free and it is hard to imagine that Harvard would want to pay for all the DOIs for all the users of Dataverse.

In summary, Dataverse serves as a functional system to publishing research data. The software is of good quality but the documentation and instructions related to setting up the system could be better and the test coverage could be improved. The mechanical processes of uploading and searching datasets from the system work, but both of them could be made more intuitive. Computerized upload of datasets is a must have feature when dealing with datasets that contain a large amount of files. In addition to taking care of these technical matters, training for research data management and integration of research data publishing to the workflow of scientists is required.

[26] `http://ezid.cdlib.org/`

Chapter 6

Discussion

6.1 Methods

In this thesis we have examined the challenges around research data management, sharing and publishing. The methods used are literature reviews, interviews, questionnaires, technical benchmarks and user tests in the form of contextual interviews and lead user tests.

Literature reviews on the subject show that while the open access way of publishing research papers has been studied, the scientific contribution to the problem of open access research data has been tackled by a small number of researchers. Statistical studies of open access benefits with research papers promise good results for publishing in open access style, but the methodology and sample size quality varies. The metrics and numbers about research data sharing and management are all different within the research papers published, making it harder to interpret results from them. This is to be expected, since research data publication and sharing is a relatively new phenomenon. A challenge for the future would be to rigorously quantify the benefits of research data publication and sharing. Not many concrete numbers about the benefits of research data publishing and sharing are found from the literature. Many of the research papers in the field are also quite new, so it will take time to find out which ones of them provide to be the most valuable.

Interviews were used to find out the current situation of research data management and sharing in Finland. The interviewees represented many groups of stakeholders, but it is possible that the people chosen for the interviews were not the best representatives of their group. Legal issues were also brought up quite a few times during this thesis and it would have been beneficial to interview a legal expert as well. The interviews were conducted in places chosen by the interviewees to make the interview process easier for them. One possible problem with the interviews was that they were conducted alone, but this was taken into account by recording the interviews. Similar questions were asked from all the interviewees, but with many of them the deep discussions were done on different topics. Doing more interviews to gain more data points and following up on the interviewees on later parts of the research would have been beneficial. The Complex Network group was interviewed in two phases of the thesis (initial interviews and contextual interviews in the prototyping phase) and that proved valuable - hinting that revisiting other interviews would have been valuable as well. Time constraints of the thesis prevented this.

The questionnaires were not designed for the use of this thesis [43, 44]. Instead they were used to find out the research data management needs of Aalto for planning the future of services offered by Aalto. However, the questions asked in the surveys were very relevant for this thesis as well which led to the decision to use them as is instead of doing overlapping work. The general caveats of surveys, such as the shallowness of information and the chance of misinterpretation, apply of course. To minimize the the risk of misinformation the results from the surveys were shown as is and interpretation was kept to a minimum. Repeating the questionnaires could have been valuable since they were conducted two years ago. We did not spend resources on this since the deeper user interview and test feedback was considered higher value.

Technical benchmarking was done with existing solutions to a somewhat varying degree. The Dataverse solution was inspected thoroughly, Zenodo and Hydra installations were tried, other publishing platforms were signed up for and tried that way, and iRODS was studied through the documentation and source code. There is a possibility that the conclusions drawn from the thorough inspection of Dataverse do not apply for the other solutions directly, but since the implementations are quite similar the risk is likely low. We have tried separating the learnings from the systems from the system specific features, such as the look and feel of their UIs. The cloud environment that is used as the running environment of the prototype Dataverse system would not be the installation destination of the final system. This did not generate problems for the testing of the system. More hands on testing on the different technical solutions would have been beneficial, but this was not feasible in the time frame of the thesis.

The contextual interviews and lead user tests were chosen as a tool to learn more about the research data publication systems to complement the technical examination. Questionnaires would be another option, but while you could get more data points from a survey, the information is superficial. One angle for this thesis is that Aalto University is forming a data policy to to govern research data management and publishing and gaining deeper user insights through controlled user tests would provide deeper insights for that as well. Additionally, surveys on the subject had already been made and they are introduced in Section 4.7. The sample size of 12 users (10 for the contextual interviews and 2 for the lead user tests) does not yield statistically significant results, but the depth that the examination was carried out with should provide value. Enlisting more lead users would have been beneficial but we could not find users with enough time to test the system.

It was interesting that the interactions with the users of the system brought up similar points that were raised in the literature. The lack of culture and same reasons for not sharing research data were found out within the users as were found out in the literature. The sample size is not, again, scientifically significant, but it seems likely that the problems reported in literature around the world do happen in Finland as well.

The goal of combining these methods was to gain a holistic view on the problem and not only focus on the technological side. More user engagement would have given a better view, but on the whole we feel that the overall view created by this thesis is representative of current state of the research data sharing, publishing and management in the field of science.

The chosen methodologies were chosen also partially to accommodate the studies of the writer of this thesis - I have studied user centered design as a my minor during my Masters' studies.

6.2 Combining insights

While the main focus of this thesis from the beginning was to find appropriate technical solutions for sharing research data it became clear that sharing research data is not strictly a technical problem. While there is a lot of work to be done to implement tools and systems to make the process of sharing and publishing research data easy, a point solution to do just that would fall flat. Firstly, research data publication and sharing is closely tied to research data management during the research process. If research data is not handled during the process with the goal of one day making it public, the process of eventually making it public is very hard or impossible. And even though it might be technically viable the time and effort it would take to turn datasets that have not been properly documented and maintained during the process might be prohibitive. Secondly, the culture for sharing research data is still developing. There is not enough knowledge about either research data management or sharing. Additionally, there is no incentive to share research data, since researchers' contribution is measured mainly in citation to research papers.

The problem of research data also involves people from different disciplines in unprecedented ways. The roles of libraries, university software infrastructure maintainers and the researchers themselves are undefined in the new world of research data management, publication and sharing. When you consider the academic publishers and the peer review process that is at the heart of scientific publishing it is clear that the roles of all these actors need to be defined in the research data context.

Making research data accessible is likely to promote the quality of science. To achieve this, better technical solutions need to be implemented and the culture around research data needs to be taken into a more open direction.

Examining the state and options that come with research data management, publishing and sharing four survival strategies for the data intensive world of research can be outlined. The strategies are, in order from the most recommendable to the least recommendable, international collaboration, open source solutions, national collaboration and implementing own solutions.

International collaboration entails being a part of a international initiative for research data management, sharing and publication. EUDAT is an example of an initiative like this. This makes collaboration with others that are working in the initiative easier, takes the implementation load off the institutions while giving them a chance to participate and lessens the amount of existing systems and standards. The challenge is to adapt the international solutions to conform to the local needs of research institutions.

Implementing one of the existing open source solutions in a research institution would bring that institution to the community around that solution. Example solutions include Dataverse, CKAN and Invenio. The problem with institutions implementing different open source solutions is that it increases the overall complexity of the research data landscape and puts the burden of maintaining the solution to the research institution.

National initiatives, such as the Open Research and Science initiative in Finland, take the burden of implementation from the research institutions. The problem is that research nowadays is global and the national providers are likely to enable national collaboration very well, but international collaboration would be a challenge.

Implementing a novel solution for a research institution is an option, but seems very inefficient since solutions exist already. While you might get a perfectly tuned system for your institution, it would be more cost efficient to use existing solutions. Implementing own solutions also means that the burden to integrate with other systems is on you.

This thesis has also identified many aspects of a solution that could solve the research data management, publishing and sharing problems. These aspects are boiled down to a set of design requirements. The requirements can be used as a basis to design solutions or as means to validate existing solutions.

These requirements are split into must have requirements, functional requirements, hardware requirements and user experience requirements. Other than the must have requirements that a system either does or does not fulfill the requirements contain metrics that can be used to validate if the system fulfills the requirement. The formulation of the requirements presented here is quite general, which means that in order to use them to design or validate a solution the metrics presented in the requirements should be made more specific. The requirements are also not instructions set in stone - they should be used to guide in developing a research data management, sharing and publishing system.

The must have requirements are presented in Tables 6.1 and 6.2. The other requirements are presented in Appendix B.

Table 6.1: Must have requirements for a research data publishing solution

Requirement	**Rationale**
Users can upload datasets	Without datasets, there is not research data publishing
User can upload relevant metadata for their datasets	Without metadata, finding and reusing research data is impossible

The metadata of datasets is full text searchable	In order to find datasets, the metadata has to be searchable
Published datasets are assigned a persistent identifiers	Persistent identifiers allow for referencing the data and maintaining links for longer
There are no restrictions to the type of uploaded data	Research data comes in many shapes and forms and all data must be able to be published
The datasets in the system can be made public for all the world to see	The goal of the publishing platform is to make the data public

Table 6.2: The must have requirements of a research data management solution

Requirement	**Rationale**
Users can upload their research data during the research project	Instead of per researcher storage solution for storing research data, there must be a centralized solution
The tool imposes no restrictions to the type of research data stored	Research data comes in all shapes and forms and the tool must serve all the disciplines
The research data management tool must allow for sharing data with collaborators	Research is done globally and collaboratively nowadays, meaning that data must be shared with collaborators

6.3 Future work

Future work around the subject should include both integrated technical solutions that make good research data management practices a part of the daily life of researchers. This would in turn make the publication of their data easy and help improving the culture around the subject. To change the culture more research on the benefits of open research data is required, but there might be other ways as well.

Some fields of science, such as physics, psychology and genomics have been more successful than other fields in implementing open research data practices. Research on how these fields managed that transition and how that could be applied to other fields could find ways to bring open research data to other fields as well.

It is also possible that the culture of research data could be changed with appropriate sticks and carrots. If funding bodies would make it a priority to demand public research data, there would likely be an urgency to provide such solutions for researchers as soon as possible. On the other hand, maybe citations to research data could be integrated to the h-index of researchers, thus making the already used metric reward those who publish their datasets. This reframing of the problem - instead of asking ”how should we implement systems that help with research data?” we ask ”how do we motivate people to share their research data” - is a design paradigm that has been used successfully in other system level design problems [16].

The solutions for the technical and cultural problems might also be found from analogous problems. For example, software development has benefited for a long time from the open source community, where people contribute code for the public domain without monetary compensation. It is been studied that people do not do this out of the kindness of their hearts, but instead with the goal of turning their free work now into profit in the future [30]. The tools that are used to share software, such as GitHub, could also provide a fresh look on how tools for research data management and sharing could work.

Chapter 7

Conclusions

This thesis has examined different technical solutions to managing, sharing and publishing research data, focusing on sharing and publishing. Table 7.3 shows the features of solutions presented and shows how they differ in the context of research data publishing, whereas Table 7.5 shows the same solutions but in the context of research data management during a research project. Table 7.1 shows the legend that is used in Tables 7.3 and 7.5. Research data is referred to as data in the tables for a more concise presentation.

Table 7.1: The color coding of the summary Tables 7.3 and 7.5

Color and Mark	**Meaning**
++	The feature is well implemented and could be considered a strong point in the solution
+	The feature is implemented in the solutions in some way - it might need some work to get up and running or the other features of the solution can be used to approximate the functionality
-	The feature is missing from the solution or can be found from the solution but according to user feedback is too hard to use and is a clear weak point of the solution

In Tables 7.3 and 7.5 the IDA, Etsin, Avaa and PAS are a part of the Finnish Open Science Initiative. Dataverse, Hydra, Zenodo, iRODS and CKAN are open source solutions. EUDAT refers to the B2-offering[1], of which the B2Share and B2Drop are considered to be a part of a research data management and publication process. ACRIS is the Elsevier Pure instance that is being trialed at Aalto University. GitHub is included in the comparison since it is a widely adopted tool for publishing and collaborating on software projects.
The features used in the columns in Table 7.3 are explained in Table 7.2.

[1] `http://eudat.eu/`

Table 7.2: The features being compared on Table 7.3

Feature	**Explanation**
Data Storage	Data storage means that the solution offers the chance to save the actual research data to the system and not just links to the datasets, for example
Metadata Storage	Metadata storage means that the system has a structured way of storing metadata about datasets - this does not mean that the system should contain actual research data, since some services are built just as places to store metadata and links to datasets
Open Access Data	System that implements this feature allows the datasets to be searched and downloaded by the general public without restrictions to the use of the datasets
Restricted Data	Whereas open access data is available for all the world to see, systems that implement the restricted data feature allow for a more fine grained user management allowing datasets to be shared with restricted groups of users
Long Term Archival	Long term archival is archival for datasets that lasts for tens of years - a lot longer than commodity hard drives last in active use (commodity hard drives usually last from three to five years in active use)
Full Text Search	Full text search refers to the ability to search the all the metadata available in the system to find relevant datasets
Dataset Versioning	Dataset versioning means that the system allows for uploading newer versions of the already uploaded datasets without deleting the old datasets, since someone might use and refer to the old datasets
Identifier Scheme	Identifier scheme tells which, if any, persistent identifier schemes the solution implements

From the solutions presented in Table 7.3 the systems that are designed primarily for research data publishing (Avaa, Dataverse, Hydra, Zenodo, and CKAN) have almost all the features specified. What some of them are lacking is the ability to restrict access to the research data, which is a requirement for some datasets that contain data that has some confidentiality clauses. Avaa, for example, is a platform for nothing else than completely open data making it an awkward choice for a university, for example. None of these solutions implement long term archival, which is an important feature going forward and a goal for future development. These solutions are also remarkably similar, so any one of them could be used to set up a research data publishing platform for a research institution.

The only solution for long term archival is PAS, which is in test phase as of writing of this thesis. It ticks many boxes of the features, but it has to be taken into account that the PAS system is only for long term archival and can not be used for the normal publishing activity of research institutions. It is also unclear at the moment how datasets for long term archival are chosen, what kind of metadata long term archival requires and how the system should integrate to existing and upcoming research data repositories.

The tools that have a lot of other functionality in addition to research data publishing implement the features to a varying degree. iRODS, which is a tool for managing data, is clearly not designed to be a data sharing platform for research data. ACRIS has many other tasks as well, such as maintaining a the publication profile of researchers and reporting publications. EUDAT's B2Share offering offers a serviceable research data publication platform, but lacks the ability to restrict access to the datasets in a fine grained manner. IDA is an iRODS based system, but has some features for publishing datasets from the system. Biggest problem with it is that it is perceived very hard to use by users.

Etsin is a metadata publication platform - it only contains metadata and links to the actual location of the datasets. While this is valuable, separating the research data from the metadata requires either integration between the storage and metadata systems or more work from the users. This is also the reason why Open Access Data is marked as a weakness of the system - no data, no open access.

The identifier scheme is also a point of interest. Citing datasets is something that has not been widely adopted, which means that in order to help that adoption a well known identifier scheme should be used. It is good that a Finnish identifier scheme exists, but if datasets want international recognition an internationally known identifier scheme should be used.

The technical solutions with a focus on research data publishing fill their role adequately. The user tests show that there are user experience and usability improvements that could be done - those are detailed in Section 5.5. What the user tests also show is that training related to the options when it comes to research data publishing is lacking. If the culture and willingness to share research data is to be improved, that would require a conscious effort from the university side.

Table 7.3: Existing solutions in light of research data publishing

Tool	Data Storage	Metadata Storage	Open Access Data	Restricted Data	Long Term Archival	Full Text Search	Dataset Versioning	Identifier Scheme
IDA	++	+	+	+	-	++	-	URN
Etsin	-	++	-	-	-	++	-	URN
Avaa	++	++	++	-	-	++	-	URN
Dataverse	++	++	++	++	-	++	+	DOI/Handle
Hydra	++	++	++	++	-	++	+	DOI
Zenodo	++	++	++	+	-	++	+	DOI
EUDAT	++	++	++	+	-	++	+	Handle
iRODS	++	+	-	-	-	++	+	-
CKAN	++	++	++	++	-	++	+	+
ACRIS	-	+	+	+	-	++	-	+
PAS	++	++	++	-	++	++	-	URN
GitHub	++	+	++	++	-	++	++	DOI

Research data publishing is the end goal of research data management, and Table 7.5 shows the same tools compared previously in the context of research data publishing being compared in the light of research data management during the research process. Table 7.4 explains the features that were explored from the research data management angle.

Table 7.4: The features being compared on Table 7.5

Feature	Explanation
File Sharing With Partners	The basis of collaboration is sharing results and in this context it means that the solution enables files to be shared using the system - this also implies that the files can be stored there for personal use
File Editing With Partners	For collaboration just sharing files might not be enough, and the files should be able to be edited instead of always uploading new versions to the system
Workflow Sharing	The research workflows in some fields of science are highly automated and sharing those would save a lot of time for collaborators and the person sharing them assuming he or she had to return to that workflow later
Comments On Files	Commenting on others' work makes collaboration easier
Web Interface	A system implements this feature if it has an interface online that can be accessed from anywhere so that the research data is readily accessible
Command Line Interface	A system implements a command line interface if the users can access their research data using technologies like SSH
Desktop Interface	A system implements a desktop interface if it offers software that the users can install on their personal machines or it integrates to the file system on the user's machine
Integrated Data Collection	Integrated data collection means that tools that collect research data can be integrated to the solution such that the research data is collected and put into the systems automatically

The same tools that are suitable for research data publishing are not necessarily right for research data management and vice versa. iRODS, IDA and the B2Drop side of the EUDAT service offer dedicated services to share research data during the research process and manage it in a centralized service. GitHub is a collaborative tool for projects around programming.

The tools that work well for research data publishing lack many important features that research data management and research collaboration, such as file editing and proper commenting functionality on file. Research workflows cannot be stored or shared with any of the tools, which is not surprising -

Table 7.5: Existing solutions in light of research data management

Tool	File Sharing With Partners	File Editing With Partners	Workflow Sharing	Comments On Files	Web Interface	Command Line Interface	Desktop Interface	Integrated Data Collection
IDA	++	++	-	+	++	++	++	+
Etsin	-	-	-	-	++	++	-	-
Avaa	++	-	-	-	++	++	-	-
Dataverse	++	-	-	+	++	++	-	-
Hydra	++	-	-	+	++	++	-	-
Zenodo	++	-	-	+	++	++	-	-
EUDAT	++	++	-	+	++	++	++	-
iRODS	++	++	-	+	++	++	++	++
CKAN	++	-	-	+	++	++	-	-
ACRIS	++	-	-	-	++	++	-	-
PAS	++	-	-	-	++	++	-	-
GitHub	++	++	-	++	++	++	++	-

the definition of a workflow varies a lot between disciplines and it is not a requirement for many.

The user tests and interviews also show that the research management ways of researchers could be much better. Research data management is learned by doing and from peers and the awareness of good practices and tools for research data management is low. Good practices and tools should be taught to the researchers creating and managing the research data.

When looking at the existing solutions through both the context of research data management and publishing it is clear that a big problem is that there is not a solution that would handle the entire lifecycle of the data from creation to publishing. Research data sharing and managing should not be considered separately from research data management. Without management, there can not be effective sharing or publishing and without sharing or publishing there is no need for management. Many universities, Aalto University included, have implemented neither managing or publishing research data tools in a centralized way. The lack of dedicated research data management infrastructure also affects the training and knowledge of research data management and tools in the sense that if there is no infrastructure to leverage there is no way to teach institutional best practices.

This thesis set out to answer two research questions. Firstly, how research data can be shared and published using modern tools and how these tools work in practice? Secondly, what are the non technical matters that affect sharing and publishing research data? To answer the first question we have found, presented and tested existing software solutions for research data publishing and sharing. Through user testing and technical benchmarking we have concluded that the user experience of these solutions could be enhanced to make research data sharing and publishing easier, but the solutions do serve their purpose and can be used to share and publish research data. Additionally, the lack of tools to manage research data during the research process makes publishing and sharing research data harder. As for the second question, we have concluded using user interviews and user tests that the lack of culture, incentives and know-how about research data sharing, publishing and management is a major barrier to publishing and sharing research data.

Following these conclusions we have presented steps to make research data publishing and sharing prevalent in the field of science. In order to make the tools for sharing and publishing research data better, holistic solutions that combine research data management, publishing and sharing should be implemented. Solving the non technical problems of research data management, sharing and publishing starts from proving the benefits of sharing and publishing research data and from providing education and incentives for researchers.

Appendix A

Users of Data Repositories

A.1 Researcher

- As a researcher, I want to publish selected research datasets so that they can be useful to others and I can get citations.
- As a researcher, I want to have a single citeable URL and a persistent identifier for a dataset. This way I can publish my datasets and draw attention to my work and get citations.
- As a researcher, I want to publish a massive dataset in a way that others can access it, and not have to duplicate metadata entries to other hosting services.
- As a researcher, I want my data to be linked to my own pages and possibly a data profile page, so that I gain visibility to myself from releasing data.
- As a person applying for funding, I want a ready-made data publication solution that I can put into my applications, so that it saves me time when applying for funding and increases my chances to gain funding.
- As a researcher, I want to store metadata or data in the system, but in a way that I can guarantee that it will never be published or viewable to anyone unauthorized by accident and without explicit consent.
- As a researcher I want to easily find datasets online and downloading them should be easy.
- As a researcher, I want to know what other researchers are doing and what kind of data they are using. I might want to collaborate with them or use their data to make my research.
- As a researcher, I want my papers and research to have as much impact as possible, which is aided by publishing my datasets free online.
- As a researcher, I want to be sure that my data is safe so that I can focus on more important work.
- As a researcher, I want data management during my research project to be as easy as possible so that it does not cause unnecessary overhead.
- As a researcher, I want to follow the best practices of my field to comply facilitate collaboration with others.
- As a researcher, I want to be educated in research data management so that I can save time while doing my research.
- As a researcher and a new member of a research group I want to easily get a grasp of the research data and the practices the group is using to get to work faster.
- As a researchers, I might want to use my dataset first before making it public so that I get the first benefit of my dataset.

A.2 Course

- As a course instructor in a data-driven course, I want to put several small datasets online for use of students in my class, so that I have less data management to do myself in the long run.
- As a course instructor in a data-driven course, I want for my students to be able to find other datasets online, so that they may find other data which engages them more

than my own.

A.3 Research group

- As a research group, we want to be able to put dataset metadata in one central place, so that we do not lose memory of data and metadata over time as people come and go. This can be both public and private data.
- As a research group, we want to put metadata about our data where it can be browsed by others, so that we can find collaborators who may want to work with us. Some of this data should be visible for all and some should be restricted to smaller groups.
- As a research group, we want to have a data collection with permissions such that it is very easy to add new members to our group. Preferably, this is automatic using university user management system.
- As a research group, we want different levels of privacy of data to coexist in our data collection so that we need only one collection. For example, all data in our collection should be private by default, but some can be published.
- As a research group, we want to plug our research data collection devices to the research data management systems to automate our research process.
- As a research group, we want the research data system to integrate to our research workflows.

A.4 Librarian

- As a librarian, I want to be a part of the digital publishing of our institution's datasets.
- As a librarian, I want to use my metadata and description expertise to aid the researchers and to have their datasets properly documented and metadata properly used.
- As a librarian, I want the dataset publication system to communicate with other electronic publishing systems already in use in our institution
- As a librarian, I want to pass on my knowledge of metadata and electronic publishing also to the scientists so that they can both take that into account while they work and so that my work in helping them becomes easier.

A.5 Student

- As a student, I want easy access to any materials that my course requires me to use with my course work.
- As a student working on any level of thesis I would like to know if relevant datasets that I could use with my thesis exist and would like to have easy access to them.
- As a student working on different student projects that generate data, I would like to know the best practices that allow me to use the least time to manage my data and allow me to save them in a convenient location.

A.6 IT staff

- As an administrator, I want to be able to see the data produced and released by my unit in the last N years, so that I can document productivity and better apply for funding.

- As the research data repository administrator, I want tangible benefits to researchers, so that they feel that using this is worth their time and continue using long-term.
- As an administrator, I want the system to be easy to maintain, so that it do not cause extraneous overhead with my other maintenance tasks.
- As an administrator, I want the data management plans researchers make to contain the whole lifecycle of the research data in order to preserve the research data the best way possible.

A.7 Others

- As a developer of applications, I want to have access to interesting datasets that could potentially be used as a basis of applications.
- As a funding body, I want that the access to the datasets that should be public is easy and that the publishing system does not allow them to be published in a way that they are hard to find.

Appendix B

Requirements for Research Data Management and Publishing System

B.1 Functional Requirements for Research Data Publishing

Requirement	**Metric**	**Rationale**
The publishing platform can host and serve big files	The file maximum size is in the order of tens of gigabytes	The platform does not need to serve big data, but even ”normal” data can be gigabytes in size
Users can sort the search results by different factors	The search results can be sorted by various sorting criteria, e.g. dataset publication date, dataset language, dataset creation date	If the repository contains many datasets, searching for relevant ones needs to be made easier
The metadata templates offered by the system satisfy the needs of different fields of science	The metadata templates follow the existing metadata standards or best practices	Metadata should be descriptive of the data and understandable by others than the creators of the data
Data can be uploaded programmatically	The platform offers and API with associated documentation	If you possess a lot of data, you cannot upload it all manually
Data can in the system can be visualized	Data visualization follows the best practices of the field	Good visualizations give an easy overview to the data
Access control to datasets should be fine grained	Access to the datasets can be controlled by e.g. user account, groups of users, IP address range or embargoes	Sometimes you need the data to be only visible to a subset or there might be an embargo on the data – for example, students of Aalto University should see a certain dataset but not everyone in the world

The data repository must be secure	The data repository satisfies industry standard requirements for security	The data repository may contain confidential information which should be safe
Users in the system can be given different roles	The different roles are e.g. administrator, curator, contributor and guest	Publishing research data involves many parties, such as librarians and IT staff, and their roles are not about creating data but manage the system and curate the data
The system should be integrated to the already existing user management system	The system's user management module is extensible to integrate external user management system	Not only do new students come in every year, staff changes all the time – and keeping multiple systems up to date is a cause for problems
Unknown vocabulary should be made clear to the users	When encountered with unknown vocabulary, the user must get help within the same screen	Research data has a lot of vocabulary that is not all known for researchers – helping them to understand that is important
Users can add tags to their datasets and files	Tags can be added to both datasets and files	Tags allow the system to group similar datasets and files to help find relevant ones
Users can to filter search results	The filtering can be done by e.g. field of science, dataset size or creator of the data	If the repository contains many datasets, finding the relevant ones is easier from filtered results
The system can be integrable with systems for long term archival of selected important datasets	Selected research datasets can be stored for over 20 years in a long term archival system	Long term archival allows for the persistence of science and continued use of the datasets
The system enables different versions of the dataset	The system can store and display different versions of the dataset	Datasets can be worked on and changed, but the old versions should be available since someone might use the older versions of the data
The system allows for downloadable citations to the dataset	The system gives citations in common formats	Citing datasets is much easier when you can get citations right out of the system

B.2 Functional Requirements for Research Data Management

Requirement	Metric	Rationale
Research data management tool must not interfere with the research work	Research data management must not take more than 5% of researcher's time	If the tool causes problems it is not going to be used
Publishing data from the research data management tool should be easy	It should take fewer than 5 interactions with the system to publish a dataset from the research data management tool	There is no point in separating the systems and making publishing easy it makes publishing more likely
The tool must be able to store metadata in addition to storing the actual data	The metadata in the system should follow the metadata standards of the publishing platform	Data without metadata is incomplete and making metadata a part of the research data management tool would promote metadata even during the research process
The tool should have a graphical user interface	There is a graphical user interface to the system that is available through the Internet	Not all researchers like a command line interface
The tool should have a command line interface	There is a command line interface to the system that can be accessed for example with SSH	Not all researchers like graphical user interfaces
The tool should be integrable to data collection devices	Once integrated, the users do not need to manually do anything to upload the raw data to the system	The more computerized the research process is the better it is for the researchers
The system can be integrated to research workflow systems	The system offers APIs to integrate it to existing workflow systems	Some research projects have research workflows that automate the research process and that should be accommodated
The system allows for comments on shared files	The system has a commenting function for the objects in the system	Comments help with collaboration on the files

The system uses the existing user management system	The system integrates to the existing user management system of the institution	We do not want to add more user accounts or points of confusion to the user
The research data should be accessible from anywhere	The system does not filter traffic based on an IP address range, but the system should use the user management system to authenticate users	Researchers might want to work from home, for example

B.3 Hardware Requirements

Requirement	**Metric**	**Rationale**
Data is stored within the legal geographical limits	The hardware of the system is located such that the legal constraints are satisfied (for example, in the case of Finland the hardware is in Finland)	Some data cannot leave borders of countries
The data must be backed up for disaster recovery	The data storage is designed following good standards and principles	The research data is sensitive and should be backed up in case of failures – full back up on tapes is not feasible, but good enough replication is required
The hardware must adhere to the safety requirements of the data concerning privacy and other confidentiality	The hardware conforms to the safety standards of the data	Some data must be safer than other data – however, the most extreme cases should be handled separately
The hardware should be virtualized by the software on top of it	The hardware can be changed and the system on top of it does not need to be changed	The research data lifespan is likely to outlive several compute and storage hardware generations and hardware will always eventually fail

B.4 User Experience Requirements

The user experience requirements shown here are not exhaustive - in order to serve users best they should be involved in designing the system.

Requirement	Metric	Rationale
The system clearly shows what metadata is needed for the upload of datasets	Uploaded datasets contain all the relevant metadata	Metadata is important and making sure that users input all of it makes the quality of data better
Search results are displayed in a way that even if there are a lot of search results the user is not confused	The data deluge from search does not hide the dataset the user is actually searching for	When the system contains many datasets, displaying search results becomes a pain
The user is provided with examples on how to write clear descriptions of their datasets	The descriptions are short and follow the best practices of the field	The free form descriptions of datasets are the most important when searching and trying to find relevant datasets to gain a basic understanding of the dataset
The research data management tool is as easy to use as the commercial cloud services	When compared in user tests, the research data tool gets as good scores as the commercial ones	The commercial cloud services will be the clear benchmark for the research data management tool and thus should be as good or better than those services
When searching, the user should be prompted to narrow down the search if there are too many results	If there are over a hundred search results, the user should be prompted to narrow down the search	Users get frustrated if they can't find what they are looking for and prompting them to narrow down the search helps them find results
The users have access to an overview of their data, published and unpublished	The publishing platform and the research data management tool are integrated to a common dashboard view	The users want to keep track of their projects and maybe want to even show off their current contribution
The data visualization should be interactive when applicable	The user can define the visualization parameters and subsets of the data for the visualization	Visualizing data is a way to understand the data

The user should be rewarded for using the systems	The system generates reports or badges the user can show on his or her online profiles	Sharing research data needs to give a reward to the user
The system can highlight published datasets from time to time	The front page of the system has dedicated space for highlighted datasets	Highlighting datasets is a way of rewarding users and getting exposure for the datasets

Bibliography

[1] Alawi A. Alsheikh-Ali, Waqas Qureshi, Mouaz H. Al-Mallah, and John P. A. Ioannidis. Public availability of published research data in high-impact journals. *PLoS ONE*, 6(9):e24357, 09 2011. URL http://dx.doi.org/10.1371%2Fjournal.pone.0024357.

[2] Peter W. Arzberger, Peter Schroeder, Anne Beaulieu, Geoffrey C. Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, and Paul Wouters. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3:135–152, 2004. URL http://dx.doi.org/10.2481/dsj.3.135.

[3] Chris Awre, Tom Cramer, Richard Green, Lynn McRae, Bess Sadler, Tim Sigmon, Thornton Staples, and Ross Wayland. Project Hydra: Designing & building a reusable framework for multipurpose, multifunction, multi-institutional repository-powered solutions. 2009. URL http://hdl.handle.net/1853/28496.

[4] Fran Berman, Ross Wilkinson, and John Wood. Building global infrastructure for data sharing and exchange through the research data alliance. *D-Lib Magazine*, 20(1/2), 2014. URL http://www.dlib.org/dlib/january14/01guest_editorial.html.

[5] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001. URL http://iir.ruc.edu.cn/pdf/The%20Semantic%20Web.pdf.

[6] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009. URL http://dx.doi.org/10.4018/jswis.2009081901.

[7] Christine L. Borgman. The conundrum of sharing research data. *JASIST*, 63(6):1059–1078, 2012. URL http://dx.doi.org/10.1002/asi.22634.

[8] Jerome Caffaro and Samuele Kaplun. Invenio: A modern digital library for grey literature. Technical Report CERN-OPEN-2010-027, CERN, Geneva, Dec 2010. URL http://cds.cern.ch/record/1312678.

[9] James J Cimino and Elaine J Ayres. The clinical research data repository of the US National Institutes of Health. *Studies in health technology and informatics*, 160(Pt 2):1299, 2010. URL http://dx.doi.org/10.3233/978-1-60750-588-4-1299.

[10] Anna Clements and Valerie McCutcheon. Research data meets research information management: Two case studies using (a) Pure CERIF-CRIS

and (b) EPrints repository platform with CERIF extensions. In Keith G. Jeffery, Anna Clements, Pablo de Castro, and Daniela Luzi, editors, *12th International Conference on Current Research Information Systems, CRIS 2014 - Managing data intensive science - The role of Research Information Systems in realising the digital agenda, Rome, Italy, May 13-15, 2014*, volume 33 of *Procedia Computer Science*, pages 199–206. Elsevier, 2014. URL `http://dx.doi.org/10.1016/j.procs.2014.06.033`.

[11] Melissa H Cragin, Carole L Palmer, Jacob R Carlson, and Michael Witt. Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1926):4023–4038, 2010. URL `http://dx.doi.org/10.1098/rsta.2010.0165`.

[12] Iain D. Craig, Andrew M. Plume, Marie E. McVeigh, James Pringle, and Mayur Amin. Do open access articles have greater citation impact?: A critical review of the literature. *J. Informetrics*, 1(3):239–248, 2007. URL `http://dx.doi.org/10.1016/j.joi.2007.04.001`.

[13] Philip M Davis, Bruce V Lewenstein, Daniel H Simon, James G Booth, Mathew JL Connolly, et al. Open access publishing, article downloads, and citations: Randomised controlled trial. *BMj*, 337:a568, 2008. URL `http://dx.doi.org/10.1136/bmj.a568`.

[14] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, and Cees de Laat. Addressing big data challenges for scientific data infrastructure. In *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, CloudCom 2012, Taipei, Taiwan, December 3-6, 2012*, pages 614–617. IEEE Computer Society, 2012. ISBN 978-1-4673-4511-8. URL `http://dx.doi.org/10.1109/CloudCom.2012.6427494`.

[15] Peter Doorn, Ingrid Dillo, and René van Horik. Lies, damned lies and research data: Can data sharing prevent data fraud? *IJDC*, 8(1):229–243, 2013. URL `http://dx.doi.org/10.2218/ijdc.v8i1.256`.

[16] Kees Dorst. *Frame Innovation: Create New Thinking by Design*. MIT Press, 2015.

[17] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002. URL `http://dx.doi.org/10.1093/nar/30.1.207`.

[18] Gunther Eysenbach. Citation advantage of open access articles. *PLoS Biol*, 4(5):e157, 05 2006. URL `http://dx.doi.org/10.1371%2Fjournal.pbio.0040157`.

[19] Stephen E Fienberg, Margaret E Martin, Miron L Straf, et al. *Sharing research data.* National Academies, 1985.

[20] Beth A. Fischer and Michael J. Zigmond. The essential nature of sharing in science. *Science and Engineering Ethics*, 16(4):783–799, 2010. URL `http://dx.doi.org/10.1007/s11948-010-9239-x`.

[21] Matias Frosterus, Eero Hyvönen, and Joonas Laitio. DataFinland - A semantic portal for open and linked datasets. In Grigoris Antoniou, Marko Grobelnik, Elena Paslaru Bontas Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Z. Pan, editors, *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29 - June 2, 2011, Proceedings, Part II*, volume 6644 of *Lecture Notes in Computer Science*, pages 243–254. Springer, 2011. ISBN 978-3-642-21063-1. doi: 10.1007/978-3-642-21064-8_17. URL `http://dx.doi.org/10.1007/978-3-642-21064-8_17`.

[22] Yassine Gargouri, Chawki Hajjem, Vincent Larivière, Yves Gingras, Les Carr, Tim Brody, and Stevan Harnad. Self-selected or mandated, open access increases citation impact for higher quality research. *CoRR*, abs/1001.0361, 2010. URL `http://arxiv.org/abs/1001.0361`.

[23] Jeremy Goecks, Anton Nekrutenko, James Taylor, et al. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010. URL `http://genomebiology.com/2010/11/8/R86`.

[24] Pieter Van Gorp and Steffen Mazanek. SHARE: A web portal for creating and sharing executable research papers. In Sato et al. [56], pages 589–597. URL `http://dx.doi.org/10.1016/j.procs.2011.04.062`.

[25] Marjan Grootveld and Jeff van Egmond. Peer-reviewed open research data: Results of a pilot. *IJDC*, 7(2):81–91, 2012. URL `http://dx.doi.org/10.2218/ijdc.v7i2.231`.

[26] Chawki Hajjem, Stevan Harnad, and Yves Gingras. Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *CoRR*, abs/cs/0606079, 2006. URL `http://arxiv.org/abs/cs/0606079`.

[27] Stevan Harnad and Tim Brody. Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-lib Magazine*, 10(6), 2004. URL `http://eprints.soton.ac.uk/id/eprint/260207`.

[28] Harnad, Stevan and Brody, Tim and Vallieres, Francois and Carr, Les and Hitchcock, Steve and Gingras, Yves and Oppenheim, Charles and Stamerjohanns, Heinrich and Hilf, Eberhard R. The access/impact problem and

the green and gold roads to open access. *Serials review*, 30(4):310–314, 2004. URL http://dx.doi.org/10.1080/00987913.2004.10764930.

[29] Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. Research electronic data capture (redcap) - A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42(2):377–381, 2009. URL http://dx.doi.org/10.1016/j.jbi.2008.08.010.

[30] Alexander Hars and Shaosong Ou. Working for free? - Motivations of participating in open source projects. In *34th Annual Hawaii International Conference on System Sciences (HICSS-34), January 3-6, 2001, Maui, Hawaii, USA*. IEEE Computer Society, 2001. doi: 10.1109/HICSS.2001.927045. URL http://dx.doi.org/10.1109/HICSS.2001.927045.

[31] P Bryan Heidorn. The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51(7-8):662–672, 2011. URL http://dx.doi.org/10.1080/01930826.2011.601269.

[32] Tony Hey, Stewart Tansley, and Kristin M. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009. ISBN 978-0982544204. URL http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

[33] Jorge E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, 2010. URL http://dx.doi.org/10.1007/s11192-010-0193-9.

[34] Hans Jørn Nielsen Hjørland, Birger Helene Høyrup, Hans Jørn Nielsen, and Birger Hjørland. Curating research data: The potential roles of libraries and information professionals. *Journal of Documentation*, 70(2):221–240, 2014. URL http://dx.doi.org/10.1108/JD-03-2013-0034.

[35] James Honaker and Vito D'Orazio. Statistical modeling by gesture: A graphical, browser-based statistical interface for data repositories. In Federica Cena, Altigran Soares da Silva, and Christoph Trattner, editors, *Hypertext 2014 Extended Proceedings: Late-breaking Results, Doctoral Consortium and Workshop Proceedings of the 25th ACM Hypertext and Social Media Conference (Hypertext 2014), Santiago, Chile, September 1-4, 2014.*, volume 1210 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014. URL http://ceur-ws.org/Vol-1210/datawiz2014_05.pdf.

[36] Barbara R Jasny, Gilbert Chin, Lisa Chong, and Sacha Vignieri. Again, and again, and again.... *Science*, 334(6060):1225–1225, 2011. URL http://dx.doi.org/10.1126/science.334.6060.1225.

[37] Charles W. Bailey Jr. Open access and libraries. *Collection Management*, 32(3-4):351–383, 2007. URL `http://dx.doi.org/10.1300/J105v32n03_07`.

[38] Jane Kaye. The tension between data sharing and the protection of privacy in genomics research. *Annual review of genomics and human genetics*, 13:415, 2012. URL `http://dx.doi.org/10.1146%2Fannurev-genom-082410-101454`.

[39] Gary King. An introduction to the Dataverse network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2):173–199, 2007. URL `http://dx.doi.org/10.1177/0049124107306660`.

[40] Bartha Maria Knoppers, Jennifer R Harris, Anne Marie Tassé, Isabelle Budin-Ljøsne, Jane Kaye, Mylène Deschênes, and Ma'n H Zawati. Towards a data sharing code of conduct for international genomic research. *Genome Med*, 3(7):46, 2011. URL `http://www.genomemedicine.com/content/3/7/46`.

[41] Ilari Korhonen and Miika Nurminen. Development of a native cross-platform iRODS GUI client. pages 21–28, 2015. URL `http://urn.fi/URN:NBN:fi:jyu-201509213186`.

[42] Mikael Laakso, Patrik Welling, Helena Bukvova, Linus Nyman, Bo-Christer Björk, and Turid Hedlund. The development of open access journal publishing from 1993 to 2009. *PLoS ONE*, 6(6):e20961, 06 2011. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0020961`.

[43] Ilari Lähteenmäki. Questionnaire: Aalto-yliopiston tietotekniikkapalveluiden tutkimustallennuksen esiselvitys- projekti, 2013.

[44] Ilari Lähteenmäki and Janne Torkkel. Questionnaire: Tutkimustiedon tallennus - kyselytuloksia ja tilannekatsaus, 2013.

[45] Damien Lecarpentier, Alberto Michelini, and Peter Wittenburg. The building of the EUDAT cross-disciplinary data infrastructure. In *EGU General Assembly Conference Abstracts*, volume 15 of *EGU General Assembly Conference Abstracts*, page 7202, April 2013. URL `http://adsabs.harvard.edu/abs/2013EGUGA..15.7202L`.

[46] Luis Martinez-Uribe and Stuart Macdonald. User engagement in research data curation. In Maristella Agosti, José Luis Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, editors, *Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings*, volume 5714 of *Lecture Notes in Computer Science*, pages 309–314. Springer, 2009. ISBN 978-3-642-04345-1. URL `http://dx.doi.org/10.1007/978-3-642-04346-8_30`.

[47] Geontae Noh, Ji Young Chun, and Ik Rae Jeong. Sharing privacy protected and statistically sound clinical research data using outsourced data storage. *J. Applied Mathematics*, 2014:381361:1–381361:12, 2014. URL `http://dx.doi.org/10.1155/2014/381361`.

[48] Piotr Nowakowski, Eryk Ciepiela, Daniel Harezlak, Joanna Kocot, Marek Kasztelnik, Tomasz Bartynski, Jan Meizner, Grzegorz Dyk, and Maciej Malawski. The collage authoring environment. In Sato et al. [56], pages 608–617. doi: 10.1016/j.procs.2011.04.064. URL `http://dx.doi.org/10.1016/j.procs.2011.04.064`.

[49] Roger D Peng. Reproducible research in computational science. *Science (New York, Ny)*, 334(6060):1226, 2011. URL `http://dx.doi.org/10.1126%2Fscience.1213847`.

[50] Luca Pireddu, Simone Leo, Nicola Soranzo, and Gianluigi Zanetti. A Hadoop-Galaxy adapter for user-friendly and scalable data-intensive bioinformatics in Galaxy. In Pierre Baldi and Wei Wang, editors, *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14, Newport Beach, California, USA, September 20-23, 2014*, pages 184–191. ACM, 2014. ISBN 978-1-4503-2894-4. URL `http://doi.acm.org/10.1145/2649387.2649429`.

[51] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3):e308, 03 2007. URL `http://dx.plos.org/10.1371/journal.pone.0000308`.

[52] Heather A Piwowar, Todd J Vision, and Michael C Whitlock. Data archiving is a good investment. *Nature*, 473(7347):285–285, 2011. URL `http://dx.doi.org/10.1038/473285a`.

[53] Jean-Baptiste Poline, Janis L. Breeze, Satrajit S. Ghosh, Krzysztof J. Gorgolewski, Yaroslav O. Halchenko, Michael Hanke, Christian Haselgrove, Karl G. Helmer, David B. Keator, Daniel S. Marcus, Russell A. Poldrack, Yannick Schwartz, John Ashburner, and David N. Kennedy. Data sharing in neuroimaging research. *Front. Neuroinform.*, 2012, 2012. URL `http://dx.doi.org/10.3389/fninf.2012.00009`.

[54] Arcot Rajasekar, Reagan Moore, Chien-Yi Hou, Christopher A. Lee, Richard Marciano, Antoine de Torcy, Michael Wan, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Paul Tooby, and Bing Zhu. *iRODS Primer: Integrated Rule-Oriented Data System.* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2010. URL `http://dx.doi.org/10.2200/S00233ED1V01Y200912ICR012`.

[55] David De Roure, Carole A. Goble, and Robert Stevens. The design and realisation of the my$_{\text{experiment}}$ virtual research environment for social sharing of workflows. *Future Generation Comp. Syst.*, 25(5):561–567, 2009. URL `http://dx.doi.org/10.1016/j.future.2008.06.010`.

[56] Mitsuhisa Sato, Satoshi Matsuoka, Peter M. A. Sloot, G. Dick van Albada, and Jack Dongarra, editors. *Proceedings of the International Conference on Computational Science, ICCS 2011, Nanyang Technological University, Singapore, 1-3 June, 2011*, volume 4 of *Procedia Computer Science*, 2011. Elsevier. URL `http://www.sciencedirect.com/science/journal/18770509/4`.

[57] Caroline J. Savage and Andrew J. Vickers. Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE*, 4(9):e7078, 09 2009. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0007078`.

[58] Li Si, Yueting Li, Xiaozhe Zhuang, Wenming Xing, Xiaoqin Hua, Xin Li, and Juanjuan Xin. An empirical study on the performance evaluation of scientific data sharing platforms in China. *Library Hi Tech*, 33(2):211–229, 2015. URL `http://dx.doi.org/10.1108/LHT-09-2014-0093`.

[59] Dylan A. Simon, Andrew S. Gordon, Lisa Steiger, and Rick O. Gilmore. Databrary: Enabling sharing and reuse of research video. In Paul Logasa Bogen II, Suzie Allard, Holly Mercer, Micah Beck, Sally Jo Cunningham, Dion Hoe-Lian Goh, and Geneva Henry, editors, *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries, Knoxville, TN, USA, June 21-25, 2015*, pages 279–280. ACM, 2015. ISBN 978-1-4503-3594-2. URL `http://doi.acm.org/10.1145/2756406.2756951`.

[60] Peter Suber. Open access overview. 2007. URL `http://legacy.earlham.edu/~peters/fos/overview.htm`.

[61] A. Swan, Yassine Gargouri, M. Hunt, and Stevan Harnad. Open access policy: Numbers, analysis, effectiveness. *CoRR*, abs/1504.02261, 2015. URL `http://arxiv.org/abs/1504.02261`.

[62] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6):e21101, 06 2011. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0021101`.

[63] Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*, 10(8):e0134826, 08 2015. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0134826`.

[64] Michael C Whitlock. Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, 26(2):61–65, 2011. URL `http://dx.doi.org/10.1016/j.tree.2010.11.006`.

[65] Jelte M Wicherts, Denny Borsboom, Judith Kats, and Dylan Molenaar. The poor availability of psychological research data for reanalysis. *American Psychologist*, 61(7):726, 2006. URL `http://dx.doi.org/10.1037/0003-066X.61.7.726`.

[66] Jelte M. Wicherts, Marjan Bakker, and Dylan Molenaar. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, 6(11):e26828, 11 2011. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0026828`.

[67] Joss Winn et al. Open data and the academy: An evaluation of CKAN for research data management. 2013. URL `https://rd-alliance.org/system/files/documents/CKANEvaluation.pdf`. IASSIST, Cologne.

[68] Michael Witt. Institutional repositories and research data curation in a distributed environment. *Library Trends*, 57(2):191–201, 2008. URL `http://dx.doi.org/10.1353/lib.0.0029`.

[69] Jingfeng Xia and Katie Nakanishi. Self-selection and the citation advantage of open access articles. *Online Information Review*, 36(1):40–51, 2012. URL `http://dx.doi.org/10.1108/14684521211206953`.